Investigação Operacional 2013

Atas do

XVI Congresso

da Associação Portuguesa de Investigação Operacional

Bragança
3 a 5 de junho de 2013

Editado por
José F. Oliveira
Clara B. Vaz

Escola Superior de Tecnologia e Gestão
Instituto Politécnico de Bragança

Atas do XVI Congresso da Associação Portuguesa de Investigação Operacional

Editores:

José Fernando Oliveira

Clara Bento Vaz

Com a colaboração de:

Ana Isabel Pereira

Instituto Politécnico de Bragança
3 a 5 de junho 2013

Este volume contém artigos submetidos e apresentados no XVI Congresso da Associação Portuguesa de Investigação Operacional, realizado em Bragança, Portugal, de 3 a 5 de junho de 2103.

Título: **Livro de Atas do XVI Congresso da Associação Portuguesa de Investigação Operacional**

Editores:
José Fernando Oliveira
Clara Bento Vaz

Colaboração:
Ana Isabel Pereira

Primeira edição, em formato eletrónico, junho 2013
ISBN: 978-972-745-154-8

Comissão de Programa

José Fernando Oliveira (Presidente), Universidade do Porto, Faculdade de Engenharia
Agostinho Agra, Universidade de Aveiro, Departamento de Matemática
Ana Isabel Pereira, Instituto Politécnico de Bragança, Escola Superior de Tecnologia e Gestão
Ana Paula Teixeira, Universidade de Trás-os-Montes e Alto Douro, Departamento de Matemática e CIO
Ana Viana, Instituto Politécnico do Porto, Instituto Superior de Engenharia
Clara Bento Vaz, Instituto Politécnico de Bragança, Escola Superior de Tecnologia e Gestão
Filipe Alvelos, Universidade do Minho, Escola de Engenharia
Isabel Gomes, Universidade Nova de Lisboa, Faculdade de Ciências e Tecnologia
João Luís Soares, Universidade de Coimbra, Faculdade de Ciências e Tecnologia
Joaquim Borges Gouveia, Universidade de Aveiro, Dep. de Economia, Gestão e Engenharia Industrial
Jorge Orestes Cerdeira, Universidade Nova de Lisboa, Faculdade de Ciências e Tecnologia
José Manuel Valério de Carvalho, Universidade do Minho, Escola de Engenharia
Margarida Vaz Pato, Universidade Técnica de Lisboa, Instituto Superior de Economia e Gestão
Maria Antónia Carravilla, Universidade do Porto, Faculdade de Engenharia
Maria Eugénia Captivo, Universidade de Lisboa, Faculdade de Ciências e CIO
Maria João Alves, Universidade de Coimbra, Faculdade de Economia
Marília Pires, Universidade do Algarve, Faculdade de Ciências e Tecnologia
Miguel Constantino, Universidade de Lisboa, Faculdade de Ciências
Mónica Oliveira, Universidade Técnica de Lisboa, Instituto Superior Técnico
Susana Relvas, Universidade Técnica de Lisboa, Instituto Superior Técnico

Comissão Organizadora

Clara Bento Vaz, Instituto Politécnico de Bragança (Presidente)
Ana Isabel Pereira, Instituto Politécnico de Bragança
António Jorge Trindade Duarte, Instituto Politécnico de Bragança
Carla Alexandra Soares Geraldes, Instituto Politécnico de Bragança
Carla Maria Carneiro Alves, Instituto Politécnico de Bragança
Carla Sofia Renca da Cruz, Instituto Politécnico de Bragança
Carla Sofia Veiga Fernandes, Instituto Politécnico de Bragança
Carlos Jorge da Rocha Balsa, Instituto Politécnico de Bragança
Edite Martins Cordeiro, Instituto Politécnico de Bragança
Elisa Margarida Correia de Barros, Instituto Politécnico de Bragança
Florbela Alexandra Pires Fernandes, Instituto Politécnico de Bragança
Francisco José Pires Peito, Instituto Politécnico de Bragança
Ilda Marisa de Sá Reis, Instituto Politécnico de Bragança
João Paulo Pais de Almeida, Instituto Politécnico de Bragança
José Fernando Oliveira, Universidade do Porto, Faculdade de Engenharia
José Mário Escudeiro de Aguiar, Instituto Politécnico de Bragança
José Paulo Macedo Matias, Instituto Politécnico de Bragança
Maria de Fátima Silva Pacheco, Instituto Politécnico de Bragança
Maria Prudência Gonçalves Martins, Instituto Politécnico de Bragança
Paula Maria Pereira de Barros, Instituto Politécnico de Bragança

Índice

Editorial **8**

CSR of Portuguese Companies Listed on Euronext Lisbon: a Multivariate Analysis **9**
Sandra Afonso, Paula Fernandes, Ana Paula Monte

Production Planning of Perishable Food Products by Mixed-Integer Programming **16**
Pedro Amorim, Bernardo Almada-Lobo

A stochastic model for a multi-period multi-product closed loop supply chain **27**
Susana Baptista, Maria Isabel Gomes, Ana Paula Barbosa-Póvoa

Genetic Algorithms for the SearchCol++ framework: application to drivers' rostering **38**
Vítor Barbosa, Ana Respício, Filipe Alvelos

A Comparative Study of Two Optimization Clustering Techniques on Unemployment Data **48**
Elisa Barros, Alcina Nunes, Carlos Balsa

Otimização das visitas domiciliárias das equipas de profissionais de saúde nos Centros de Saúde **58**
Bruno Bastos, Tiago Heleno, António Trigo, Pedro Martins

Aproximação de cálculos iterativos por redes neuronais em sistemas de equações diferenciais ordinárias **67**
Ana S. R. Brásio, Andrey Romanenko, Natércia C. P. Fernandes

Computational comparison of algorithms for a generalization of the node-weighted Steiner tree and forest problems **77**
Raul Brás, J. Orestes Cerdeira

A multi-objective and multi-period approach for planning the delivery of long-term care services **88**
Teresa Cardoso, Mónica Oliveira, Ana Barbosa-Póvoa, Stefan Nickel

Design and planning of resilient closed-loop supply chains **98**
Sónia R. Cardoso, Ana Paula F. D. Barbosa-Póvoa, Susana Relvas

Benchmarking* dos Serviços dos Hospitais Portugueses: Uma Aplicação de *Data Envelopment Analysis **108**
Ricardo A. S. Castro, Conceição Silva Portela, Ana S. Camanho

Routing and assignment of clients of garden maintenance services **120**
J. Orestes Cerdeira, Manuel Cruz, Ana Moura

Discrete lot sizing and scheduling on parallel machines: description of a column generation approach **126**
António J.S.T. Duarte, J.M.V. Valério de Carvalho

A criação de horários no Ensino Superior Português: uma solução real para o problema real **135**
Pedro Fernandes, Carla Sofia Pereira, Armando Barbosa

Análise da eficiência das microempresas do setor do retalho no interior de Portugal: uma aplicação Data Envelopment Analysis **145**
António B. Fernandes, Maurício A. Vaz

Incorporação da resistência ao fogo na gestão florestal à escala da paisagem: uma aplicação à Mata Nacional de Leiria **154**
L. Ferreira, M. Constantino, J. G. Borges, J. Garcia-Gonzalo

A tool to manage tasks of R&D projects **162**
Joana Fialho, Pedro Godinho, João Paulo Costa

An optimisation model for the warehouse design and planning problem **172**
Carla A. S. Geraldes, Sameiro Carvalho, Guilherme Pereira

Assessing residential building sustainability in the operation phase **185**
I. M. Horta, A. S. Camanho, T. G. Dias

Optimization of a Humanoid Robot gait: multilocal optimization approach **194**
José Lima, Ana I. Pereira, Paulo Costa, José Gonçalves

Um método híbrido de pontos interiores e *branch-and-bound* aplicado ao modelo multi-objetivo de custo de colheita, coleta e aproveitamento de resíduos da cana-de-açúcar **201**
Camila de Lima, Antonio Roberto Balbo, Helenice de Oliveira Florentino Silva, Thiago Pedro Donadon Homem

Development of a Multicriteria Decision Aiding Model for monitoring and evaluating the performance of Health Care Units **212**
Diana F. Lopes

On numerical testing of the regularity of Semidefinite problems **223**
Eloísa Macedo

Dynamic location problem with uncertainty: a branch&bound approach **233**
Maria do Céu Marques, Joana Matos Dias

Dantzig-Wolfe reformulations for the forest harvest scheduling subject to maximum area restrictions **244**
Isabel Martins, Filipe Alvelos, Miguel Constantino, Ricardo Magalhães

Numerical Experiments with a Modified Regularization Scheme for Mathematical Programs with Complementarity Constraints **254**
Teófilo Miguel M. Melo, João Luís H. Matias, M. Teresa T. Monteiro

Extending the Resource-Task Network (RTN) for industrial scheduling problems **262**
Samuel Moniz, Ana P. Barbosa-Póvoa, Jorge P. Sousa

Investment Projects: Evaluation Tools and Methods **271**
Nuno Moutinho, Helena Mouta

Planeamento de rotas marítimas e estiva de contentores **275**
Jorge Oliveira, Ana Moura

Tactical and Operational Planning in Reverse Logistics Systems with Multiple Depots **286**
Tânia Ramos, Maria Isabel Gomes, Ana Paula Barbosa-Póvoa

Downstream oil products distribution planning **296**
Nuno Mota, Susana Relvas, Jorge Gonçalves

Distribution based artificial fish swarm in continuous global optimization **306**
Ana Maria A.C. Rocha, M. Fernanda P. Costa, Edite M.G.P. Fernandes

Recolha de Resíduos Sólidos Urbanos-otimização de rotas **313**
Ana Maria Rodrigues, José Soeiro Ferreira

Energy Efficient Routing for Telecommunication Networks with Multiperiod Traffic **323**
Dorabella Santos, Carlos Lopes, Amaro de Sousa, Filipe Alvelos

Método Previsor Corretor Primal Dual de Pontos Interiores Aplicado a um Problema de Despacho Econômico com Restrições Ambientais **333**
Amélia de Lorena Stanzani, Antonio Roberto Balbo

Cell-free Layer Measurements in Bifurcating Microchannels: a global approach **341**
B. Taboada, D. Bento, D. Pinho, A.I. Pereira, R. Lima

OR/MS EDUCATION: an overview of the 2003-2012 decade **347**
Ana Paula Teixeira, João Miranda

Framework for performance assessment of wind farms **356**
Clara Bento Vaz, Ângela Paula Ferreira

Multi-period and multi-product inventory management model with lateral transshipments **367**
Joaquim Jorge Vicente, Susana Relvas, Ana Barbosa-Póvoa

Scheduling batch processes using the RTN discrete time formulation: a case study **378**
Miguel Vieira, Tânia Pinto-Varela, Ana Paula Barbosa-Póvoa

The construction of composite indicators with undesirable outputs using DEA models **386**
Andreia Zanella, Ana S. Camanho, Teresa Galvão Dias

Índice de autores **396**

Agradecimentos **398**

Editorial

Na sequência do XIV Congresso da APDIO, na Caparica, onde pela primeira vez se publicou um livro de atas com artigos completos, e do XV Congresso da ADPIO, em Coimbra, onde novo livro foi editado e publicado, apresenta-se neste volume o Livro de Atas do XVI Congresso da APDIO, em Bragança, e que tem edição eletrónica.

Este livro contém 41 artigos que passaram por um rigoroso processo de avaliação por pares. À Comissão de Programa, que tão vivamente se envolveu neste processo de avaliação e revisão, e aos avaliadores que se empenharam de forma notável neste trabalho, o nosso muito obrigado.

Adicionalmente, teremos este ano pela primeira vez a publicação de um livro na série CIM Series in Mathematical Sciences, editado pela Springer Verlag, dedicado ao IO2013. Os autores de um conjunto selecionado de artigos publicados nestas atas serão convidados a fazer a submissão dos seus trabalhos, que serão alvo de um novo processo de avaliação e revisão, agora com avaliadores internacionais e gerido pelos editores convidados deste livro, segundo um processo normal de aceitação ou rejeição, com aprovação final pelos editores regulares da série de livros.

Uma última palavra é dirigida aos autores dos trabalhos publicados neste volume. Uma palavra de agradecimento pelo seu interesse e disponibilidade em publicarem nestas atas, uma palavra de felicitações pela qualidade dos trabalhos produzidos e uma palavra de incentivo, para que no próximo congresso possamos ter novamente um livro de atas com artigos de qualidade que sejam uma boa demonstração da vitalidade e da qualidade do trabalho cientifico da comunidade portuguesa de Investigação Operacional.

José Fernando Oliveira e Clara Bento Vaz

CSR of Portuguese Companies listed on Euronext Lisbon: a multivariate analysis

Sandra Afonso*, Paula Fernandes‡, Ana Paula Monte‡

*Instituto Politécnico de Bragança, Portugal.
E-mail:a24528@alunos.ipb.pt

‡ Instituto Politécnico de Bragança, Portugal & NECE[1] , Universidade da Beira Interior, Portugal.
E-mail:{pof,apmonte}@ipb.pt

Abstract

The purpose of this paper is to present a cluster analysis applied to group companies by their social performance and to compare the results. The results indicate that companies with better social performance are not the ones with better economic performance, and it suggests that the middle path might provide a good relation CSR-Economic performance, as a basis to sustainable development. The study focused on 19 quoted Portuguese companies and the analysis covered a period of five years, between 2005 and 2009. The results indicated that three clusters were classified in CSR Low (3 companies), CSR Medium (12 companies) and CSR High (4 companies). According to the cross validated classification based on discriminant analysis, the results reveal that 94.7% of the cases were classified correctly.

Keywords: Corporate Social Responsibility. Multivariate analysis. PSI-20 companies.

1 Introduction

In recent years, there have been many negative cases involving corporations and theirs leaders, cases of corruption involving fraudulent accounting, the growing gap between the salaries of top managers and their employees, abusive practices, marketing of products harmful to public health, violation of human rights and environmental standards [Lama and Muyzenberg, 2008].

In a world that is characterized by increasing economic and social asymmetries, it is necessary to find a way to promote stability that can be sustained in a virtuous cycle. "We have to choose between a global market driven only by calculation of short-term profit, and one which has a human face (...). Between a selfish free-for-all in which we ignore the fate of the losers, and a future in which the strong and successful accept their responsibilities, showing global vision and leadership" (Kofi Annan as cited in World Business Council for Sustainable Development [WBCSD, 2000]. CSR can be the way to a successful acceptance of responsibilities and to build a sustainable development.

Although the concept of Corporate Social Responsibility (CSR) has gained a prominent position in the general management literature, there is still uncertainty about how to adequately define the term [Bakker et al., 2005, Dahlsrud, 2006].

Bearing these facts in mind, the objective of this paper is to contribute to a better understanding of CSR in a Portuguese context. Therefore, it was intended to identify homogeneous groups of companies listed on EURONEXT belonging to the PSI-20, according to social responsibility, based on two multivariate analyses, namely cluster analysis and discriminant analysis.

The paper is organized as follows: after this introduction, section 2 will provide a synopsis of CSR definitions and measurement of CSR; section 3 will present the methodology used in this research. Section 4 presents the results and the discussion where cluster and discriminant analyses provide insights into the main determinants of CSR strategies and differences between companies. The final section presents the main conclusions and some future research direction.

2 Background

Social business concerns have existed for a long time, but the CSR debate began in the United Sates in 1953, when Howard Bowen argued that businessmen had the obligation to conduct business according

[1]R&D Centre funded by the Multiannual Funding Programme of R&D Centres of FCT - Portuguese Foundation for Science and Technology, Ministry of Education and Science.

to society's goals and values [Carroll and Shabana, 2010]. For Wartick and Cochran the CSR concept as suggest by Bowen, has two main premises: (i) the business exists to serve society, and its behaviour must be ruled by society's guidelines, in this context business assumes a social contract with society, which is the vehicle which brings business behaviour to conformity with society patterns; (ii) the business acts as a moral agency in society, and should act consistently according to society's values, many concepts followed, but until now without any consensual definition [Wartick and Cochran, 1985].

In Europe the CSR debate emerged later, in 1993, with the appeal made by Jacques Delors at the time when he was President of the European Community Commission (COM), to company's social intervention, which had a good acceptance [COM, 2001]. In 2001 the COM launched the "Green Paper" aiming to promote a European framework for CSR and considering that CSR could contribute to achieve the goal set at the European Council of Lisbon 2000: European economy becoming the world's most dynamic and competitive, based on knowledge and setting the basis for a sustainable development [COM, 2001].

The World Business Council for Sustainable Development, an organization created in 1995 that addresses the commitment to a sustainable development, also considers that sustainable development is based on "three fundamental and inseparable pillars: the generation of economic wealth, environmental improvement and social responsibility" [WBCSD, 2000] and that "CSR is an integral part of sustainable development" [WBCSD, 2000].

Although there isn't a consensual definition of sustainable development or a CSR definition, we can consider the tree dimensions (economic, social and environmental) as common two both concepts. But if CSR emerges as a way for an organization to assume it's responsibilities and contribute to a sustainable development, it can also raise the question of how can CSR positively affect economic performance in a way that can generate resources to continually invest in social and environmental demands.

According to Waddock and Graves (1997) high levels of financial performance can provide the resources necessary to invest in CSR practices. Also Ullmann argued that in periods of low economic return, companies have other priorities than investment in CSR, which may suggest that a satisfactory financial performance can have a positive influence in future commitment with social responsibility practices [Balabanis et al., 1998]. But CSR can also improve the economic performance, providing greater availability of resources. Orlitzky, Schmidt and Rynes (2003) suggest that social performance and financial and economic performance influence each other through a "virtuous cycle", since companies with good financial and economic performance invest more in social performance because they can do it, but at the same time the social performance also helps them increase financial success.

There are companies that invest in CSR despite the fact that this investment in the short term reduces the present value of their cash flows. According to Mackey and Mackey (2007), that can be explained by the conditions of supply and demand for CSR investments opportunities. When the demand is greater than supply even reducing the present value of cash-flows, the investments may generate economic value for companies [Mackey and Mackey, 2007].

In an attempt to relate CSR and economic and financial performance, many researches have achieved different and opposing results (e.g. [Waddock and Graves, 1997], [Griffin and Mahon, 1997], [Balabanis et al., 1998], [McWilliams and Siegel, 2000], [Poddi and Vergalli, 2009]). Waddock and Graves (1997) argue that difficulties in the measurement of social performance are the main reason for the uncertainty of the results obtained. Some of the measurement criteria often used are the content analysis of annual reports, expert evaluations, the index developed by the rating agency Kinder, Lydenderg, Domini (KLD), or the indexes of Fortune and Moskowitz, based on reputation (e.g. [Waddock and Graves, 1997], [Griffin and Mahon, 1997], [Balabanis et al., 1998], [Orlitzky et al., 2003], [McWilliams and Siegel, 2000], [Poddi and Vergalli, 2009], [Stanwick and Stanwick, 1998], [Goss and Roberts, 2011], [Harjoto and Jo, 2011]).

Szekeley and Knirsch (2005), analyzed the best metrics used by German companies to measure sustainable performance, and conclude that different methods were used, but many have adopted the guidelines of the Global Reporting Initiative (GRI). The authors considered it to be a good start, and a tool that needs to be improved, but is not enough for the structural changes that companies need to undertake internally to become more sustainable, and that requires a strong and visionary leadership [Szekeley and Knirsch, 2005].

In Portugal, although many companies were using CSR practices in a informal way, the CSR systematic practices only begun after the celebration of international agreements, and more specifically the European Lisbon Conference of 2000 [CECOA, 2004]. Researches done in the Portuguese context also refer to the lack of information to study Portuguese reality, and also lack of formulas for CSR implementation [Leite and Rebelo, 2010].

In this research, a different frame from previous researches was followed for the measurement of CSR. The measurement of social performance was based on the COM guidelines, and namely the two dimensions of CSR, internal and external dimensions [COM, 2001].

3 Methodology

This paper aims to identify homogeneous groups of companies listed on EURONEXT belonging to the PSI-20, according to social responsibility, based on two multivariate analyses: cluster analysis and discriminant analysis. 19 of 20 companies were selected, excluding EDP renewable. This company was excluded by the fact that its reports were published on the official website by EDP which comprised EDP Renewable information. So, the study focus on 19 companies (N = 19) and a period of 5 years (2005-2009) was taken into account. Therefore, the companies under study are Altri, BCP; BES, BPI, Brisa, Cimpor, EDP, Galp, Jeronimo Martins (JM), Mota-Engil (MOTEN), Portucel, Portugal Telecom (PT), REN, Semapa, Sonae Industria (SOIN), Sonae, SonaeCom (SOCOM), Teixeira Duarte (TEIXDU), Zon. Noted that these companies are obliged to report their accounts according to International Accounting Standards - International Financial Reporting Standards (IAS IFRS standard), since 2005. It was chosen a five years period of analysis because a long period can provide more reliable information about companies' commitment with CSR and also allows an evolution analysis of the adoption of CSR.

The variables chosen to measure the multiple dimensions of social performance (see Table 1), were based and adapted from Green Paper guidelines [COM, 2001], considering as well diverse literature on the subject, and also the GRI guidelines used by several Portuguese companies that report their social performance.

The analysis and measurement of social performance was done through content analyses from companies' sustainability and annual reports, available on companies' official websites. An index was built with 239 items considering the relevant aspects for each of the variables defined for measuring social performance. It was also considered the fact that most of Portuguese companies set their CSR goals according to the tree dimensions of the sustainable development: Economic, Environmental and Social. To each item was attributed a score: 0 (to a negative answer); 1 (to a positive answer); 0,5 (to an incomplete answer).

Table 1: Variables of Research.

Internal dimension	External dimension	Other variables
Responsible management	Local communities	
Human resources management	Stakeholders	
Health and safety at work	Human rights	CSR instruments
Environmental and natural resources management	Environmental and philanthropic global concerns	
Business ethics		

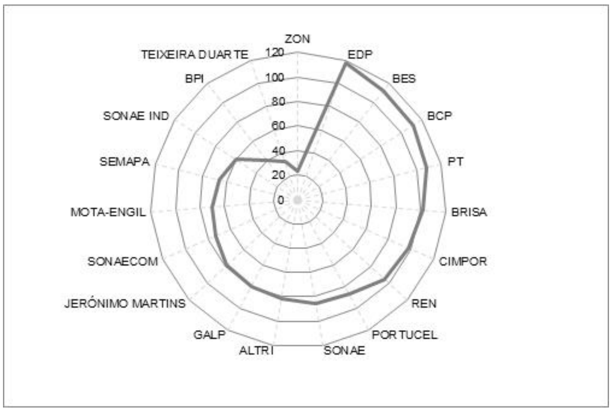


Figure 1: CSR Index for each company under analysis.

The final result, which is the total of all variables scores, was named CSR Index and allowed to positioning the companies according to their social performance (Fig. 1). Through the analysis of Fig. 1, it can be seen that EDP, BES, BCP, PT and BRISA obtained the highest values on the CSR Index, which may indicate that these companies should be more predisposed to adopt CSR practices.

A hierarchical cluster analysis was applied using the method proposed by Ward (1963), for being the one who has made the solution more consistent with other studies and applied to quantitative variables measured on a ratio scale. This analysis was produced in order to identify homogeneous groups of companies based on the variables chosen to measure social performance. A discriminant analysis was also applied to assess the adequacy of classification produced with a hierarchical cluster analysis.

4 Results and Discussion

The hierarchical cluster analysis, using the Ward method and the squared euclidean distance, produced the dendogram in Fig. 2. Trough the analysis of dendogram we can clearly see two main clusters, although the division into three clusters presents a more homogeneous distribution of cases. To decide on the optimal number of clusters the r-square criterium was used and the graph of the relativized distance between clusters. A solution of tree clusters was chosen, explaining 62% of the total variance.

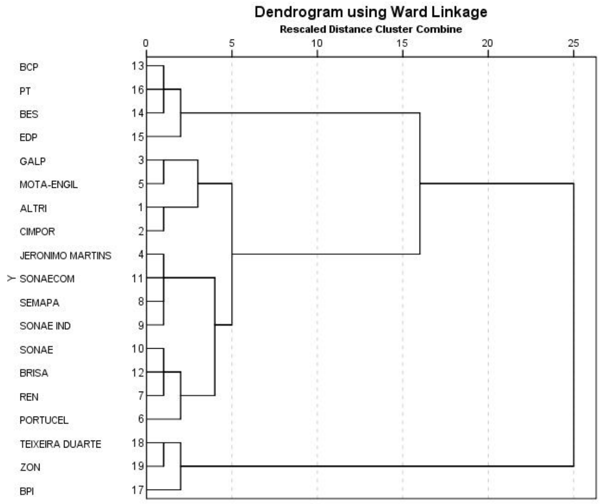


Figure 2: Dendogram using the Ward linkage method.

Each cluster was named according to the social performance of the companies that composed it: Cluster 1 - CSR Medium; Cluster 2 - CSR High; Cluster 3 - CSR Low (see Table 2). The practices uses by companies in these areas were important to define their social performance.

Table 2: Cluster Composition.

Cluster	Companies	
Clusters 1: CSR Medium	BRISA	GALP
	CIMPOR	JERONIMO MARTINS
	REN	SONAECOM
	PORTUCEL	MOTA-ENGIL
	SONAE	SEMAPA
	ALTRI	SONAE IND
Clusters 2: CSR High	BES	
	BCP	
	EDP	
	PT	
Clusters 3: CSR Low	BPI	
	TEIXEIRA DUARTE	
	ZON	

A Discriminant Analysis was conducted to predict and classify whether the companies have a low, medium or high CSR. It was also used to confirm the results produced by hierarchical cluster analysis. According to the results produced by Discriminant Analysis it was possible to observe significant mean differences for all predictors - independent variables on the dependent variable. Box's M test indicated that the assumption of equality of covariance matrices wasn't violated. This can be concluded by the insignificance of the differences observed (Box's M is 9.243 with $F = 1.032$ and $p - value = 0,404$). The discriminate function revealed a significant association between groups and all predictors. Regarding the results, the 1st function with an eigenvalue of 4.71 corresponds to 81% of variance explained in terms of differences between clusters, thus explaining the greater proportion of the variance, and 2nd function 19%. The correlation between the canonical functions and clusters enables us to observe a greater correlation to the 1st function 0,91. Wilks' lambda indicates the significance of the discriminant function and a highly significant function ($p - value < 0.001$). Closer analysis of the structure matrix (Table 3) revealed the significant predictors, namely for the 1st discriminant function of Human resources management (0.891), Environmental and natural resources management (0.778), Local communities (0.621) and for the 2nd discriminant function of CSR instruments (0.495) and Health and safety at work (0.468).

Table 3: Structure Matrix.

	Discriminant Function	
	1	2
Human resources management	0,891	-0,454
Environmental and natural resources management	0,778	0,628
Local communities	0,621	0,271
Responsible management	0,440	0,091
Environmental and philanthropic global concerns	0,427	0,118
Stakeholders	0,314	0,257
Business ethics	-0,171	-0,149
Human rights	0,112	-0,105
CSR instruments	0,350	0,495
Health and safety at work	0,426	0,468

The cross validated classification results reveal that 94.7% of the cases were classified correctly into 'Cluster 1 - CSR Medium', 'Cluster 2 - CSR High' and 'Cluster 3 - CSR Low'.

Fig. 3 shows the clusters classification and it was possible to corroborate the results obtained in the cluster analysis. It was also observed that the centroids of each cluster are quite distant from each other allowing the separation of clusters.

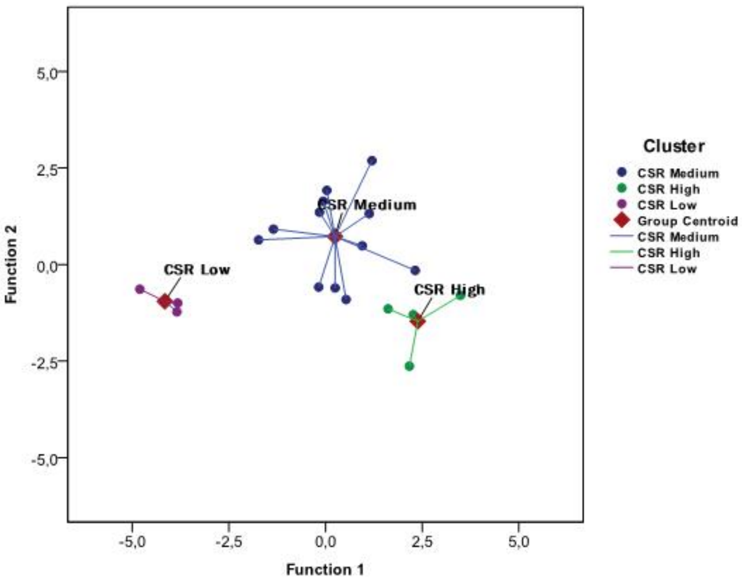


Figure 3: Canonical Discriminant Functions.

Briefly, it can be concluded that the three clusters of companies differ with regards to their dominant motives and culture for pursuing CSR strategies.

5 Conclusions and further research

The objectives of this research were to identify homogeneous groups of companies listed on EURONEXT belonging to the PSI-20, according to social responsibility, and based on two multivariate analyses, hierarchical cluster analysis and discriminant analysis.

According to the outputs produced based on both multivariate analysis it was possible to identified three homogenous groups of companies: (1) CSR Low cluster consisting of three companies, (2) CSR Medium cluster consisting of 12 companies, and (3) CSR High group consisting of 4 companies. It also indicated that the companies in question, which are companies of reference in the Portuguese context, mostly have an average performance in relation to CSR, although some companies present a high level. So it can be concluded that most denote a growing sensitivity to CSR practice.

This research adds an important contribution to the definition of Portuguese companies listed on EURONEXT belonging to the PSI-20, and it also provides a new measurement model that takes into account ten dimensions, a broad view of the multiplicity of CSR and is based on the guidelines of the European Commission. For future research it is suggested that similar methodology is used in a comparative analysis by sectors, taking into consideration the national or international context, with distinct legal and institutional frameworks.

References

[Bakker et al., 2005] Bakker, F., Groenewegen, P. and Hond, F. (2005), A bibliometric analysis of 30 years of research and theory on corporate social responsibility and corporate social performance, *Business & Society*, 44(3): 283-317.

[Balabanis et al., 1998] Balabanis, G., Phillips, H. and Lyall, J. (1998). Corporate social responsibility and economic performance in the top British companies: are they linked?, *European Business Review*, 98(1): 25-44.

[Carroll and Shabana, 2010] Carroll, A. and Shabana, K. (2010). The business case for corporate social responsibility: A review of concepts, research and practice, *International Journal of Management Reviews*, 1: 85-105.

[CECOA, 2004] CECOA (2004). Centro de Formação Profissional para o Comércio e Afins, Responsabilidade social das empresas – Estado de arte em Portugal. Retrieved March 1, 2013, from `http://www.sairdacasca.com/recursos/docs/CECOA.pdf`.

[COM, 2001] COM (2001). Commission of the European Community, Green Paper: Promoting a European framework for Corporate Social Responsibility, COM 366, Brussels, 2001. Retrieved October 15, 2010, from `http://eur-lex.europa.eu/LexUriServ/site/en/com/2001/com2001_0366en01.pdf`.

[COM, 2002] COM (2002). Commission of the European Community, Corporate Social Responsibility: A business contribution to Sustainable Development, COM 347, Brussels, 2002. Retrieved March 1, 2013, from `http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2002:0347:FIN:en:PDF`.

[Dahlsrud, 2006] Dahlsrud, A. (2006), How corporate social responsibility is defined: An analysis of 37 definitions, *Corporate Social Responsibility and Environmental Management*, 15(1): 1-13.

[Goss and Roberts, 2011] Goss, A. and Roberts, G. (2011). The impact of Corporate Social Responsibility on the cost of bank loans. *Journal of Banking & Finance*, 35(7): 1794-1810.

[Griffin and Mahon, 1997] Griffin, J. and Mahon, J. (1997). The corporate social performance and corporate financial performance debate, *Business and Society*, 36(1): 5-31.

[Harjoto and Jo, 2011] Harjoto, M. and Jo, H. (2011). Corporate governance and CSR nexus. *Journal of Business Ethics*, 100: 45-67.

[Lama and Muyzenberg, 2008] Lama, D. and Muyzenberg, L. (2008). O caminho para a liderança, *Transl. S. Oliveira*. Ed. Alfragide: Publicações Dom Quixote.

[Leite and Rebelo, 2010] Leite, C. and Rebelo, T. (2010). Explorando, caracterizando e promovendo a responsabilidade social das empresas em Portugal, in *Proc. of the Eighth National Symposium on Research in Psychology*, Universidade do Minho, Portugal.

[Mackey and Mackey, 2007] Mackey, A. and Mackey, T. (2007). Corporate social responsibility and firms performance: investor preferences and corporate strategies, *Academy of Management Review*, 32 (3): 817-835.

[McWilliams and Siegel, 2000] McWilliams, A. and Siegel, D. (2000). Corporate social responsibility and financial performance: correlation or misspecification, *Strategic Management Journal*, 21: 603-609.

[Orlitzky et al., 2003] Orlitzky, M., Schmidt, F. and Rynes, S. (2003). Corporate social and financial performance: A meta-analysis, *Organization Studies*, 24 (3): 403-441.

[Poddi and Vergalli, 2009] Poddi, L. and Vergalli, S. (2009). Does corporate social responsibility affect the performance of firms?, Fondazione Enri Enrico Mattei in 14th Coalition Theory Network Workshop, Masstricht, Netherlands, Retrieved March 1, 2013, from `http://ageconsearch.umn.edu/bitstream/52531/2/52-09.pdf`.

[Masstricht, Netherlands, 2013] Masstricht, Netherlands, Retrieved March 1, 2013, from `http://ageconsearch.umn.edu/bitstream/52531/2/52-09.pdf`.

[Stanwick and Stanwick, 1998] Stanwick, P. and Stanwick, S. (1998). The relationship between corporate social performance, and organizational size, financial performance and environmental performance: An empirical examination, *Journal of Business Ethics*, 17: 195-204.

[Szekeley and Knirsch, 2005] Szekeley, F. and Knirsch, M. (2005). Responsible leadership and corporate social responsibility metrics for sustainable performance, *European Management Journal*, 23 (6): 628-647.

[Waddock and Graves, 1997] Waddock, S. and Graves, S. (1997). The corporate social performance - Financial performance link, *Strategic Management Journal*, 18 (4): 303-319.

[Ward, 1963] Ward, J. (1963). Hierarchical grouping to optimize and objective function. *Journal of the American Statistical Association*, 58: 236-244.

[Wartick and Cochran, 1985] Wartick, S., and Cochran, P. (1985). The evolution of the corporate social performance model, *Academy of Management Review*, 10 (4): 758-769.

[WBCSD, 2000] WBCSD (2000). World Business Council for Sustainable Development, Corporate Social Responsibility: making good business sense, 2000. Retrieved March 1, 2013, from `http://www.wbcsd.org/DocRoot/IunSPdIKvmYH5HjbN4XC/csr2000.pdf`.

Production Planning of Perishable Food Products by Mixed-Integer Programming

Pedro Amorim*, Bernardo Almada-Lobo*

* *INESC TEC, Faculdade de Engenharia, Universidade do Porto, Portugal*
E-mail:`amorim.pedro@fe.up.pt`
E-mail(second author):`almada.lobo@fe.up.pt`

Abstract

In this paper, the main complexities related to the modelling of production planning problems of food products are addressed. We start with a base model and build a roadmap on how to incorporate key features of food production planning. The different "ingredients" are organized around the model components to be extended: constraints, objective functions and parameters. We cover issues such as expiry dates, customers' behaviour, discarding costs, value of freshness and age-dependent demand. To understand the impact of these "ingredients", we solve an illustrative example for each model and analyse the changes on the solution structure of the production plan.

Keywords: Production Planning, Food Industry, Perishability, Mathematical Programming.

1 Introduction

The supply chain planning of food products is ruled by the dynamic nature of its products. Throughout the planning horizon, the characteristics of these products go through significant changes. The root cause for these changes may be related to, for example, the physical nature of the products or the value that the customer lends to them. Without acknowledging the perishable nature of food products, one may incur in avoidable spoilage costs (for example, in the case of meat products) or, on the other hand, sell the product before it is close enough to its best state (for example, in the case of cheese products). In this paper we focus on perishable food products that start losing their properties after being produced.

[Fleischmann et al., 2008] define planning as the activity that supports decision-making by identifying the potential alternatives and making the best decisions according to the objective of the planners. Let us look into the challenges of engaging in this production planning activity in the context of food products.

In order to identify the alternatives it is important to understand the decisions that the decision maker wants to make. It is common to organize the supply chain planning according to two dimensions: the supply chain process and the hierarchical level. The scope of this paper is the production supply chain process and we deal with problems arising at the operational decision level. Therefore, we will address food production planning problems that have to decide about the size of the lots to be produced and about the schedule of these production lots. In this problem, we usually determine the size of lots to be produced while trading off the changeover and stock holding costs. In food production, expiry dates may enforce constraints related to the upper bounds on lot-sizes and consequently the need of scheduling more often a given family of products (increasing the difficulty of sequencing). Expiry dates relate to the concept of perishability that is defined by [Amorim et al., 2011] as: "A good, which can be a raw material, an intermediate product or a final one, is called 'perishable' if during the considered planning period at least one of the following conditions takes place: (1) its physical status worsens noticeably (e.g. by spoilage, decay or depletion), and/or (2) its value decreases in the perception of a(n internal or external) customer, and/or (3) there is a danger of a future reduced functionality in some authority's opinion.". In this paper, we will consider goods that suffer a physical deterioration, for which customers' attribute a decreasing value and for which authorities usually limit the commercialization period.

The second part of [Fleischmann et al., 2008] definition of planning relates to the objectives of the planners. The literature in production planning tackles most of the problems with traditional single objective models. The goal is usually related either to an operational measure, such as makespan, or to some monetary measure, such as cost or profit. In this paper we extend these objectives by including factors related to the food industry, such as spoilage costs. Moreover, a multi-objective approach is described in order to account for the customer willingness for fresher products. Therefore, we think that besides avoiding the products spoilage, there may exist a substantial intangible gain from delivering fresher products to customers. Such considerations are closely related to the consumer purchasing behaviour of perishable goods that should be addressed by any planner in a (food) company with a supply chain orientation.

In the remainder of the paper, we present how a traditional base model dealing with the production planning of food products has to be changed in order to accommodate the characteristics of the products it has to deal with. Therefore, we start by presenting a base model in Section 2. In Section 3, we understand how the constraints have to be extended to incorporate key aspects, such as the fact that products have a shelf-life or that customers pick up the fresher available products. Section 4 analyses the possible changes in the objective function: discarding costs of perished goods and valuing in a different objective function the freshness. Finally, Section 5 brings the possibility of having more information on key parameters – dependency between price and age and between demand and age. The "ingredients" presented throughout this paper can be mixed together in various ways to form the "recipe" suitable for the production environment. In order to help understanding the implications of these "ingredients" in the solution structure, all models are solved for an illustrative example in Section 6. Finally, in Section 7 the conclusions are presented.

2 Base model

We start by presenting a base model for production planning in food industries. This model focuses on the packaging stage and has no considerations about the perishable nature of the products.

One important concept in the fast moving consumer food goods is the recipe. Usually, products belong to a certain recipe that requires a major setup and the products within the recipes are sequenced a priori and just need a minor setup. We use an adaptation of the block planning formulation [Günther et al., 2006] that was designed for similar production environments to that of the food production. To make it clearer, a block corresponds to a recipe and within and between recipes the sequence of products is set *a priori*. Therefore, the only decision to be made for each block / product, besides the sizing the lots, is to produce it or not. This modelling approach increases the application potential of decision support systems in production planning, because decision makers are comfortable with the definition of the recipes and, simultaneously, the scheduling complexity is fairly reduced. In Figure 1 a production schedule with two blocks, A and B, is depicted. Notice that before producing products of a given recipe a major setup is necessary. Afterwards, all products within the same recipe are produced after preparing a minor setup.

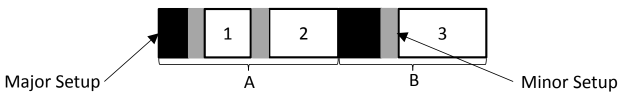


Figure 1: Adapted block planning concept.

For the base model we also use an adaptation of the simple plant location (SPL) reformulation to model inventory and demand fulfilment decision variables. In the traditional SPL reformulation, it is known for which period the production of a given period refers to. In a food production planning context we are more interested in the actual age of the product. This will be rather helpful in limiting the usage of stock based on the shelf-life of the products. Moreover, it can be also used to keep track of the freshness of the products delivered to the clients. Figure 2 shows how traditional decision variables are transformed through the adapted SPL reformulation (the meaning of these variables and parameters is given below).

The deployment of these two adpated concepts (block planning and simple plant location) results in a base model flexible enough to cope with the exigencies of production planning in food industries.

Let us now move to a formal description of the problem. Consider a set of products $k = 1, ..., K$ that are produced based on a certain recipe / block $j = 1, ..., N$. There is only one recipe to produce each product and, therefore, a product is assigned to one block only. Hence, for each block j there is a set $\mathcal{KJ}$ of products k related to it. Blocks are to be scheduled on $l = 1, ..., L$ parallel production lines over a finite planning horizon consisting of periods $t = 1, ..., T$ with a given length. This length is related to the company practice of measuring external elements, such as demand (thus, periods correspond to days, weeks or months in most of the cases). According to the block structure, all scheduling decisions are made *a priori* for both recipes and products. Hence, the production sequence is determined beforehand, minimizing the setup times and costs according to the planner expertise [Günther et al., 2006].

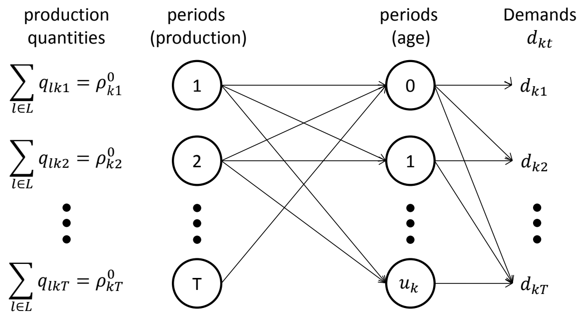


Figure 2: Schematic representation of the adaptation of the simple plant location reformulation with an emphasis on the inventory age.

Consider the following indices, parameters, and decision variables that are used hereafter.

Indices

$l \in \mathcal{L}$	parallel production lines
$j \in \mathcal{N}$	blocks
$k \in \mathcal{KJ}$	products
$t \in \mathcal{T}$	periods
$a \in \mathcal{A} = \{a \in \mathbb{Z}_0^+ \| a \leq t-1\}$	age (in periods)

Parameters

C_{lt}	capacity (time) of production line l available in period t
e_{lk}	capacity consumption (time) needed to produce one unit of product k on line l
c_{lk}	production costs of product k (per unit) on line l
u_k	shelf-life of product k just after production (time)
p_k	price of product k
h_k	inventory carrying cost of product k
m_{lj}	minimum lot size (units) of block j on line l
$\bar{s}_{lj}(\bar{\tau}_{lj})$	setup cost (time) of a changeover to block j on line l
$\underline{s}_{lk}(\underline{\tau}_{lk})$	setup cost (time) of a changeover to product k on line l
d_{kt}	demand for product k in period t (units)

Decision Variables

$\rho^a_{kt} \geq 0$	initial inventory of product k with age a available at period t
$\psi^a_{kt} \geq 0$	fraction of the maximum demand for product k delivered with age a at period t
$q_{lkt} \geq 0$	quantity of product k produced in period t on line l
$p_{lkt} \in \{0,1\}$	equals 1, if line l is set up for product k in period t (0 otherwise)
$y_{ljt} \in \{0,1\}$	equals 1, if line l is set up for block j in period t (0 otherwise)

Figure 3 presents a schematic relation between the main decision variables.

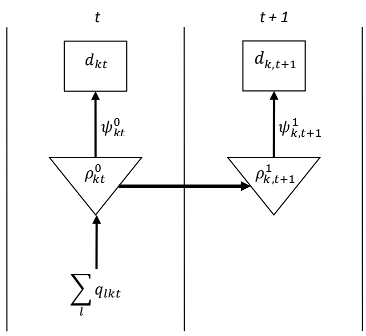


Figure 3: Schematic representation of the relation between decision variables.

The base production planning model of food products (B-PP-FP) reads:

B-PP-FP

$$\max \sum_{k,t,a} p_k\, d_{kt}\, \psi^a_{kt} - \sum_{l,j,t} \bar{s}_{lj}\, y_{ljt} - \sum_{l,k,t} (\underline{s}_{lk}\, p_{lkt} + c_{lk}\, q_{lkt}) - \sum_{k,t,a} h_k\, (\rho^a_{kt} - d_{kt}\psi^a_{kt}) \tag{1}$$

subject to:

$$\sum_a \psi^a_{kt} \leq 1 \quad \forall k \in \mathcal{K}, t \in \mathcal{T} \tag{2}$$

$$\rho^a_{kt} = \rho^{a-1}_{k,t-1} - d_{k,t-1}\psi^{a-1}_{k,t-1} \quad \forall k \in \mathcal{K}, t \in \mathcal{T}, a \in \mathcal{A} \setminus \{0\} \tag{3}$$

$$\sum_l q_{lkt} = \rho^0_{kt} \quad \forall k \in \mathcal{K}, t \in \mathcal{T} \tag{4}$$

$$p_{lkt} \leq y_{ljt} \quad \forall l \in \mathcal{L}, j \in \mathcal{N}, k \in \mathcal{KJ}, t \in \mathcal{T} \tag{5}$$

$$q_{lkt} \leq \frac{C_{lt}}{e_{lk}} p_{lkt} \quad \forall l \in \mathcal{L}, k \in \mathcal{K}, t \in \mathcal{T} \tag{6}$$

$$\sum_j \bar{\tau}_{lj}\, y_{ljt} + \sum_k (\underline{\tau}_{lk}\, p_{lkt} + e_{lk}\, q_{lkt}) \leq C_{lt} \quad \forall l \in \mathcal{L}, t \in \mathcal{T} \tag{7}$$

$$\sum_{k \in \mathcal{KJ}} q_{lkt} \geq m_{lj} y_{ljt} \quad \forall l \in \mathcal{L}, j \in \mathcal{N}, t \in \mathcal{T} \tag{8}$$

$$\psi^a_{kt},\, \rho^a_{kt},\, q_{lkt} \geq 0;\, p_{lkt},\, y_{ljt} \in \{0, 1\} \tag{9}$$

The objective function (1) maximizes the profit of the producer over the planning horizon. Therefore, revenue is subtracted by setup costs of blocks, setup costs of products, variable production costs and inventory costs. Note that the setup structure considers major and minor setup for the first product to be produced in a given block. For example, in the yoghurt production when changing from one kind of yoghurt to the other a major setup might correspond to cleansing the lines and linking the new yoghurt tank, while the minor setup may correspond to setting up the machine to fill the yoghurt in the new packages. These two operations can not be done in parallel.

Equations (2) forbid the sum of all sold products of different ages to exceed the demand. Equations (3) establish the inventory balance constraints, ageing the stock throughout the horizon. Constraints (4) link the production variables to the inventory ones, setting all production in a given period to the initial stock with age 0. Equations (5) and (6) ensure that a product can only be produced if both the correspondent block and product are set up, respectively. Limited capacity in the lines is to be reduced by setup times between blocks, setup times between products and also by the time consumed producing products (6). Constraints (7) introduce minimum lot-sizes for each block.

Final constraints (9) define the domain of the decision variables.

3 Extending the constraints

Two realistic factors may impact the production plans: the fact that inventory that is beyond the expiry date can no longer be sold (product-related), and the fact that customers in face of inventories with different shelf-lives, choose products with the farthest expiry date (customer-related). These issues are addressed by limiting the feasibility domain as follows.

3.1 Inventory expiry

In order to make sure that no expired product is used to satisfy demand it suffices to redefine the constraints dealing with these variables.

The production planning model of food products with inventory expiry constraints (IE-PP-FP) reads:

IE-PP-FP

$$\max \sum_{k,t,a} p_k\, d_{kt}\, \psi^a_{kt} - \sum_{l,j,t} \bar{s}_{lj}\, y_{ljt} - \sum_{l,k,t} (\underline{s}_{lk}\, p_{lkt} + c_{lk}\, q_{lkt}) - \sum_{k,t,a} h_k\, (\rho^a_{kt} - d_{kt}\psi^a_{kt})$$

subject to:

$$\sum_{a \le u_k - 1} \psi^a_{kt} \le 1 \quad \forall k \in \mathcal{K}, t \in \mathcal{T} \tag{10}$$

$$\rho^a_{kt} = \rho^{a-1}_{k,t-1} - d_{k,t-1}\psi^{a-1}_{k,t-1} \quad \forall k \in \mathcal{K}, t \in \mathcal{T}, a \in \mathcal{A} \setminus \{0\} : a \le u_k \tag{11}$$

$$(4)\text{-}(8)$$

$$\psi^a_{kt},\ \rho^a_{kt},\ q_{lkt} \ge 0;\ p_{lkt},\ y_{ljt} \in \{0,1\}$$

The remaining constraints are exactly the same as in the base model of Section 2.

3.2 Consumer behaviour

In a context where the production process is tightened to the downstream supply chain processes satisfying final customers demand it may be important to better incorporate the instinctive behaviour of consumers. Regarding food products, usually a last-expired-first-out policy is put in practice by customers. This behaviour may guide production plans towards a more just-in-time philosophy.

It is necessary to add a new decision variable θ^a_{kt} in order to model this behaviour that equals 1, if inventory of product k with age a is used to satisfy demand in period t (0 otherwise).

The production planning model of food products incorporating consumer behaviour (CB-PP-FP) reads:

CB-PP-FP

$$\max \sum_{k,t,a} p_k\, d_{kt}\, \psi^a_{kt} - \sum_{l,j,t} \bar{s}_{lj}\, y_{ljt} - \sum_{l,k,t} (\underline{s}_{lk}\, p_{lkt} + c_{lk}\, q_{lkt}) - \sum_{k,t,a} h_k\, (\rho^a_{kt} - d_{kt}\psi^a_{kt})$$

subject to:

$$\psi^a_{kt} \le \theta^a_{kt} \quad \forall k \in \mathcal{K}, t \in \mathcal{T}, a \in \mathcal{A} : a \le u_k - 1 \tag{12}$$

$$\rho^{a-1}_{kt} - d_{kt}\psi^{a-1}_{kt} \le M(1 - \theta^a_{kt}) \quad \forall k \in \mathcal{K}, t \in \mathcal{T}, a \in \mathcal{A} \setminus \{0\} : a \le u_k - 1 \tag{13}$$

$$(10)\text{-}(11),\ (4)\text{-}(8)$$

$$\psi^a_{kt},\ \rho^a_{kt},\ q_{lkt} \ge 0;\ \theta^a_{kt},\ p_{lkt},\ y_{ljt} \in \{0,1\} \tag{14}$$

In the previous models, it is assumed that the seller is able to assign optimal inventory quantities of different ages to customers in order to maximize profit. With constraints (12) and (13) this situation is avoided by mimicking the more instinctive consumer purchasing behaviour of perishable products that will drive customers to pick up products with the highest degree of freshness. Thus, constraints (12) turn the value of θ^a_{kt} to 1, whenever inventory of a given product k in period t with age a is used to satisfy demand. The value of this variable θ^a_{kt} is used in equations (13) to ensure that a fresher inventory can only be used after depleting the older inventory. Note that parameter M denotes a big number.

4 Extending the objective function

The most expedite approach to grasp the perishability phenomena is to penalize the spoiled products with a discard cost in the objective function. This penalty cost makes sense if we acknowledge that products have a limited shelf-life. Another approach derives from the awareness of the customers' willingness for fresher products while, simultaneously, the level of information regarding the detailed values of this willingness is low. In this case, a new objective function is added to the one maximizing profit. The new objective function aims at maximizing the freshness of the products delivered.

4.1 Discarding costs

This objective function extends the traditional production planning objective function by incorporating perishability related costs. Remark that the cost of spoiled products are valued through the opportunity cost $\bar{p}_k$. This opportunity cost corresponds to the revenue yielded by the best alternative that could have been produced and sold instead of producing product k that got spoiled. However, it may also be regarded, in a more tangible manner, as a disposal cost for each unit of perished inventory that has to be properly discarded.

The production planning model of food products including discarding costs (DC-PP-FP) reads:

DC-PP-FP

$$\max \sum_{k,t,a} p_k\, d_{kt}\, \psi^a_{kt} - \sum_{l,j,t} \bar{s}_{lj}\, y_{ljt} - \sum_{l,k,t} (\underline{s}_{lk}\, p_{lkt} + c_{lk}\, q_{lkt}) - \sum_{k,t,a} h_k\, (\rho^a_{kt} - d_{kt}\psi^a_{kt}) - \sum_{k,t,a \geq u_k} \bar{p}_k \rho^a_{kt} \tag{15}$$

subject to:

(2)-(8)

$$\psi^a_{kt},\ \rho^a_{kt},\ q_{lkt} \geq 0;\ p_{lkt},\ y_{ljt} \in \{0, 1\}$$

The only difference to the model presented in Section 2 is reflected by the cost of spoilage tracked by the last term of (15).

4.2 Measuring freshness

In this model, the economic tangible profit is separated from the customer intangible value of having fresher products in two distinct objective functions. The first objective continues to be the maximization of profit and the second one maximizes the average freshness of delivered products [Amorim et al., 2011b]. These two objectives are certainly conflicting since achieving a higher freshness of products delivered has to be done at the expense of higher production costs, for example, through the splitting of some production lots. Therefore, we acknowledge the complete different nature of the two complementary objectives and the difficulty to attribute different monetary values to different degrees of freshness. As a result, the decision maker will be offered a trade-off between freshness of delivered products and total profit. This trade-off can be represented by a set of solutions which do not dominate one another regarding both objectives (non-dominated or Pareto optimal front). We need to define the following additional parameter $[d_{kt}]$ that is the number of non-zero occurrences in the demand matrix.

The model that accounts for a measure of freshness (MF-PP-FP) reads:

MF-PP-FP

$$\max \sum_{k,t,a} p_k\, d_{kt}\, \psi^a_{kt} - \sum_{l,j,t} \bar{s}_{lj}\, y_{ljt} - \sum_{l,k,t} (\underline{s}_{lk}\, p_{lkt} + c_{lk}\, q_{lkt}) \tag{16}$$

$$\max \frac{1}{[d_{kt}]} \sum_{k,t,a} \frac{u_k - a}{u_k} \psi^a_{kt} \tag{17}$$

subject to:

(2)-(8)

$$\psi_{kt}^{a},\, \rho_{kt}^{a},\, q_{lkt} \geq 0;\, p_{lkt},\, y_{ljt} \in \{0,1\}$$

The first objective function (16) maximizes profit very similarly to the base model. In the second objective (17) the mean freshness of products to be delivered is maximized. The number of periods before spoilage is estimated by $u_k - a$. The cardinality of the non-zero demand occurrences is used to normalize this objective function between 0 and 1. This cardinality, for a given input set data, is constant and easily computed.

This approach for modelling the production planning for food products has an interesting aspect to consider regarding inventory costs. When maximizing freshness in the second objective we are already trying to minimize stocks since we try to produce as late as possible. Hence, if we had also included inventory costs in the first objective we would be somehow duplicating the inventory carrying cost effect and, therefore, (16) and (17) would be correlated.

5 Extending the parameters

The last form of differentiating the base model of food production planning is by changing or detailing the input parameters, namely: price and demand. The key reasoning is that with more accurate information and more transparency across the supply chain partners, it would be possible to discriminate either price or demand based on the actual age of the products.

5.1 Value of freshness

In this model it is assumed that, for example, a retailer will be willing to pay a different price for products with different standards of freshness. Therefore, the price parameter is extended to $\hat{p}_k^a$, price of product k paid when product has age a.

The production planning model of food products with different freshness values (VF-PP-FP) reads:

VF-PP-FP

$$\max \sum_{k,t,a} \hat{p}_k^a\, d_{kt}\, \psi_{kt}^a - \sum_{l,j,t} \bar{s}_{lj}\, y_{ljt} - \sum_{l,k,t} (\underline{s}_{lk}\, p_{lkt} + c_{lk}\, q_{lkt}) - \sum_{k,t,a} h_k\, (\rho_{kt}^a - d_{kt}\psi_{kt}^a) \tag{18}$$

subject to:

(2)-(8)

$$\psi_{kt}^{a},\, \rho_{kt}^{a},\, q_{lkt} \geq 0;\, p_{lkt},\, y_{ljt} \in \{0,1\}$$

The only difference to the model presented in Section 2 is reflected in the dependency of the revenue to the age of the delivered products.

5.2 Demand parameter

In this model we assume that according to the information about the customer purchasing behaviour, it is possible to determine a parameter $\hat{d}_{kt}^a$ for the demand for product k with age a in period t. For understanding how this parameter may be generated the readers are referred to [Amorim et al., 2013].

The production planning model of food products with an extended demand parameter (DP-PP-FP) reads:

DP-PP-FP

$$\max \sum_{k,t,a} p_k\, \hat{d}_{kt}^a\, \psi_{kt}^a - \sum_{l,j,t} \bar{s}_{lj}\, y_{ljt} - \sum_{l,k,t} (\underline{s}_{lk}\, p_{lkt} + c_{lk}\, q_{lkt}) - \sum_{k,t,a} h_k\, (\rho_{kt}^a - \hat{d}_{kt}^a\psi_{kt}^a) \tag{19}$$

subject to:

$$\hat{d}_{kt}^0 \psi_{kt}^a \leq \hat{d}_{kt}^a \quad \forall k \in \mathcal{K}, t \in \mathcal{T}, a \in \mathcal{A} \tag{20}$$

$$\rho^a_{kt} = \rho^{a-1}_{k,t-1} - \hat{d}^0_{k,t-1}\psi^{a-1}_{k,t-1} \quad \forall k \in \mathcal{K}, t \in \mathcal{T}, a \in \mathcal{A} \setminus \{0\} \tag{21}$$

(2), (4)-(8)

$$\psi^a_{kt},\ \rho^a_{kt},\ q_{lkt} \geq 0;\ p_{lkt},\ y_{ljt} \in \{0, 1\}$$

The formulation incorporating different demand levels according to the age of the product is very similar to the base model presented in Section 2, but in this model the demand parameter is replaced by its extensive form. Moreover, equation (20) do not allow the quantity of sold products of a given age to be above the demand curve derived for the respective product.

6 Illustrative Example

The setting for the illustrative example consists in a production line ($L = 1$) that has to produce 2 blocks ($N = 2$), each with two products ($K = 4$). For all products/blocks $e_{lk} = 1$, $m_{lj} = 3$ and $\underline{s}_{lj} = \underline{\tau}_{lj} = 1$. The setup costs and times between blocks ($\bar{s}_{lj}$ and $\overline{\tau}_{lj}$) are (5,5) and (1,2), respectively (Block 1, Block 2). The considered planning horizon has 4 periods ($T = 4$) and the capacity C_{lt} equals 35 for all periods. The remaining parameters are given in Table 1.

						d_{kt}			
Block	Product	u_k	p_k	c_{lk}	h_k	1	2	3	4
1	1	1	3	1	0.2	5	0	5	5
1	2	1	3	1	0.2	0	10	10	5
2	3	2	4	2	0.1	10	5	0	10
2	4	3	4	2	0.1	5	15	10	5

Table 1: Remaining parameters for the illustrative example.

We further consider, for the model of Section 4.1, that discarding costs $\bar{p}_k$ equal to p_k. In order to obtain one solution for the multi-objective model presented in Section 4.2, a weight of 200 was given to the freshness objective. For the case in which a decreasing value is considered for the price paid (Section 5.1), we consider that for products with an age higher than 0, $\hat{p}^a_k = 1$. Finally, for the last model (Section 5.2), all products suffer from a 50% rate of decrease in the demand for each period older ($\hat{d}^{a+1}_{kt} = 0.5\hat{d}^a_{kt}$).

6.1 Results and discussion

Table 2 shows the results for the key decision variables under analysis q_{lkt}, ρ^a_{kt}, ψ^a_{kt} for all models from Sections 2 − 5. All instances were solved to optimality in less than two seconds by the solver IBM ILOG CPLEX 12.4 and the models were coded in the IBM ILOG OPL IDE. We purposely omitted the objective function values as they are not relevant for our discussion.

Overall, results seem to indicate that even for an illustrative example, by incrementally introducing different features and methods for better tackling the food production planning, different solutions are obtained for almost every model tested. The only production plan leading to spoiled products is the base model (B-PP-FP) when products 1 and 2 reach an age of 1 in period 4 ($\rho^1_{14} = \rho^1_{24} = 5$). All the other models are able to avoid that these products reach the expiry dates by different reasons. For example, while the model considering inventory expiry (IE-PP-FP) avoids spoilage by the fact that we limit the demand fulfilment to products with a significant remaining shelf-life, the model introducing discarding costs (DC-PP-FP) is able to achieve the same solution by strongly penalizing the occurrence of expired inventory.

One interesting analysis lies on the different solutions between the inventory expiry model (IE-PP-FP) and the consumer behaviour one (CB-PP-FP) that have as the only difference constraints (12) and (13) in the CB-PP-FP model, which mimic the fact that customers pick up the fresher available products. For product 4 in period 2 when 20 units are produced in both models ($q_{142} = 20$), model IE-PP-FP is able to allocate to satisfy demand part of the production of period 2 and part of the production of period 1 with age 1 ($\psi^0_{42} = 73\%$ and $\psi^1_{42} = 27\%$). On the contrary, the CB-PP-FP model is forced to satisfy all demand in period 2 with the production executed in the same day ($\psi^0_{42} = 100\%$ and $\psi^1_{42} = 0\%$). These differences ultimately lead to the fact that customers in period 3 are penalized in the CB-PP-FP model as they will be satisfied with less fresh products ($\psi_{43}^{2} = 40\%$). This fact could potentially lead to lost sales and it reflects the importance of proper inventory control when dealing with perishable products.

From the 7 models, it is clear that the last 3 are able to better incorporate the consumer eagerness for fresher products. In special the model measuring freshness (MF-PP-FP) and the model having an extended demand parameter (DP-PP-FP) have an equivalent behaviour. Both models incorporate explicitly the importance of satisfying customers with a high degree of freshness. The difference between them relies more on the amount and quality of information the decision maker has when setting up the model (less information for the MF-PP-FP and more for the DP-PP-FP).

		q_{lkt}				ρ^a_{kt}										ψ^a_{kt}								
	Period (t)	1	2	3	4	1	2	2	3	3	3	4	4	4	4	1	2	2	3	3	3	4	4	4
Model	Product (k) \Age (a)	0	0	0	0	0	0	1	0	1	2	0	1	2	3	0	0	1	0	1	2	0	1	2
B-PP-FP (Base)	1	5	0	10	0	5	0	0	10	0	0	0	5	0	0	100%	0%	0%	100%	0%	0%	0%	100%	0%
	2	8	0	15	0	8	0	8	15	0	0	0	5	0	0	0%	0%	80%	100%	0%	0%	0%	100%	0%
	3	10	5	0	10	10	5	0	0	0	0	10	0	0	0	100%	100%	0%	0%	0%	0%	100%	0%	0%
	4	5	25	0	5	5	25	0	0	10	0	5	0	0	0	100%	100%	0%	0%	100%	0%	100%	0%	0%
IE-PP-FP (Inventory expiry)	1	5	0	5	5	5	0	0	5	0	0	5	0	0	0	100%	0%	0%	100%	0%	0%	100%	0%	0%
	2	0	10	10	5	0	10	0	10	0	0	5	0	0	0	0%	100%	0%	100%	0%	0%	100%	0%	0%
	3	**15**	**0**	0	10	**15**	**0**	**5**	0	0	0	10	0	0	0	100%	**0%**	**100%**	0%	0%	0%	100%	0%	0%
	4	**9**	**20**	**0**	5	**9**	**20**	**4**	**0**	**9**	**0**	5	0	0	0	100%	**73%**	**27%**	**0%**	**90%**	**0%**	100%	0%	0%
CB-PP-FP (Consumer behaviour)	1	5	0	5	5	5	0	0	5	0	0	5	0	0	0	100%	0%	0%	100%	0%	0%	100%	0%	0%
	2	0	10	10	5	0	10	0	10	0	0	5	0	0	0	0%	100%	0%	100%	0%	0%	100%	0%	0%
	3	**15**	**0**	0	10	**15**	**0**	**5**	0	0	0	10	0	0	0	100%	**0%**	**100%**	0%	0%	0%	100%	0%	0%
	4	**9**	**20**	**0**	5	**9**	**20**	**4**	**0**	**5**	**4**	5	0	0	0	100%	**100%**	**0%**	**0%**	**50%**	**40%**	100%	0%	0%
DC-PP-FP (Discarding costs)	1	5	0	5	5	5	0	0	5	0	0	5	0	0	0	100%	0%	0%	100%	0%	0%	100%	0%	0%
	2	0	10	10	5	0	10	0	10	0	0	5	0	0	0	0%	100%	0%	100%	0%	0%	100%	0%	0%
	3	**15**	**0**	0	10	**15**	**0**	**5**	0	0	0	10	0	0	0	100%	**0%**	**100%**	0%	0%	0%	100%	0%	0%
	4	**9**	**20**	**0**	5	**9**	**20**	**4**	**0**	**9**	**0**	5	0	0	0	100%	**73%**	**27%**	**0%**	**90%**	**0%**	100%	0%	0%
MF-PP-FP (Measuring freshness)	1	5	0	5	5	5	0	0	5	0	0	5	0	0	0	100%	0%	0%	100%	0%	0%	100%	0%	0%
	2	0	10	10	5	0	10	0	10	0	0	5	0	0	0	0%	100%	0%	100%	0%	0%	100%	0%	0%
	3	**10**	**5**	0	10	**10**	**5**	**0**	0	0	0	10	0	0	0	100%	**100%**	**0%**	0%	0%	0%	100%	0%	0%
	4	**6**	**14**	**10**	5	**6**	**14**	**1**	**10**	**0**	**0**	5	0	0	0	100%	**93%**	**7%**	**100%**	**0%**	**0%**	100%	0%	0%
VF-PP-FP (Value of freshness)	1	5	0	5	5	5	0	0	5	0	0	5	0	0	0	100%	0%	0%	100%	0%	0%	100%	0%	0%
	2	0	10	10	5	0	10	0	10	0	0	5	0	0	0	0%	100%	0%	100%	0%	0%	100%	0%	0%
	3	**10**	**5**	0	10	**10**	5	**0**	0	0	0	10	0	0	0	100%	**100%**	**0%**	**0%**	0%	0%	100%	0%	0%
	4	**5**	**14**	**10**	5	**5**	14	**0**	**10**	**0**	**0**	5	0	0	0	100%	**93%**	**0%**	**100%**	**0%**	**0%**	100%	0%	0%
DP-PP-FP (Demand parameter)	1	5	0	5	5	5	0	0	5	0	0	5	0	0	0	100%	0%	0%	100%	0%	0%	100%	0%	0%
	2	0	10	10	5	0	10	0	10	0	0	5	0	0	0	0%	100%	0%	100%	0%	0%	100%	0%	0%
	3	**10**	**5**	0	10	**10**	**5**	**0**	0	0	0	10	0	0	0	100%	**100%**	**0%**	0%	0%	0%	100%	0%	0%
	4	**6**	**14**	**10**	5	**6**	**14**	**1**	**10**	**0**	**0**	5	0	0	0	100%	**93%**	**7%**	**100%**	**0%**	**0%**	100%	0%	0%

Table 2: Results for decision variables q_{lkt}, ρ^a_{kt}, ψ^a_{kt} for all models (Sections 2 – 5). Results that are the same across all models are in light grey and results that differ between the models (except the base one – B-PP-FP) are in bold. Notice that the relation between the time and age index is respected in this table $a \in \mathcal{A} = \{a \in \mathbb{Z}^+_0 | a \leq t-1\}$.

7 Conclusions

In this paper, we have formulated several ways of integrating different challenges related to exogenous factors (such as customer behaviour and the perishable nature of the products) arising in the production planning of food products. The formulations have the same base model as starting point and we have organised them based on the extensions of the model components required: constraints, objective function and parameters. In particular, we have analysed how to limit the inventory age based on an adapted simple plant location reformulation, how to incorporate the consumer behaviour within the inventory policy, how to include discarding costs in the objective function, how to model customer willingness for fresh products in a multi-objective framework and how to value freshness either in the price or demand parameters. To analyse the implications of each of these "ingredients", an illustrative example is presented and solved, exposing the different solution structures achieved. The differences across the solutions show the importance of choosing a model suitable to the particular business setting, in order to accommodate the multiple challenges present in these industries. Moreover, acknowledging the perishable nature of the products and evaluating the amount and quality of information at hands may be crucial in lowering disposal costs and achieving higher service levels.

There are other ingredients that may not be so straightforward of incorporating that were not presented in this discussion for a matter of space constraints. For example, [Wang et al., 2009] deals with the incorporation of batch traceability that is increasingly important with the recent cases of products recall.

It is also possible to extend these "ingredients" to other supply chain planning problems dealing with food products. For example, [Amorim and Almada-Lobo, 2013b] uses a multi-objective framework for a vehicle routing problem of highly perishable food products. In a more integrated perspective [Rong et al., 2011] and [Amorim et al., 2012] show the importance of food quality across the production and distribution processes.

References

[Amorim et al., 2011] Amorim, P., Meyr, H., Almeder, C. and Almada-Lobo, B. (2011). Managing perishability in production-distribution planning: a discussion and review. *Flexible Services and Manufacturing Journal*, 10.1007/s10696-011-9122-3: 1–25

[Amorim et al., 2011b] Amorim, P., Antunes, C. H. and Almada-Lobo, B. (2011). Multi-Objective Lot Sizing and Scheduling Dealing with Perishability Issues. *Industrial Engineering & Chemistry Research*, 50:3371–3381.

[Amorim et al., 2012] Amorim, P. and Günther, H.-O. and Almada-Lobo, B. (2012). Multi-objective integrated production and distribution planning of perishable products. *International Journal of Production Economics*, 138(1):89–101.

[Amorim et al., 2013] Amorim, P., Costa, A. M. and Almada-Lobo, B. (2013). Influence of Consumer Purchasing Behaviour on the Production Planning of Perishable Food. *ORSpectrum*, To appear.

[Amorim and Almada-Lobo, 2013b] Amorim, P. and Almada-Lobo, B. (2013). The impact of Food Perishability Issues in the Vehicle Routing Problem. *Computers & Industrial Engineering*, To appear.

[Fleischmann et al., 2008] Fleischmann, B., Meyer, H. and Wagner, M. (2008). Advanced Planning. *Planning Supply Chain Management and Advanced Planning, Springer, Berlin, Heidelberg*: 81–106.

[Günther et al., 2006] Günther, H.-O., Grunow, M. and Neuhaus, U. (2006). Realizing block planning concepts in make-and-pack production using MILP modelling and SAP APO. *International Journal of Production Research*, 44(18):3711–3726.

[Rong et al., 2011] Rong, A. and Akkerman, R. and Grunow, M. (2011). An optimization approach for managing fresh food quality throughout the supply chain. *International Journal of Production Economics*, 131(1):421–429.

[Wang et al., 2009] Wang, X. and Li, D. and O'Brien, C. (2009). Optimisation of traceability and operations planning: an integrated model for perishable food production. *International Journal of Production Research*, 47(11):2865–2886.

A stochastic model for a multi-period multi-product closed loop supply chain

Susana Baptista*†, Maria Isabel Gomes*, Ana Paula Barbosa-Póvoa*‡

* *Centro de Matemática e Aplicações, FCT, Universidade Nova de Lisboa, 2829-516 Caparica, Portugal,*
E-mail:`sbb@fct.unl.pt`

† *Centro de Matemática e Aplicações, FCT, Universidade Nova de Lisboa, 2829-516 Caparica, Portugal,*
E-mail:`mirg@fct.unl.pt`

‡ *CEG-IST, Instituto Superior Técnico, Universidade Técnica de Lisboa, Av. Rovisco Pais, 1049-001 Lisboa, Portugal.*
E-mail:`apovoa@ist.utl.pt`

Abstract

In this work we propose a stochastic model for the design and planning of closed-loop supply chains. Uncertainties in demand and return volumes are modelled together with uncertain transportation costs. A two-stage stochastic programming is developed and a sensitivity analysis to the worst-case probability is performed in order to test the solution robustness. Finally, in order to prove the goodness of the stochastic approach, the value of the stochastic solution and the value of perfect information are computed. An example based on a real case shows the model applicability.

Palavras chave: Closed-Loop Supply Chain, Design and Planning, Two-stage Stochastic Optimization.

1 Introduction

Nowadays the integration of forward and reverse flows in supply chains is a major concern for industries resulting in a high interest at the academia level. In such context, as points out [Fleischmann et al., 2001], the simultaneous design of the forward and reverse channels may lead to significant cost savings. Closed-loop chain network design has been an area of intensive research in the past decade as shown in the detailed literature review of [Akçali et al., 2009]. Authors have addressed strategic and tactical decisions for a single product in a single period [Fleischmann et al., 2001], for a single product in a multi-period setting [Beamon and Fernandes, 2004], for multi-products in a single period [Uster et al., 2007] and for multi-products in a multi-period setting [Salema et al., 2010]. Notice that in the aforementioned models, all the parameters are assumed to be deterministic. However, given the strategic nature of design decisions, several sources of uncertainty play a major role in the network behavior and therefore should not be disregarded. The stochastic closed-loop network design problem has been addressed by fewer authors when compared with the deterministic case. One of the first works was accomplished by [Inderfurth, 2005] for a single period in a multi-period setting and by [Listes, 2007] for a single product network in a single period. [Salema et al., 2007] and [Chouinard et al., 2008] proposed a two-stage stochastic model for a multi-product product network in a single period. [Pishvaee and Rabbani, 2011] address uncertainties in the closed-loop supply chain through a robust optimization approach which avoids, according to the authors, the fitting of probability distributions for the uncertainty sources (transportation costs, demand and return volumes). [Zeballos, 2012] proposed a two-stage stochastic MILP model to study the quantity and quality of customer returned products. In this work, both sources are modelled simultaneously considering the maximization of total expected supply chain profit. Very recently, [Amin and Zhang, 2013] studied the design and planning of closed-loop supply chain in a three-step approach. Firstly, a qualitative approach is used to identify possible entities that will integrate the network design phase. Follows the evaluation of entities and the network configuration that in parallel approaches provide the inputs for the final step. In this third and last phase, a multi-objective MILP model selects and allocates customers' orders to the network entities. Uncertainties are addressed in phase two: a fuzzy approach models the uncertainty concerning the "importance" of entities to the network; a stochastic nonlinear MILP model tackles the network configuration where demand is the source of uncertainty. In this work different sources of uncertainties are tackled but not in an integrated way. [Cardoso et al., 2013] proposed a model for the design and planning of closed-loop supply chains and studied the impact that different network configurations have on the Net Present Value (NPV). In addition, authors address the demand uncertainty in a multi-period setting. Demand uncertainty is modelled by a discrete probability distribution, which in a three time period planning horizon leads to a total of nine scenarios. A sensitivity analysis is performed on demand volumes and on the scenario probability distribution considering two different cases: the network flow structure is fixed or changes with the scenario. [Ramezani et al., 2013] proposed a multi-objective two-stage stochastic model where several sources of uncertainty such as selling prices, costs and demand and return rates are modelled. The ϵ-constraint methodology is used to approximate the Pareto Front considering three objectives: the maximization of profit and customer's service level and the minimization of defective parts acquired from the suppliers.

With two exceptions, all the above works addressed stochasticity in a single period context and in a single source of uncertainty. In this work we study three sources of uncertainties: transportation costs, demand and return volumes. The large fluctuations of diesel prices observed during the last 4 years: -21%, 15%, 19% and 5% (price increase in Portugal and when compared with the year before) and the fact that transportation costs represent a large fraction of the supply chain costs (in [Cardoso et al., 2013] transportation costs account on average for 38% of the total costs), turn the capturing of this uncertainty as one of the key features of the present model. Notice that since operational decisions are stochastic decisions, in the sense that they depend on the randomness realizations, and network design decisions are not stochastic, since they're to be taken in a unique way given all circumstances, a two-stage stochastic programming is adopted as a modelling framework. Further, random parameters are assumed to be discretely distributed with two possible realizations: normal case and worst case. Given the complexity of assessing a value for the worst case probability a sensitivity analysis to the worst-case probability is performed in order to test the solution robustness and support the network design decisions. Finally, in order to prove the goodness of the stochastic approach, the value of the stochastic solution and the value of perfect information are computed.

The paper is structured as follows. In the next section a detailed description and complete formulation of the model developed is presented. Section 3 is devoted to the computational tests and result analysis regarding a multi-period and multi-commodity network case study and finally, section 4 states the main conclusions.

2 The modelling approach

The model here proposed is an extension of the previous work of [Salema et al., 2010] in the sense that it addresses the uncertainty problem by considering transportation costs, customers' demands and customers' returns to be stochastic.Other than stochasticity, further refinements regarding the previous model include the fact that sales revenues are also considered, all the monetary values involved in future time periods are updated to their present value and that manufacturing and remanufacturing processes are now distinguishable in the sense that, beyond the recycled raw materials/recovered components that feed production, the model also accounts for the amount of brand-new materials/components to be acquired, so that the production costs are fully captured.

All entities composing the supply chain (Figure 1) act as product transformation points. Factories manufacture new products and/or remanufacture used ones. Warehouses execute postponement operations, which customize products to meet customers' demand. Customers' disposed products are collected by disassembly centres which, after sorting and disassembling operations, send components to be remanufactured or to be properly disposed.

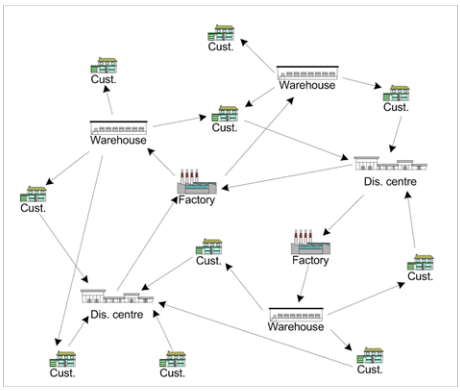


Figure 1: Closed-loop supply chain network structure.

Since operational decisions are to be undertaken for an entire time horizon, a two-unit scale is adopted for time modelling.Given a time horizon (e.g. 10 years), customers' demand has to be satisfied in some predefined time units, named the macro-period (e.g. yearly), while all planning decisions are to be taken in a smaller time scale, as months or weeks (the micro-period).

2.1 Stochastic modelling

In order to address stochasticity, we adopted a two-stage recourse model, where first stage decisions concern the design of the supply chain while second stage decisions regard the planning of the supply chain. We assumed that the random vector ξ that pieces together the stochastic components has finite support so that all the possible realizations of ξ are completely described by a set of scenarios. First stage decisions, involving the location of the four echelon network (plants, warehouses, customers and disassembly centres), are to be defined in order to maximize the expected net present value over all scenarios considered, and second stage decisions are to be defined for each one of the scenarios considered, so that different levels of production, storage and distribution flows will be obtained for each scenario.

As mentioned, three sources of uncertainty are modelled. Though a similar modelling approach is followed for each stochastic component, it is important to stress that each plays a different role in the model: demand is one of the models independent parameters, return uncertainties are variable constraint coefficients and transportation costs are objective function coefficients. As for the modelling, in particular for the demand modelling, if d denotes a product demand value and δ_s^d the demand variability factor for scenario s, the demand to be satisfied will be given by $d+\delta_s^d d = (1+\delta_s^d)d$. Regarding the returned volume, since it is assumed that demand doesn't have to be totally satisfied, the return volumes will depend on the met demand. Thus, if x_d denotes the customer product demand volume that is satisfied, ρ the expected return rate and δ_s^r the return variability factor of scenario s, the total returned volume is given by $(1+\delta_s^r)\rho x_d$. Finally for the modelling of transportations costs, where c denotes the transportation cost, x the transported amount and δ_s^t the transportation cost variability factor in scenario s, the total transportation cost will be given by $(1+\delta_s^t)cx$.

2.2 Model formulation

Consider the sets and parameters defined in Appendix A and B and the following decision variables.

Continuous variables:

$X_{spijt'}$ amount of product p dispatched from entity i to entity j at micro-period t' under scenario s
$Y^{Sup}_{spit'}$ ($Y^{dc}_{spit'}$) amount of product p manufactured (remanufactured) by factory i at micro-period t' under scenario s
$S_{spit'}$ amount of product p stocked at entity i at micro-period t' under scenario s
U_{spit} unmet demand of product p in customer i at macro-period t under scenario s.

Binary variables:

Z_i equal to 1 if entity i is opened/included in the supply chain.

The problem can then be stated as:

Max F=

$$\sum_{s\in S}\sum_{t'\in T_m}\sum_{i\in I_{dc}}\sum_{j\in I_c}\sum_{p\in P_{(I_{dc})}} prob_s \xi_p/(1+Rt)^{1+MacroP(t')} X_{spijt'}$$

$$-\sum_{i\in I\setminus I_c} c_i^{fix} Z_i$$

$$-\sum_{s\in S}\sum_{t'\in T_m}\sum_{i\in I_f}\sum_{p\in P_{(I_f)}} prob_s c^{pr_1}_{pit'}/(1+Rt)^{1+MacroP(t')}Y^{Sup}_{spit'}$$

$$-\sum_{s\in S}\sum_{t'\in T_m}\sum_{i\in I_f}\sum_{p\in P_{(I_f)}} prob_s c^{pr_2}_{pit'}/(1+Rt)^{1+MacroP(t')}Y^{dc}_{spit'}$$

$$-\sum_{s\in S}\sum_{t'\in T_m}\sum_{i,j\in I i\neq j}\sum_{p\in P_{(Set(i))}} prob_s(1+\delta^t_{sMacroP(t')})c^{tr}_{pit'}/(1+Rt)^{1+MacroP(t')}d_{ij}X_{spijt'}$$

$$-\sum_{s\in S}\sum_{t'\in T_m}\sum_{i\in I_{dc}}\sum_{p\in P_{(I_{dc})}} prob_s(1+\delta^t_{sMacroP(t')})c^{tr}_{pit'}/(1+Rt)^{1+MacroP(t')}d_{iDisp}X_{spiDispt'}$$

$$-\sum_{s\in S}\sum_{t'\in T_m}\sum_{i\in I\setminus I_c}\sum_{p\in P_{(Set(i))}} prob_s c^{st}_{pit'}/(1+Rt)^{1+MacroP(t')}S_{spit'}$$

$$-\sum_{s\in S}\sum_{t\in T}\sum_{i\in I_c}\sum_{p\in P_{(I_{dc})}} prob_s c^{u}_{pit'}/(1+Rt)^{1+t}U_{spit} \qquad (1)$$

s.t.

$$\sum_{\tilde p\in P_{I_{sup}}}\beta_{\tilde p p}X_{s\tilde p i(t'-\tau_{Sup,i})} = Y^{Sup}_{spi(t'+\phi^{Sup}_{\tilde p p})} \qquad s\in S, p\in P_{(I_f)}, i\in I_f, t'\in T_m : (t'-\tau_{Sup,i})\in T_m \qquad (2)$$

$$\sum_{\tilde p\in P_{I_{dc}}}\beta_{\tilde p p}X_{s\tilde p i(t'-\tau_{dc,i})} = Y^{dc}_{spi(t'+\phi^{dc}_{\tilde p p})} \qquad s\in S, p\in P_{(I_f)}, i\in I_f, t'\in T_m : (t'-\tau_{dc,i})\in T_m \qquad (3)$$

$$S_{spi(t'-1)} + Y^{Sup}_{spit'} + Y^{dc}_{spit'} = S_{spit'} + \sum_{j\in I_{dc}} X_{spijt'} \qquad s\in S, p\in P_{(I_f)}, i\in I_f, t'\in T_m : (t'-1)\in T_m \quad (4)$$

$$S_{spi(t'-1)} + \sum_{j\in I_f}\sum_{\tilde p\in P_{I_f}}\beta_{\tilde p p}X_{s\tilde p ji(t'-\tau_{ji}-\phi_{\tilde p p})} = S_{spit'} + \sum_{j\in I_f} X_{spijt'}$$
$$s\in S, p\in P_{(I_w)}, i\in I_w, t'\in T_m : (t'-1)\in T_m \qquad (5)$$

$$S_{spi(t'-1)} + \sum_{j\in I_c}\sum_{\tilde p\in P_{I_c}}\beta_{\tilde p p}X_{s\tilde p ji(t'-\tau_{ji}-\phi_{\tilde p p})} = S_{spit'} + \sum_{j\in I_f} X_{spijt'} + X_{spiDispt'}$$
$$s\in S, p\in P_{(I_{dc})}, i\in I_{dc}, t'\in T_m : (t'-1)\in T_m \qquad (6)$$

$$U_{spit} \geq Z_i(1+\delta^d_{st})d_{pit} - \sum_{j\in I_w}\sum_{t'\in MicroP(t)} X_{spji(t'-\tau_{ji})} \qquad s\in S, p\in P_{(I_w)}, i\in I_c, t\in T \qquad (7)$$

$$\sum_{j\in I_w}\sum_{\tilde p\in P_{I_w}}(1+\delta^r_{s\tilde p p})\rho_{\tilde p p}X_{s\tilde p ji(t'-\tau_{ji}-\phi_{\tilde p p})} = \sum_{j\in I_{dc}} X_{spijt'}$$
$$s\in S, p\in P_{(I_c)}, i\in I_c, t'\in T_m : (t'-\tau_{ji}-\phi_{\tilde p p})\in T_m \qquad (8)$$

$$X_{spiDispt'} \leq (1-\alpha_p)\sum_{j\in I_c}\sum_{\tilde p\in P_{I_c}}\beta_{\tilde p p}X_{s\tilde p ji(t'-\tau_{ji}-\phi_{\tilde p p})}$$
$$s\in S, p\in P_{(I_{dc})}, i\in I_{dc}, t'\in T_m : (t'-\tau_{ji}-\phi_{\tilde p p})\in T_m \qquad (9)$$

$$\sum_{p\in P_{(I_f)}}(Y^{Sup}_{spit'} + Y^{dc}_{spit'}) \leq MaxCap_i \times Z_i \qquad s\in S, i\in I_f, t'\in T_m \qquad (10)$$

$$\sum_{p \in P_{(I_f)}} (Y^{Sup}_{spit'} + Y^{dc}_{spit'}) \geq MinCap_i \times Z_i \qquad s \in S, i \in I_f, t' \in T_m \tag{11}$$

$$\sum_{p \in P_{(Set(i))}} X_{spijt'} \leq MaxF_{ij} \times Z_i \qquad s \in S, i, j \in I : i \neq j, t' \in T_m \tag{12}$$

$$\sum_{p \in P_{(Set(i))}} X_{spijt'} \geq MinF_{ij} \times Z_i \qquad s \in S, i, j \in I : i \neq j, t' \in T_m \tag{13}$$

$$\sum_{p \in P_{(Set(i))}} S_{spit'} \leq MaxSt_i \times Z_i \qquad s \in S, i, j \in I \backslash I_c : i \neq j, t' \in T_m \tag{14}$$

$$\sum_{i \in I_w} Z_i \geq \sum_{i \in I_f} Z_i \tag{15}$$

$$X_{spijt'} \geq 0 \qquad s \in S, p \in P_{(Set(i))}, (i, j) \in A, t' \in T_m \tag{16}$$

$$Y^{Sup}_{spit'} \geq 0, Y^{dc}_{spit'} \geq 0 \qquad s \in S, p \in P_{(I_f)}, i \in I_f, t' \in T_m \tag{17}$$

$$S_{spit'} \geq 0 \qquad s \in S, p \in P_{(Set(i))}, i \in I \backslash I_c, t' \in T_m \tag{18}$$

$$U_{spit} \geq 0 \qquad s \in S, p \in P_{(I_c)}, i \in I_c, t' \in T_m \tag{19}$$

$$Z_i \in \{0, 1\} \qquad i \in I \tag{20}$$

The objective function (1) expresses the total expected supply chain profit composed by: the expected revenue (first term), the opening fixed costs (second term), the expected productions costs for manufactured products (third term) and for remanufactured products (fourth term), expected transportation costs between all entities (fifth term), expected disposal costs (sixth term), expected stock costs (seventh term) and finally the expected penalty cost for not serving the demand of the included customers. All monetary values are reduced to their present value and fixed investment costs are equally divided over the assets useful life. Since the time horizon considered (e.g. five years) is assumed to be smaller than the assets useful life (e.g. fifteen years), the assets value at the end of the time horizon is not considered. Notice that a long asset useful life assumption is only appropriate if the addressed industry is not of short life-cycle products, otherwise a procedure similar to the one proposed by [Cardoso et al., 2013] must be performed. Equation (2) ensures that for every scenario, for each manufactured product p, all the components needed in the factories at time t' are transformed into p, so that p is available at time $t' + \Phi^{Sup}_{\tilde{p}p}$, where $\Phi^{Sup}_{p\tilde{p}}$ is the respective production time. Equation (3) establishes an equivalent result for remanufactured products. Equations (4), (5) and (6) are the balance equations at factories, warehouses and disassembly centres that ensure that the total flow dispatched from the entity (outbound flow) is equal to the level stock changes plus the flow that is dispatched to the entity (inbound flow). Notice that in equations (5) and (6), for a given entity the outbound flow of product p at time t' depends on the inbound flow at time $t' - \tau_{ji} - \Phi_{p\tilde{p}}$ where $\Phi_{p\tilde{p}}$ is the processing time of product $\tilde{p}$ into p, and τ_{ji} is the shipping time between the origin and destination entities. A main feature of this model is the fact that product demands are considered to be stochastic. Since a recourse programming context was adopted, penalties for not meeting customer's demands were considered in the objective function. The amount of unmet demand is precisely defined by equation (7), which states that for every scenario and for every client, the unmet demand of product p is the difference between the demand of the product and the amount that was delivered to the client during the macro-period t. Notice that a product that reaches a client at time t', must have been dispatched from the warehouse at time $t' - \tau_{ji}$, thus the amount of p delivered to customer i is given by $X_{spji(t'-\tau_{ji})}$ summed over all warehouses j and all micro-time periods t' that occur between the macro-time periods $[t-1, t]$. Regarding the demand of p, $Z_i(1 + \delta^d_{st})d_{pit}$ ensures that only demands of clients that belong to the network are considered. Since d_{pit} is the expected demand, δ^d_{st} is the demand variability factor (measured as a percentage of the expected demand and dependent of t) under scenario s. Another new feature of this model regards the stochasticity of product returns which is handled by equation (8). This equation is the balance equation at customers that ensures that for every scenario, customer and product returned from customers, the amount of product p shipped to all disassembly centers at time t' must be equal to the amount of products $\tilde{p}$ that are returned as p. Notice that the amount of products $\tilde{p}$ to be returned as p, is the amount of $\tilde{p}$ dispatched from all the warehouses at time $t' - \tau_{ji} - \phi_{(\tilde{p}p)}$ (where the parameters are defined as before) times the return rate of product $\tilde{p}$ as p which is given by $(1 + \delta^r_{\tilde{p}p})\rho_{\tilde{p}p}$. Since $\rho_{\tilde{p}p}$ is the expected return rate of $\tilde{p}$ as p, $\delta^r_{\tilde{p}p}$ is the return variability factor (measured as a percentage of the expected return rate) under scenario

s. Equation (9) assures that disassembly centres can only send to disposal less than a fraction of the returned products, in order to comply with the recovery targets set by legislation. This equation can be easily modified to cover the case where the disposal represents a third player (e.g. external recycling) so that $(1 - \alpha_p)$ is the fraction of products not suitable for remanufacturing. The maximum and minimum production capacities of factories are defined by equations (10) and (11) respectively, the maximum and minimum flow capacities are defined by equations (12) and (13) respectively and finally (14) sets the maximum stock levels at the different entities where products may be stored, i.e. factories, warehouses and disassembly centres. Constraint (15) ensures that the number of warehouses should be at least as large as the number of plants, so that the distribution structure is more decentralized in order to better encompass the transportation costs variations. Constraints (16) to (20) establish variables' domains.

3 Example details and results analysis

The computational tests were performed on a multi-period and multi-commodity network based on the work addressed by [Salema et al., 2010] where a glass supply chain network was studied. This network superstructure was defined with possible five plants, eight warehouses, 18 customers and eight disassembly centres. Three different products were considered in the flows plants-warehouses (F1 to F3), six for the flows warehouses-customers (A1 to A6), one for the flows customers-disassembly centres (R) and finally two for the closing loop flows disassembly centres-plants (C1 and C2) and one for the flows suppliers-plants (S). A three-year time horizon, with one year macro-time unit and three months micro-time unit was considered (parameters' values may be found in Appendix C). Regarding the number of scenarios, a normal case (NC) and a worst case (WC) scenario were identified.

As mentioned, transportation costs were defined as $(1 + \delta^{t}_{sMacroP(t')})c^{tr}_{pit'}$ and an annual variability factor of 5% and 10% increase was set for the normal and worst cases respectively. Customers' demands were assumed to increase annually 2% under a normal scenario and decrease 10% in the worst case. Customers' returns were defined as $(1 + \delta^{r}_{s\tilde{p}p})\rho_{\tilde{p}p}$ where the values of the parameters $\rho_{\tilde{p}p}$ and $\delta^{r}_{s\tilde{p}p}$ were set as in Table 1. $\delta^{r}_{s\tilde{p}p}$ was assumed to have a null value under the normal scenario $\delta^{r}_{NC\tilde{p}p}$ so that no variability is incurred, while for the worst case scenario $\delta^{r}_{WC\tilde{p}p}$, the values were defined in order to ensure that fewer products are collected and that products with a larger return rate also present a smaller variability.

Table 1: Values for Return fraction and variability factor.

			A1	**A2**	**A3**	**A4**	**A5**	**A6**
Return fraction		$\rho_{\tilde{p}p}$	0.45	0.7	0.5	0.8	0.4	0.9
Variability factor	Normal Case	$\delta^{r}_{NC\tilde{p}p}$	0	0	0	0	0	0
	Worst Case	$\delta^{r}_{WC\tilde{p}p}$	0.2	0.1	0.2	0.05	0.2	0.05

The model was implemented in OPL using CPLEX 12.4 as solver. All the tests reported were conducted on a laptop with a 2.4 GHz Core i5 processor and all recourse problems were solved on average in 2079 seconds with computing times ranging from 1200 till 2871 CPU seconds. Recourse problems have 22066 constraints and 26505 variables from which 39 are binary. All monetary values presented were rescaled by a 1/1000 monetary units (m.u.) factor.

In order to analyse the solution robustness to the worst case scenario probability, the net present value and the number of entities were computed for probability values ranging from 0 to 100% following an increment of 20%. The results obtained for the expected value of perfect information (EVPI) and the value of the stochastic solution (VSS) were also analysed. Since the problem under analysis is a maximization problem, it is well known (see [Birge and Louveaux, 1997]) that the expected value of perfect information is the defined as EVPI = WS - RP and the value of the stochastic solution as VSS = RP - EMV, where RP denotes the recourse program presented above, WS the wait-and-see problem and EMV the expected result of the mean problem. Thus, RP = $min_x E_{\xi} z(x, \xi)$ will define the recourse problem, WS = $E_{\xi}(min_x z(x, \xi))$ the wait-and-see problem and EMV = $E_{\xi}(z(min_x z(x, E(\xi)), \xi))$ the expected result of the mean problem.

Figure 2 shows that, for whatever problem considered, the net present value decreases when the worst case probability increases and that globally, for worst case probabilities levels above 60% the supply chain system becomes unprofitable. Figure 3 exhibits the fact that, for the expected mean value problem, the network structure has two different configurations and that again the 60% worst case probability level is the turning point. Notice that the graphs of the number of plants and distribution centres are overlapped in figure 3. When analysing the NPV decrease trends (Fig. 3), it becomes clear that the WS problem solution exhibits a linear trend due to the fact that the probabilities changes only affect the objective function coefficients. As for the EMV, the NPV decrease follows from the network topology decrease depicted by figure 3. Regarding the recourse problem, two different decreasing rates trends can be identified. A steepest NPV decrease for probability values up to 60% with an average loss of 500 (m.u.$\times 10^3$) per a 20% probability increase, while for probability values above 60%, the average loss is only of 200 (m.u.$\times 10^3$) per a 20% probability increase. Such trend difference may be explained by the fact that the system topologies of probability levels up to 40% include 14 customers, while for probability levels beyond 60%, only 7 customers are considered (see Table 2). Notice that as theoretically expected the RP solution value is bounded above by the WS solution value and below by the EMV solution value.

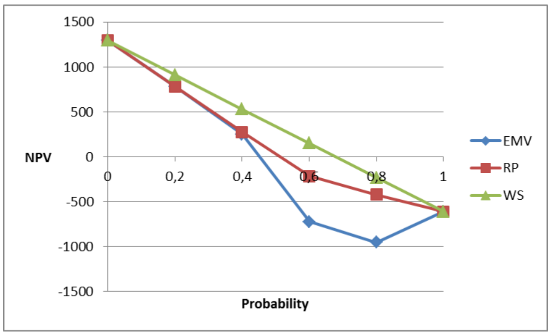


Figure 2: Relation between models' solutions and worst case probability.

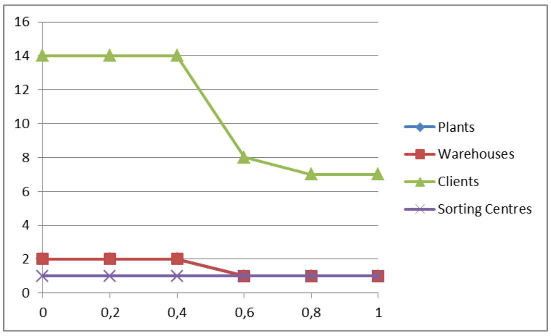


Figure 3: Relation between EMV network entities number and worst case probability.

Regarding the network structure solutions of the EMV and RP problems, Figure 3 and Table 2 show that as the worst case probability increases, the network becomes smaller, but in all cases the network design is quite robust to the worst case probability changes since only two different network configurations come up from the six probability levels analysed. As Table 3 points out no network modification occurs at probability levels up to 40%, at 60% the topology changes dramatically by reducing from two to one plant, from two warehouses to a single one, and cutting 7 of the 14 clients. Notice that for networks designed for worst case probability levels higher than 60% only small changes occur with a single customer being replaced by another one.

Table 2: Hamming distance between the EMV an RP solution.

Probability	**0**	**0.2**	**0.4**	**0.6**	**0.8**	**1**
Plants	0	0	0	0	0	0
Warehouses	0	0	0	0	0	0
Clients	0	0	0	1	0	0
Sorting Centres	0	0	0	0	0	0

Table 3: Hamming distance for RP solutions pairs.

Probability	**(0,0.2)**	**(0.2,0.4)**	**(0.4,0.6)**	**(0.6,0.8)**	**(0.8,1)**
Plants	0	0	1	0	0
Warehouses	0	0	1	0	0
Clients	0	0	7	1	1
Sorting Centres	0	0	0	0	0

Figure 4 presents the expected value of perfect information and the value of the stochastic solution. From Figure 2, it is clear that the deviation of the RP solution regarding the WS solution, and thus the EVPI value, increases for worst case probabilities up to 60% and decreases for probabilities that exceed that level, which explains the EVPI trend shown in Figure 4. Also the fact that the deviations of the EMV and RP solutions are larger for worst case probabilities ranging from 60 to 80% as it can be seen in Figure 2, explains the VSS trend depicted by Figure 4. Globally, both plots show that the 60% worst case probability level is the probability level for which it pays the most to access accurate information about the future (EVPI plot) and for which the cost of ignoring uncertainty is one of the largest (VSS plot). Another important feature is the fact that the value of the stochastic solution is almost insignificant for probability levels below 40%.

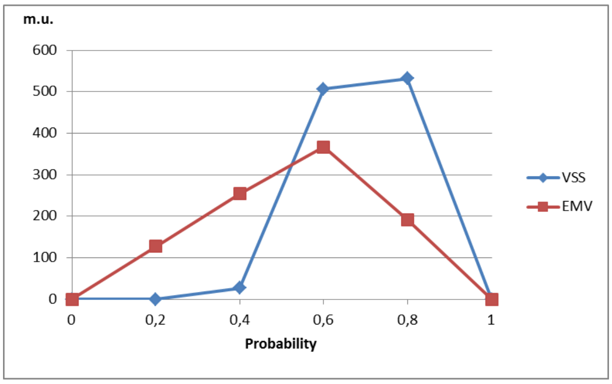


Figure 4: Relations between EVPI, VSS and worst case probability.

Finally, it is important to notice that, from a management perspective, the probability levels that are relevant are all that do not exceed the 60% level, since that is the range where the system is profitable. From a methodological perspective, any system analysis performed for probability levels below 40% can be achieved by applying the simple deterministic expected mean value problem, while for a worst case probability above that level, a stochastic programing approach should be adopted.

4 Conclusions

In this work we propose a two-stage MILP model for the design and planning of closed-loop supply chains accounting for different sources of uncertainty: product demands, return volumes, and transportation costs. Moreover, a multi-period and multi-product context is also contemplated.

The proposed formulation was applied to case based on a Portuguese glass company, where two scenarios (normal and worst case) were solved and results compared. A sensitive analysis was performed to assess the solution robustness regarding the worst case scenario probability. The expected value of perfect information and the value of the stochastic solution were also analysed. Results have shown that for values between 40% and 60% of the worst case probability, the network stops being profitable. It is between these same values that the network structure suffers the largest change. Half of the customers are no longer supplied and it can be observed a reduction to half on the number of opened factories and warehouses.

Given the fact that the range of the network profitability has been identified, a deeper analysis to the network robustness should be conducted. Such analysis will certainly involve a larger number of scenarios and thus the development of a decomposition strategy that will allow such modelling stands as future work. Furthermore, in this work we considered decisions to be taken in a risk neutral context, but in a context with a large uncertainty as it happens in the supply chain environment, risk management models should be encompassed. Thus, an extension of the present model that incorporates risk-averse measures such as the conditional value at risk will be undertaken. Finally, let us note that though we restricted the network design decisions to the first stage, a two-stage stochastic modelling approach would still be appropriate even if network decisions had to be taken throughout the time horizon, as long as those decisions had to be the same. On the contrary, if network decisions depended on the outcomes up to the moment where they're to be taken, then a multi-stage stochastic modelling approach should be considered. Such extension will also be considered in the future so that the real multi-stage management decisions of the supply-chain system are fully captured.

Acknowledgement

This work was partially supported by the Portuguese National Science Foundation through the project PTDC/SEN-ENR/102869/2008 and by CMA/FCT/UNL, under the project PEst OE/MAT/UI0297/2011.

References

[Akçali et al., 2009] Akcali, E., Çetinkaya, S., Üster, H., (2009). Network design for reverse and closed-loop supply chains: An annotated bibliography of models and solution approaches. *Networks*, 53:231–248.

[Amin and Zhang, 2013] Amin, S.H., Zhang, G., (2013). A three-stage model for closed-loop supply chain configuration under uncertainty. *International Journal of Production Research*, 51:1405–1425.

[Birge and Louveaux, 1997] Birge,J., Louveaux, F.,(1997). Introduction to Stochastic Programming. Springer

[Beamon and Fernandes, 2004] Beamon,B.M., Fernandes, C., (2004). Supply-chain network configuration for product recovery. *Production Planning & Control*, 15: 270–281.

[Cardoso et al., 2013] Cardoso, S.R., Barbosa-Póvoa, A.P.F.D., Relvas, S.,(2013). Design and Planning of Supply Chains with Integration of Reverse Logistics Activities under Demand Uncertainty. *European Journal of Operational Research* 226: 436–451.

[Chouinard et al., 2008] Chouinard, M.,D'Amours, S., Aït-Kadi, D., (2008). A stochastic programming approach for designing supply loops. *International Journal of Production Economics*, 113: 657–677.

[Fleischmann et al., 2001] Fleischmann, M.,Beullens, P., Bloemhof-Ruwaard, J.M., Wassenhove, L.N. Van, (2001). The impact of product recovery on logistics network design. *Production and Operations Management*, 10: 156–173.

[Inderfurth, 2005] Inderfurth, K (2005) Impact of uncertainties on recovery behavior in a remanufacturing environment: a numerical analysis. *International Journal of Physical & Logistics Management*, 35:318–336.

[Listes, 2007] Listes, O., (2007). A generic stochastic model for supply-and-return network design. *Computers & Operations Research*, 34: 417–442.

[Pishvaee and Rabbani, 2011] Pishvaee, M.S., Rabbani, M., (2011). A robust optimization approach to closed-loop supply chain network design under uncertainty. *Applied Mathematical Modelling*, 35:637–649.

[Ramezani et al., 2013] Ramezani, M., Bashiri,M., Tavakkoli-Moghaddam, R., (2013). A new multi-objective stochastic model for a forward/reverse logistic network design with responsiveness and quality level. *Applied Mathematical Modelling*, 37: 328–344.

[Salema et al., 2007] Salema, M.I.G., Barbosa-Povoa,A.P., Novais, A.Q., (2007). An optimization model for the design of a capacitated multi-product reverse logistics network with uncertainty. *European Journal of Operational Research*, 179: 1063–1077.

[Salema et al., 2010] Salema, M.I.G.,Barbosa-Povoa, A.P., Novais, A.Q., (2010). Simultaneous design and planning of supply chains with reverseflows: A generic modelling framework. *European Journal of Operational Research*, 203: 336–349.

[Uster et al., 2007] Uster, H., Easwaran, G.,Akçali, E., Çetinkaya, S., (2007). Benders decomposition with alternative multiple cuts for a multi-product closed-loop supply chain network design model. *Naval Research Logistics (NRL)*, 54: 890–907.

[Zeballos, 2012] Zeballos, L.J., Gomes, M.I.,Barbosa-Povoa, A.P., Novais, A.Q.,(2012). Addressing the uncertain quality and quantity of returns in closed-loop supply chains. *Computers & Chemical Engineering*, 47: 237–247.

Appendix A:
The following **sets** were considered:
I_f, I_w, I_c, I_{dc} potential location for factories, warehouses,clients and disassembly centres, respectively
$I = I_f \bigcup I_w \bigcup I_c \bigcup I_{dc}$
$P_{(I_*)}$ products/components supplied by entities of set $I_* \subseteq I$
P_{Sup} components for manufactured products purchased at suppliers
S scenarios considered
T_m, T micro-time and macro-time periods, respectively
Set of all possible network flows:
$A = \{(i,j) : \{i \in I_f \wedge j \in I_w\} \bigcup \{(i,j) : i \in I_w \wedge j \in I_c\} \bigcup \{(i,j) : i \in I_c \wedge j \in I_{dc}\} \bigcup \{(i,j) : i \in I_{dc} \wedge j = Disp\} \bigcup \{(i,j) : i \in I_{dc} \wedge j \in I_f\}$
where $Disp$ is the disposal option made available at disassembly centres.

Appendix B:
The following **parameters** were considered:
<u>*Revenues*</u>
ξ_p sell price of product p
<u>*Costs*</u>
R_t interest rate
c_i^{fix} opening fixed cost of entity i
$c_{pit'}^{pr_1}$ unit production cost of product p manufactured by factory i at time-period t'
$c_{pit'}^{pr_2}$ unit production cost of product p manufactured by factory i at time-period t'
$c_{pit'}^{tr}$ unit transportation cost of product p supplied by entity i at time-period t'
$c_{pit'}^{st}$ unit cost of product p stocked at entity i at time-period t'
c_{spit}^{u} unit penalty cost of unmet demand of product p at customer i at macro-time t
<u>*Product parameters*</u>
$\beta_{\tilde{p}p}$ conversion product rate of component $\tilde{p}$ into product p
$\phi_{\tilde{p}p}^{Sup}$ manufacturing time of component $\tilde{p}$ into product p
$\phi_{\tilde{p}p}^{dc}$ remanufacturing time of component $\tilde{p}$ into product p
$\phi_{\tilde{p}p}$ processing time of product $\tilde{p}$ into product p
$\rho_{\tilde{p}p}$ product $\tilde{p}$ expected return rate as product p
d_{pit} demand of product p at customer i at macro-time t
α_p recovery target of product p
<u>*Time parameters*</u>
τ_{ji} transportation time between entities j and i
<u>*Distance parameter*</u>
d_{ji} distance between entities j and i
<u>*Scenario parameters*</u>
$prob_s$ scenario probability
$\delta_{sMacroP(t')}^{t}$ transportation cost variability factor at macro-time t for scenario s
δ_{st}^{d} demand variability factor at macro-time t for scenario s
$\delta_{s\tilde{p}p}^{r}$ product $\tilde{p}$ return rate as product p variability factor for scenario s
<u>*Capacity parameters*</u>
$MaxCap_i$ maximum production capacity at factory i
$MinCap_i$ minimum production capacity at factory i
$MaxF_{ij}$ maximum flow between entities i and j
$MinF_{ij}$ minimum flow between entities i and j
$MaxSt_i$ maximum stock level at entity i
<u>*Functions of parameters*</u>
$Set(i) = I_*$ if $i \in I_*$
$MicroP(t) = \{t'_k, t'_{k+1}, \dots, t'_{k+n}\}$ set of micro-time periods between macro-time periods $(t-1)$ and t
$MacroP(t') = \lfloor \frac{t'-1}{n} \rfloor = t$ macro-time t at which micro-time t' belongs.

Appendix C:
I_f = {Évora, Leiria, Lisboa, Porto, Setúbal}
I_w = {Braga, Coimbra, Leiria, Lisboa, Porto, Santarém,Setúbal,Viseu}
I_c = {Aveiro, Beja, Braga, Bragança, Castelo Branco, Coimbra, Évora, Faro, Guarda, Leiria Lisboa, Portalegre, Porto, Santarém, Setúbal,Viana Castelo, Vila Real, Viseu}

$I_d c$ = {Braga, Coimbra, Leiria, Lisboa, Porto, Santarém,Setúbal,Viseu}
$P_{(I_f)}$ = {F1, F2, F3} $P_{(I_w)}$ = {A1,A2,A3,A4,A5,A6}
$P_{(I_c)}$ = {R} $P_{(I_{sc})}$ = {C1,C2} P_{Sup} = {M1,M2}
$T_m = 4$ $T = 3$
$Rt = 3\%$

Genetic Algorithms for the SearchCol++ framework: application to drivers' rostering

Vítor Barbosa*†, Ana Respício◇†, Filipe Alvelos⊗⋆

* *Escola Superior de Ciências Empresariais, Instituto Politécnico de Setúbal*
† *Centro de Investigação Operacional, Faculdade de Ciências da Universidade de Lisboa*
*E-mail:*vitor.barbosa@esce.ips.pt

◇ *Departamento de Informática, Faculdade de Ciências da Universidade de Lisboa*
*E-mail(second author):*respicio@fc.ul.pt

⊗ *Departamento de Produção e Sistemas, Escola de Engenharia da Universidade do Minho*
⋆ *Centro Algoritmi, Universidade do Minho*
*E-mail(third author):*falvelos@dps.uminho.pt

Abstract

This paper presents a new genetic algorithm included in the SearchCol++ framework. The new genetic algorithm includes an elitism strategy and a local search procedure to improve the quality of solutions and performance. The new algorithm is tested in a Bus Driver Rostering Problem decomposition model included in the framework in order to build valid rosters combining subproblems' solutions, obtained previously by using column generation. Each subproblem solution is a valid work-schedule for the driver corresponding to the subproblem. Computational tests show relevant improvement in the effectiveness and efficiency of the new algorithm to build valid rosters to the BDRP.

Keywords: genetic algorithm, hybrid optimization methods, column generation, rostering.

1 Introduction

"Personnel scheduling, or rostering, is the process of constructing work timetables for its staff so that an organisation can satisfy the demand for its goods or services"[Ernst et al., 2004]. The Bus Driver Rostering Problem (BDRP), in particular, consists of defining the work-schedule for the bus company's set of drivers for a defined rostering period, fulfilling all tasks (sequence of trips and rest periods defining a complete work day) on each day and respecting existing labour and company rules. For each driver, on each day, a task or a day off is assigned.

The BDRP and most rostering problems are NP-Hard combinatorial optimization problems [Dorne, 2008, Moz et al., 2009], being computationally difficult to obtain optimal solutions. Many authors address rostering problems with heuristic methods which are usually faster in the achievement of good solutions comparatively to exact methods. Examples of the use of non-exact methods are presented in [Ruibin et al., 2010, Moz et al., 2009, Lucic and Teodorovic, 2007, Burke et al., 2003] and include the use of metaheuristics, tabu-search, simulated annealing hyper-heuristic and evolutionary algorithms.

The drivers rostering is the last phase of the transportation planning system which also includes timetabling, vehicle scheduling and crew scheduling in order to know the drivers demand in each day [Leone et al., 2011, Nurmi et al., 2011, Xie et al., 2012]. Our concern is in the assignment of tasks/duties to each driver, assuming that the construction of those tasks was previously done by joining trips and rest times to obtain a complete daily task ready to assign to a driver.

The work presented in [Moz et al., 2009] proposes a new formulation to the BDRP where all the labour rules for a real bus company are forced by hard constraints and where other hard constraints assure the assignment of all tasks in each day and also assure the assignment of a task or a day off to each driver on each day of the rostering period (four weeks are considered in the model). Given the dificulty of the problem, particularly because they propose a bi-objective function to be optimized, the proposed approach is the use of non-exact methods: evolutionary algorithms.

Starting from the model proposed in [Moz et al., 2009], a new decomposition model for the BDRP was proposed in [Barbosa et al., 2013], where most of the constraints are considered in subproblems each of which associated to a single driver. Each subproblem has the variables and constraints related only to the driver it represents and is responsible to obtain valid lowest cost solutions corresponding to work-schedules for that driver. The binary variables of the subproblem define which task is assigned to the driver on each day of the rostering period. The tasks in which the duration goes beyond the regular daily labour time incur in overtime, which is paid with different costs for groups of driver. The day off is represented by an additional task without cost. The construction of valid work-schedules is restricted by constraints responsible for respecting the maximum work time in each week and in all the rostering period, for preventing the assignment of consecutive work days (without a day off in the middle) beyond a limit defined, assuring a minimum number of days off on Sundays and assuring a minimum number of days off in each week, assuring the assignment of a task in each day (eventually the task that represents taking the day off on that day). In addition, the assignment of incompatible tasks for consecutive days is also prevented (sequences of tasks late-early without respecting the minimum rest time). Every time a work-schedule assigns at least one work task (not the day off) a fixed cost is added to the corresponding solution, representing the cost of using the driver. A restricted master problem (RMP) including a set of constraints from the original model used to assure the assignment of all tasks (linking constraints) redefines the original model with variables associated to feasible solutions from all the subproblems. Additional convexity constraints were added to assure a complete work-schedule for each driver. For more details about the decomposition model please see [Barbosa et al., 2013].

A roster for the BDRP is a global integer solution with the work-schedule for each driver, assigning all tasks available to accomplish by the bus company on each day of the rostering period. The paper [Barbosa et al., 2013] presents the use of column generation [Desaulniers et al., 2005, Dantzig and Wolfe, 1960] followed by genetic algorithms [Reeves, 1997, Holland, 1992] to obtain valid rosters (global solution with the work-schedule for each driver, assigning all tasks available to accomplish) by integrating the proposed decomposition model in the SearchCol framework presented in [Alvelos et al., 2010] and detailed in [Alvelos et al., 2013].

In the next section, we present an overview of the SearchCol framework, the information available after the column generation and how the metaheuristics (particularly the genetic algorithms) are used to build global valid integer solutions for the BDRP. In section 3 we present our current contribution with a different version of the genetic algorithm which consider the possibility of using elitism to keep best individuals through generations and also use a simple local search to search neighbour solutions of the best found in each iteration. Later, in section 4 computational tests are presented and the results discussed. The last section provides some concluding observations.

2 BDRP decomposition model in the SearchCol++ framework

SearchCol [Alvelos et al., 2010, Alvelos et al., 2013] is a framework that allows the use of different metaheuristics on the search of good solutions over diverse optimization problems. SearchCol++ is the computational implementation of the theoretical framework proposed. Distinct problems from several research areas are included in the framework and it offers diverse non-exact methods that may be used to solve all the problems. The connection between all the problems, and the core of the framework, is that the central method applied is the column generation, followed then by the use of one or more metaheuristics to build global integer solutions using the columns (subproblem solutions) obtained during the column generation. Column generation [Desaulniers et al., 2005] is an exact method which allows to find linear optimal solutions. The framework authors claim that the set of columns generated along the optimization is a potentially good search space as the optimal linear solution may be important to guide the metaheuristics.

Each problem included in the SearchCol framework needs to have a decomposition model on which the column generation is applied. To include the BDRP in the SearchCol framework, [Barbosa et al., 2013] proposed a new decomposition model to the problem. In the particular case of the BDRP decomposition model presented, a new class is included in the framework programming code which is responsible to initialize the various subproblems and the restricted master problem with the model constraints and with the data from particular instances. Given an implementation of a decomposition, the framework can run the column generation method over it.

When the column generation ends, the framework has in its data structures valuable information. For each column added to the restricted master problem, the subproblem solution used to create it is stored, including the identification of the subproblem that produced the solution, the identification of the column added and also the value of the solution. In the BDRP decomposition, each subproblem solution is a valid work-schedule for the driver associated to the subproblem.

The complete set of subproblem solutions defines the search space available to the metaheuristics

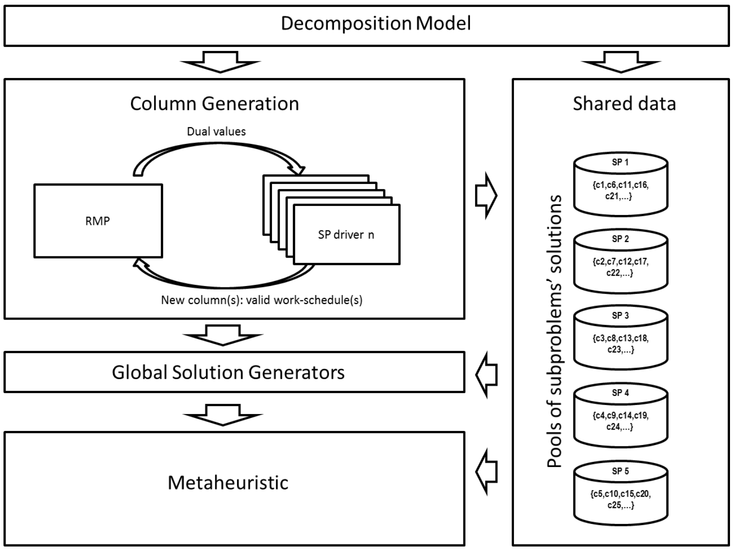


Figure 1: SearchCol base components.

to perform a search on. Furthermore, the solution of the restricted master problem, which is a linear solution, composes the solution for each subproblem by the convex combination of columns generated by that subproblem. The values of the linear solution of the final RMP are used to distinguish the subproblem solutions available in the metaheuristics search space. Subproblems solutions included in the final RMP linear solution are potentially better than the others.

For all the metaheuristics that need a starting global solution (which consists in selecting a solution from each subproblem) the framework includes distinct generators. Two of them select randomly the solution of each subproblem from the set of solutions created by that subproblem. The difference between them is that in one the probability of each solution to be chosen is uniform and on the other the probability of each solution to be chosen is biased by the value of the linear solution. All generators available are described in [Alvelos et al., 2013].

Figure 1 presents a simplified graphical view of the main components of the SearchCol framework. As mentioned above, to use the framework to solve a problem an implementation of a decomposition model of the problem is needed. Since the decomposition model initialize all objects, the column generation starts to optimize the problem storing the subproblems' generated solutions into the solutions pool. When the column generation stops, one or more global solution generators can use the pool of solutions and the last optimal solution of the RMP to compose an initial global integer solution or a population of global solutions, depending on the metaheuristic. The SearchCol framework offers other components, particularly a perturbation generator that uses all information available after the metaheuristic conclusion to add perturbations (new constraints) into the RMP by forcing the generation of new columns when column generation is restarted, allowing the repetition of the metaheuristic with a different search space.

3 Embedding Genetic Algoritms in the SearchCol with the BDRP decomposition model

In this section we present the use of Genetic Algorithms as the metaheuristic used by SearchCol to obtain global integer solutions. The use of Genetic Algorithms in the framework was introduced in [Barbosa et al., 2013]. We present an overview of the relevant aspects of the previously proposed implementation and then we introduce our current contribution to improve the Genetic Algorithms behaviour. Along the section we use the BDRP decomposition model in order to better explain the components of the algorithm but, we emphasize that they can be used with other problems included in the framework.

3.1 Genetic Algorithms core implementation

The first issue when using Genetic Algorithms (GA) to optimize solutions [Reeves, 1997] is to decide how to represent a solution, referred as chromosome in the GA concepts. Since the core of SearchCol metaheuristics is the combination of solutions from the different subproblems, the format of the representation

C1	C22	C8	C19	C10

Figure 2: GA Chromosome representing a global solution.

considers a global solution as a tuple with size equal to the number of subproblems. In the GA notation, each gene of the chromosome corresponds to a subproblem solution, and the locus of the gene identifies the subproblem (or driver).

In Figure 2 an illustrative representation of a five gene chromosome defines a global solution relying on the solutions of five subproblems. The global solution is composed by solutions 1,22,8,19 and 10 from the pools of solutions of subproblems 1 to 5, respectively.

The GA needs an initial population to evolve trough generations by using variation operators. A new generator was implemented in the SearchCol framework to allow the generation of multiple individual, depending on the population size used by the GA. This generator is parameterized to define the number of individuals generated with each global solution generator already available in the framework.

Since each subproblem solution has only the certainty to be valid to the subproblem that generated it, the variation operators proposed in [Barbosa et al., 2013] are the simple crossover operator (one and two points) and a mutation operator. A valid work-schedule for one driver can be invalid for another because information from the previous rostering period is considered for each driver. The mutation operator replaces the subproblem solution for one or more genes randomly selected by other from the same subproblem randomly chosen from the pool of solutions (shared data in Figure 1). The selection operator used to build the mating population is the tournament.

One of the most important components of the GA is the fitness function. The global solutions implementation class in the SearchCol has predefined attributes to store the feasibility and infeasibility values of a global solution. These values are normally associated with the total cost of the global solution and with the number of constraints violated in the RMP. In the BDRP, the infeasibility value counts the number of tasks not assigned and the feasibility value represents the cost of the roster. Since each subproblem solution stores its contribution (cost) to the global objective function, the total cost is easy to obtain simply summing the costs of work-schedules. Counting of unassigned tasks is also easy to obtain because all subproblems have the same number of variables, and, for each position, the simple sum of all the solution vectors (for all the drivers) identifies unassigned tasks if the sum in a position is zero. The count of the zeros gives the number of unassigned tasks.

3.2 Improving Genetic Algorithms with Elitism and Local Search

The results presented in [Barbosa et al., 2013] revealed that the GA were effective in the search for valid rosters, however, when compared with solutions obtained by the MIPSearch, they were not competitive. To improve the results of the GA, we now present the implementation of an elitism strategy and a local search in the GA. Elitism is known to increase convergence [Seijas et al., 2006, Deb and Goel, 2001]. Local search [Johnson et al., 1988] explores neighboring solutions by testing small changes in the solutions until an improvement is achieved or all neighbor solutions were tested. A neighbor solution is a solution with a different value in a limited number of variables.

The use of elitism in our current implementation is configured by a parameter defining the size of the elite pool (from now on referred to as α) used to store the best individuals found during the evolution process. If the parameter is zero, the elitism is not used. If using elitism, in each iteration the α best individuals are moved from the population to the elite pool. After, when applying the selection operator to build the mating pool, another parameter defines if the source of the individuals selected to tournament is the population or the elite pool. The variation operators are applied filling up the new population to *population size* - α, the remaining space is filled with the elite pool individuals.

At the end of each iteration, a simple local search can be applied to the best individual of the current iteration. The local search behaviour is similar to the mutation operator, but it does not change the original solution. The local search selects randomly a gene of the chromosome (equivalent to select a driver) and then other solutions from the subproblem corresponding to that gene are tested to replace the original until a better solution is found or until all solutions of that subproblem were tested. If a better solution is found, that solution is added to the next population, increasing temporarily its size by one. The final algorithm is represented by an activity diagram in Figure 3.

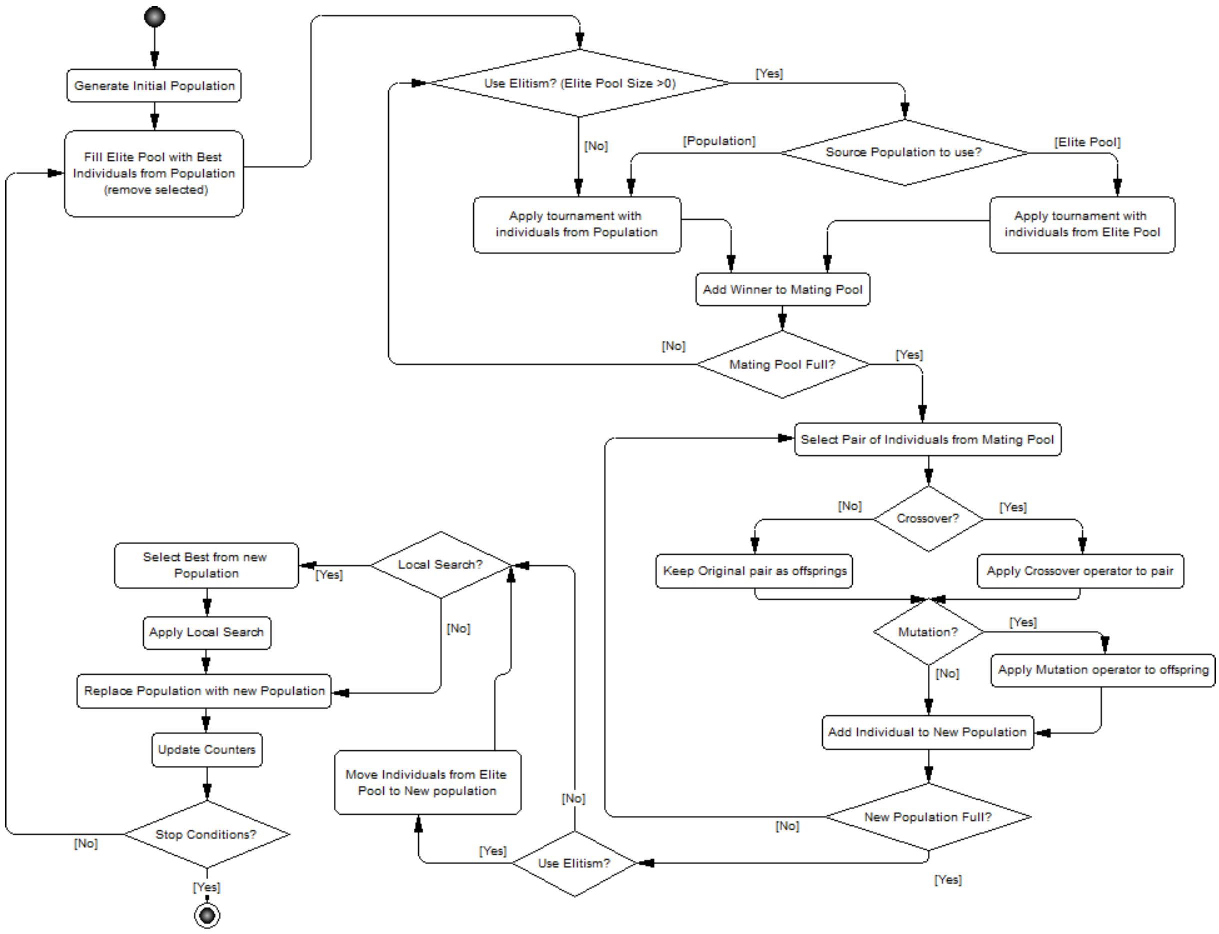


Figure 3: New GA algorithm represented with an activity diagram.

4 Computational Tests

To evaluate the impact of the current improvement to the GA a new set of tests were run with similar conditions of the tests done in [Barbosa et al., 2013]. The instances used in the computational tests are those in the set designated as P80 in [Moz et al., 2009]. The set of instances (P80_0 to P80_10) use a pool of 36 drivers. In all the instances, the group of drivers is divided in four categories of overtime cost, each group includes 9 drivers, starting from a cost factor of one and doubling for the next group, resulting that the last group overtime cost is 8 times more expensive than the first.

For each instance the tasks to assign are the same in each week day (Monday to Friday) and also in the weekend days. The number of tasks in each of the instances is presented in Table 1. Since the proposed model considers a four week rostering period, the total number of tasks is obtained considering 20 week days and 8 weekend days. The tasks count in the week and weekend days as well as the total number of tasks during all the rostering period are also shown.

The following parameters are common to all instances:

Number of available drivers	36
Maximum number of consecutive working days	6
Minimum days off by week	2
Minimum days off in a Sunday (in the rostering period)	1
Maximum total number of hours of work by week	48
Maximum total number of hours of work by rostering period	176
Contractual number of work days in the rostering period	20
Contractual number of work hours by day (before overtime)	8
Fixed cost of using a driver	100

Table 1: Number of tasks to assign in each test instance

Instance	**Week Day Tasks**	**Weekend Day Tasks**	**Total number of tasks**
P80_0	28	16	408
P80_1	18	12	296
P80_2	13	9	396
P80_3	17	14	388
P80_4	17	12	372
P80_5	16	13	348
P80_6	15	12	408
P80_7	18	12	484
P80_8	21	16	380
P80_9	17	10	516
P80_10	23	14	408

4.1 Genetic Algorithm Parameters

The configuration of parameters used in the Genetic Algorithm keeps the same values used by [Barbosa et al., 2013]. Below the value of each parameter is presented. The symbol ∗ identifies new parameters introduced in the current contribution.

GA configuration parameters:

Population Size	200
Crossover Probability	80%
Mutation Probability	15%
Stopping Criterion	5000 generations without improvement
Elite Pool Size ∗	20 (10% of Population Size)
Percentage of Individuals Selected from Elite Pool ∗	15%
Use Local Search ∗	Yes

Each initial population is composed by:

- 40% of individuals selected randomly from the pools of work-schedules available with uniform distribution for each schedule;
- 30% of individuals selected randomly from the pools of work-schedules available with biased distribution for each schedule according to the optimal solution of the RMP;
- 10% of individuals composed by rounding the linear solution of the RMP;
- 10% of individuals composed by the first solution generated by column generation;
- 10% of individuals composed by the last solution generated by column generation.

4.2 Test Conditions

To test the new GA, the same configurations of the two sets of tests presented in [Barbosa et al., 2013] were run. Basically two configuration of the SearchCol are tested where the difference between them is in the search space generation. In the first configuration, from now on referred to as GA1, in each iteration of the column generation a work-schedule from each driver is added to the RMP (if attractive) and in the second one, from now on referred to as GA2, only one schedule is added from a single driver in each iteration, changing sequentially the driver through the iterations.

Both configurations were tried because generating columns for all subproblems in each iteration results in a fast growth of the number of variables in the RMP, making it harder to optimize, but it also results in a wider search space with similar solutions to all drivers which may be useful to the GA. In [Barbosa et al., 2013] GA2 was able to obtain valid rosters for more instances, but GA1 achieved better solution values in the difficult instances.

All the tests ran on a Dell Optiplex 380 with an Intel Core 2 Duo CPU E7500, 2,93GHz, 4 Gb of RAM, operating system Windows Vista 32 bits and IBM ILOG 12.3 installed.

For each instance test, the column generation time limited was set to 1800s. The results of the GA without use of elitism are from [Barbosa et al., 2013] where 10 runs of the GA algorithm were done with the same search space. In the current configuration (with use of elitism and local search) 30 runs of the GA algorithm were done with the same search space.

4.3 Results

The results of the GA1 and GA2 without using any elitism strategy (from [Barbosa et al., 2013]) are presented in Table 2. Table 3 presents the results for the new GA version with an elitism strategy implemented and a local search procedure implemented, also for both configurations 1 and 2, GAEL1 and GAEL2.

Table 2: GA1 and GA2 results

	GA1					GA2				
	Infeasibility		Feasibility		Time	Infeasibility		Feasibility		Time
Instance	Best	Average	Best	Average	Average	Best	Average	Best	Average	Average
P80_0	29	83,6	7748	7316,4	223,93	123	134,6	6251	6376,7	117,17
P80_1	0	13,2	7970	7096,2	90,97	0	7,3	8324	8325,1	78,69
P80_2	0	0	5854	6743,9	52,56	0	0	5625	5697,2	31,93
P80_3	9	13,3	8919	9494,1	87,34	0	5,3	9329	10865,4	64,25
P80_4	5	11,1	7107	6798,1	77,39	0	2,2	7986	8583,2	64,68
P80_5	3	7,6	7820	7594,5	83,95	0	4,2	7600	8582,7	61,43
P80_6	2	3,5	8216	8304,9	66,21	0	0	8481	9156,6	38,10
P80_7	13	18	6745	6840,2	96,03	2	5,2	9343	8757,9	45,90
P80_8	31	50,6	7711	6769,4	90,05	32	53,2	8605	6891,7	67,42
P80_9	0	2,9	7017	7344	74,63	0	0,2	6661	7287	56,85
P80_10	12	53,3	6188	6026,4	132,52	43	58	6219	5923,8	99,29

Table 3: GAEL1 and GAEL2 results

	GAEL1					GAEL2				
	Infeasibility		Feasibility		Time	Infeasibility		Feasibility		Time
Instance	Best	Average	Best	Average	Average	Best	Average	Best	Average	Average
P80_0	15	19	8036	7995,6	64,58	103	127,10	7377	6756,5	57,68
P80_1	0	0	4622	5204,1	62,80	0	1,07	6805	7986,9	40,05
P80_2	0	0	3268	4030,9	45,75	0	0,00	5625	5660,5	21,56
P80_3	0	0	6401	7499,4	67,08	0	0,53	9127	10057,0	40,11
P80_4	0	0,33	4652	5952,7	64,13	0	0,23	7235	8382,8	35,43
P80_5	0	0	4420	5004,4	64,91	0	0,20	6802	7898,1	36,89
P80_6	0	0	4496	5296,4	56,34	0	0,00	7248	8186,4	34,69
P80_7	0	0	5395	6398,6	71,69	0	2,27	8428	8484,3	36,23
P80_8	0	0,07	6499	7209,8	59,78	8	38,73	8835	7375,4	46,58
P80_9	0	0	4360	5049,3	63,78	0	0,03	5864	6668,0	38,78
P80_10	0	0,1	4914	6216,6	71,58	15	38,53	6791	6298,7	46,05

Comparing the two tables, the effect of the new features is evident. For all instances the infeasibility values decreased or remained equal when this value was already zero. In addition, for all the cases where a feasible solution was attained, the cost of the best solution also improved (displayed in columns "Feasibility/Best"). It only makes sense to compare feasibility values for the cases presenting the same value of infeasibility, since if a solution has more unassigned tasks, and those tasks have overtime costs, their assignment will increase the cost of the solution. As an example, instance P80_0 had lower infeasibility values (best and average) on the new GA, but the cost of the solutions is higher, which is natural because more tasks are assigned (perhaps with overtime).

The global improvement of the solutions was predictable, however, in the earlier GA version, the GA2 was able to obtain valid rosters (with infeasibility value equal to zero) for more instances than GA1 and the introduction of elitism lead to an inversion. The GAEL1 obtained valid rosters on 10 of 11 instances, and in 7 of them it was achieved in all the runs, therefore outperforming GAEL2. In addition, for all instances where feasibility was reached, P80_1 to P80_10, the best solutions obtained by GAEL1 are always better than those produced by GAEL2. As before configuration two was faster due to a smaller computing time consumed during column generation, but for the new version GAEL1 always reached better solutions. The domination of the solutions obtained by GAEL1 is easily observable in Figure 4 and Figure 5 where the best values and the average values, respectively, are compared for both configurations (GAEL1 vs GAEL2).

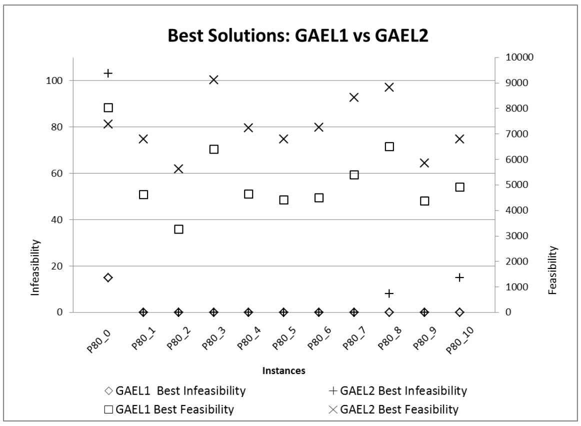


Figure 4: Best solutions comparison for both configurations of GAEL

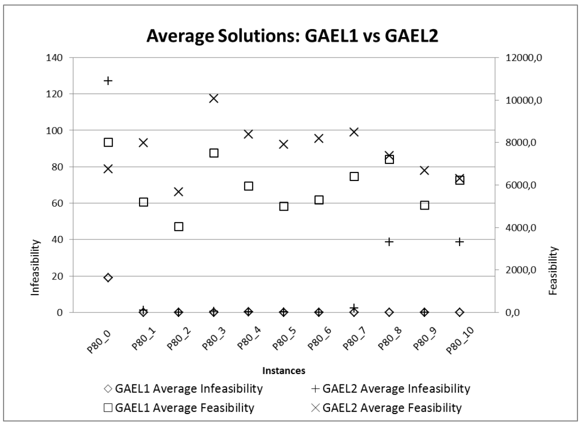


Figure 5: Average solution values comparison for both configurations of GAEL

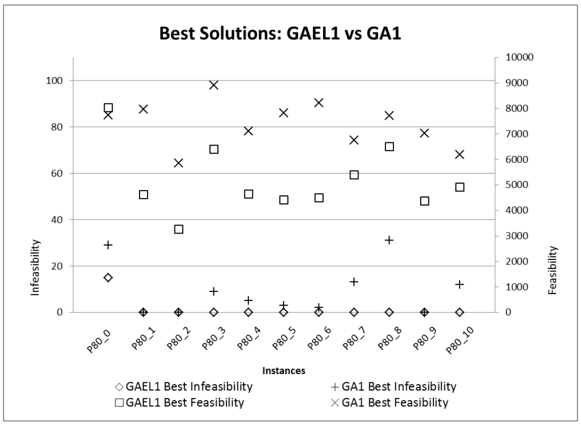


Figure 6: Best solution values of GA1 and GAEL1

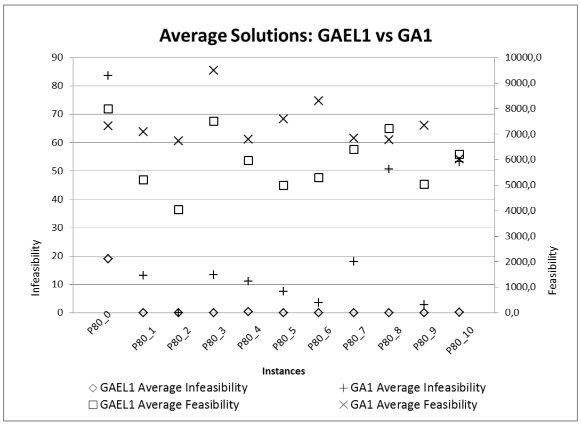


Figure 7: Average solution values of GA1 and GAEL1

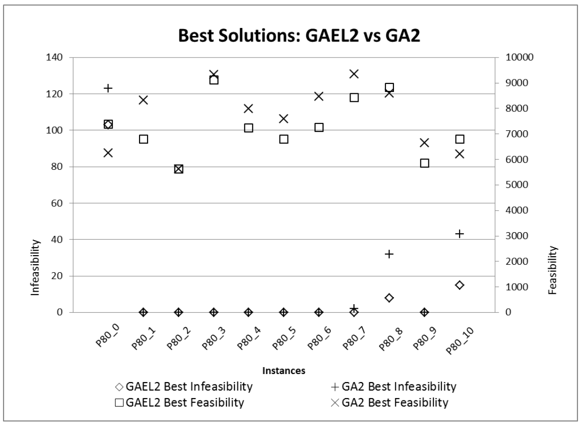


Figure 8: Best solution values of GA2 and GAEL2

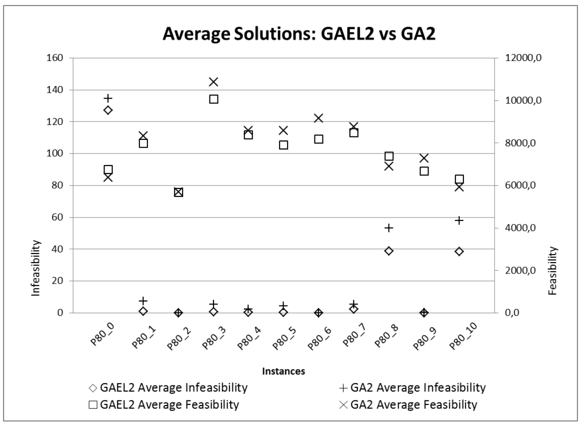


Figure 9: Average solution values of GA2 and GAEL2

In Figure 6 and Figure 7 the improvement of the best solution and the average of the solution values of the first configuration, respectively, are compared for the GA with and without the new features implementation. Figure 8 and Figure 9 show the same comparison for the second configuration.

The charts in Figures 6 and 7 disclose the greater impact of the new features in the first configuration. It is noteworthy the number of valid rosters obtained and the improvement on the solution costs.

Table 4: MIPSearch Results

Instance	Feasibility	Infeasibility	Time Search	Time CG
P80_0	-	93	5400	1800
P80_1	6477	0	5400	1800
P80_2	4628	0	6815,2	384,7
P80_3	8762	0	5400	1800
P80_4	6809	0	5400	1800
P80_5	6648	0	5400	1800
P80_6	7182	0	5400	1800
P80_7	6819	0	5400	1800
P80_8	-	12	5400	1800
P80_9	5599	0	5563,2	1636,4
P80_10	-	3	5400	1800

The SearchCol framework provides the MipSearch procedure, which proceeds to direct optimization of the resulting RMP setting the variables as binary. MipSearch considers the available columns to search for an optimal solution and therefore allows evaluating the space obtained by the column generation. MipSearch was already used with our test instances set, considering a total computing time limited to 7200s (2 hours) and the column generation computing time limited to 1800s (1/2 hour). Table 4 presents the results obtained in [Barbosa et al., 2013]. In the cases where the CG stops before the time limit, as happens in instances P80_2 and P80_9 , the remaining time can be used by the metaheuristic. A direct comparison of the best solutions found by GAEL1 with those obtained by MipSearch also reveals that GAEL1 outperforms a direct optimization procedure. All the numbers are better: infeasibility values, solutions costs and computing time. Again, GAEL1 has shown evidence of its competence in attaining good quality global solutions for the BDRP within the framework.

5 Conclusions

This paper proposes a new version of the genetic algorithm included in the SearchCol++ framework. The first addition considers the use of an elitism strategy to keep best individuals through generations and force their presence in the mating pool. The new version also allows the use of a local search procedure to explore the neighborhood of the best solution in each generation of the genetic search.

The genetic algorithm is used in the BDRP to obtain valid rosters by combining individual drivers' work-schedules. A pool of diverse valid work-schedules for all the drivers is obtained by applying column generation over a decomposition model for the problem. This concept is the core of the SearchCol framework.

The new features were tested and the results showed generalized improvements. These improvements were more prominent in the configuration where the column generation process adds attractive columns for all drivers in each iteration, hence producing a larger search space for the GA. In that scenario, the algorithm found valid solution for 10 of the 11 instances and for 7 of them this happened in all the runs.

Current results reinforce the ability of the GA to obtain valid rosters using individual solutions produced by the column generation.

Future work will include testing different parameters' settings for the GA components and refinements of the local search procedure.

Despite the current implementation of our GA is included in the SearchCol++ framework, it can be used together with other implementations of the column generation method since it is guaranteed to keep the information needed by the algorithm to evaluate the subproblems solutions efficiently and additional methods to build initial populations are provided. The use of our GA implementation can also be extended to other decomposition models for different problems, provided that global solutions are the combination of single solutions from the subproblems.

Acknowledgements

This work was partially funded by projects PTDC/EIA-EIA/100645/2008 ("SearchCol-Meta-heuristic Search by column generation" / FEDER through "COMPETE - Programa Operacional Factores de Competitividade" and FCT - Foundation for Science and Technology) and PEst-OE/MAT/UI0152 (FCT - Foundation for Science and Technology).

References

[Alvelos et al., 2010] Alvelos, F., Sousa, A., and Santos, D. (2010). Searchcol: Metaheuristic search by column generation. In Blesa, M., Blum, C., Raidl, G., Roli, A., and Sampels, M., editors, *Hybrid Metaheuristics*, volume 6373 of *Lecture Notes in Computer Science*, pages 190–205. Springer Berlin Heidelberg.

[Alvelos et al., 2013] Alvelos, F., Sousa, A., and Santos, D. (2013). Combining column generation and metaheuristics. In Talbi, E.-G., editor, *Hybrid Metaheuristics*, volume 434 of *Studies in Computational Intelligence*, pages 285–334. Springer Berlin Heidelberg.

[Barbosa et al., 2013] Barbosa, V., Respício, A., and Alvelos, F. (2013). A hybrid metaheuristic for the bus driver rostering problem. In Vitoriano, B. and Valente, F., editors, *ICORES 2013 - Second International Conference on Operations Research and Enterprise Systems*, pages 32–42, Barcelona. SCITEPRESS.

[Burke et al., 2003] Burke, E. K., Kendall, G., and Soubeiga, E. (2003). A tabu-search hyperheuristic for timetabling and rostering. *Journal of Heuristics*, 9(6):451–470.

[Dantzig and Wolfe, 1960] Dantzig, G. B. and Wolfe, P. (1960). Decomposition principle for linear programs. *Operations Research*, 8(1):101–111.

[Deb and Goel, 2001] Deb, K. and Goel, T. (2001). Controlled elitist non-dominated sorting genetic algorithms for better convergence. In Zitzler, E., Thiele, L., Deb, K., Coello Coello, C., and Corne, D., editors, *Evolutionary Multi-Criterion Optimization*, volume 1993 of *Lecture Notes in Computer Science*, pages 67–81. Springer Berlin Heidelberg.

[Desaulniers et al., 2005] Desaulniers, G., Desrosiers, J., and Solomon, M. M. (2005). *Column Generation*. Springer, New York.

[Dorne, 2008] Dorne, R. (2008). Personnel shift scheduling and rostering. In Voudouris, C., Lesaint, D., and Owusu, G., editors, *Service Chain Management*, pages 125–138. Springer Berlin Heidelberg.

[Ernst et al., 2004] Ernst, A. T., Jiang, H., Krishnamoorthy, M., and Sier, D. (2004). Staff scheduling and rostering: A review of applications, methods and models. *European Journal of Operational Research*, 153(1):3–27.

[Holland, 1992] Holland, J. H. (1992). *Adaptation in natural and artificial systems*. MIT Press.

[Johnson et al., 1988] Johnson, D. S., Papadimitriou, C. H., and Yannakakis, M. (1988). How easy is local search? *Journal of Computer and System Sciences*, 37(1):79 – 100.

[Leone et al., 2011] Leone, R., Festa, P., and Marchitto, E. (2011). A bus driver scheduling problem: a new mathematical model and a grasp approximate solution. *Journal of Heuristics*, 17(4):441–466.

[Lucic and Teodorovic, 2007] Lucic, P. and Teodorovic, D. (2007). Metaheuristics approach to the aircrew rostering problem. *Annals of Operations Research*, 155(1):311–338.

[Moz et al., 2009] Moz, M., Respício, A., and Pato, M. (2009). Bi-objective evolutionary heuristics for bus driver rostering. *Public Transport*, 1(3):189–210.

[Nurmi et al., 2011] Nurmi, K., Kyngäs, J., and Post, G. (2011). Driver rostering for bus transit companies. *Engineering Letters*, 19(2):125–132.

[Reeves, 1997] Reeves, C. R. (1997). Genetic algorithms for the operations researcher. *INFORMS Journal on Computing*, 9(3):231–250.

[Ruibin et al., 2010] Ruibin, B., Burke, E. K., Kendall, G., Jingpeng, L., and McCollum, B. (2010). A hybrid evolutionary approach to the nurse rostering problem. *Evolutionary Computation, IEEE Transactions on*, 14(4):580–590.

[Seijas et al., 2006] Seijas, J., Morató, C., and Sanz-González, J. (2006). Genetic algorithms: Two different elitism operators for stochastic and deterministic applications. In Wyrzykowski, R., Dongarra, J., Paprzycki, M., and Waniewski, J., editors, *Parallel Processing and Applied Mathematics*, volume 2328 of *Lecture Notes in Computer Science*, pages 617–625. Springer Berlin Heidelberg.

[Xie et al., 2012] Xie, L., Kliewer, N., and Suhl, L. (2012). Integrated driver rostering problem in public bus transit. *Procedia - Social and Behavioral Sciences*, 54(0):656 – 665. Proceedings of {EWGT2012} - 15th Meeting of the {EURO} Working Group on Transportation, September 2012, Paris.

A Comparative Study of Two Optimization Clustering Techniques on Unemployment Data

Elisa Barros*, Alcina Nunes*, Carlos Balsa*

* *Escola Superior de Tecnologia e Gestão, Instituto Politécnico de Bragança, Portugal*
E-mail: `ebarros@ipb.pt`, `alcina@ipb.pt`, `balsa@ipb.pt`

Abstract

An important strategy for data classification consists in organising data points in clusters. The k-means is a traditional optimisation method applied to cluster data points. Using a labour market database, we suggest the application of an alternative method based on the computation of the dominant eigenvalue of a matrix related with the distance among data points. This approach presents results consistent with the results obtained by the k-means.

Key words: Clustering methods, k-means, spectral clustering, unemployment data mining

1 Introduction

Clustering is an important process for data classification that consists in organising a set of data points into groups, called clusters. A cluster is a subset of an original set of data points that are close together in some distance measure. In other words, given a data matrix containing multivariate measurements on a large number of individuals (observations or points), the aim of the cluster analysis is to build up some natural groups (clusters) with homogeneous properties out of heterogeneous large samples [Kaufman and Rousseeuw, 1990].

Groups are based on similarities. The similarity depends on the distance between data points and a reduced distance indicates that they are more similar. Several distinct methods can be used to measure the distance among the elements of a data set. Along this work we will consider the traditional Euclidian distance, i.e., the 2-norm of the differences between data points vectors.

There are two main classes of clustering techniques: hierarchical and optimization methods. In hierarchical clustering is not necessary to know in advance the number of subsets in which we want to divide the data. The observations are successively included in groups of different dimensions depending on the level of clustering. The result is a set of nested partitions. In each step of the process, two groups are either merged (agglomerative methods) or divided (divisive methods) according to some criteria [Matinez et al, 2011]. In the agglomerative approach, single-members clusters (clusters with only one observation) are increasingly fused until all observations are in only one cluster. The divisive approach starts with a single set containing all points. This group will be increasingly divided as the distance between points is reduced. The set of nested partitions is represented graphically by a dendrogram that has a tree shape indicating the distance′s hierarchical dependence. The dendrogram can help to identify the number of clusters that should be considered during the partition of the data set. After that a partition method like k-means can be applied to identify each cluster.

The k-means [MacQueen, 1967] is an optimization method that partitions the data in exactly k clusters, previously determine. This is achieved in a sequence of steps which begins, for instance, with an initial partition randomly generated. In each step the clusters centroid (arithmetic vector mean) is computed. The minimum distance between each data point and the clusters′ different centroids will decide the formation of new clusters. The formation of a new cluster implies assigning each observation to the cluster which presents the lowest distance. After that the centroids are (re)calculated and the former step is repeated until the moment each individual belongs to a stable cluster, i.e., when the sum of the squared distances to the centroid of all data point over all the clusters is minimized. The algorithm presents a rather fast convergence, but one cannot guarantee that the algorithm finds the global minimum [Eldén, 2007].

Spectral clustering is also an optimization method. This method is becoming very popular in recent years because it has been included in algorithms used in the identification of the human genome or in web browsers. Beyond biology and information retrieval the method has other fields of application such as image analysis and, in some cases, it can perform better than standard algorithms such as k-means and hierarchical clustering [Matinez et al, 2011]. Spectral clustering methods use the k dominant eigenvectors of a matrix, called affinity matrix, based on the distance between the observations. The idea is grouping data points in a lower-dimensional space described by those k eigenvectors [Mouysset et al, 2008]. The approach may not make a lot of sense, at first, since we could apply the k-means methodology directly without going through all the matrix calculations and manipulations. However, some analyses show that mapping the points to this k-dimensional space can produce tight clusters that can easily be found applying k-means [Matinez et al, 2011].

In the present research work, spectral clustering is applied in an unusual context concerning the traditional data mining analysis. We classify 278 Portuguese mainland municipalities (*concelhos*) regarding the type/characteristics of unemployment official registers. The set of observations, $x_1, \ldots, x_{278}$, that contains 278 vectors, whose 11 coordinates are the values for some of the indicators used to characterise Portuguese unemployment (gender, age classes, levels of formal education, situation relating unemployment and unemployment duration), is divided in k clusters. The classification of observations resulting from the spectral method is than compared to the classification given by the traditional k-means method. The results are analysed from both mathematical and economic points of view. The main goal is to find evidence regarding which method produces the best cluster partition and, accordingly, to understand if the resulting clusterisation makes sense in terms of the spatial distribution of unemployment characteristics, over a country administrative territory. The idea is to understand if a particular cluster methodology for data mining analysis provides useful and suitable information that could be used to the development of national, regional or local unemployment policies.

The paper is divided as follows. The k-means method and the spectral clustering method are presented in sections 2 and 3, respectively. The methods description is followed by section 4 where data and variables analysed are also presented and described. In section 5 we move ahead toward the optimal number of clusters applying both selected methods. In section 6 the results are presented and discussed. Our concluding remarks can be found on section 7.

2 The k-means method

We are concerned with m data points $x_i \in \mathbb{R}^n$ that we want classify in k clusters, where k is predetermined. We organize the data as lines in a matrix $X \in \mathbb{R}^{m \times n}$. To describe the k-means method as proposed in [Eldén, 2007] we denote a partition of vectors $x_1, \ldots, x_m$ in k clusters as $\prod = \{\pi_1, \ldots, \pi_k\}$ where

$$\pi_j = \{\ell : x_\ell \in \text{cluster} j\}$$

defines the set of vectors in cluster j. The centroid, or the arithmetic mean, of the cluster j is:

$$m_j = \frac{1}{n_j} \sum_{\ell \in \pi_j} x_\ell \tag{1}$$

where n_j is the number of elements in cluster j. The sum of the squared distance, in 2-norm, between the data points and the j cluster′s centroid is known as the *coherence*:

$$q_j = \sum_{\ell \in \pi_j} \|x_\ell - m_j\|_2^2 \tag{2}$$

The closer the vectors are to the centroid, the smaller the value of q_j. The quality of a clustering process can be measured as the *overall coherence*:

$$Q\left(\prod\right) = \sum_{j=1}^{k} q_j \tag{3}$$

The k-means is considered an optimization method because it seeks a partition process that minimizes $Q(\prod)$ and, consequently, finds an optimal coherence. The problem of minimizing the *overall coherence* is NP-hard and, therefore, very difficult to achieve. The basic algorithm for k-means clustering is a two step heuristic procedure. Firstly, each vector is assigned to its closest group. After that, new centroids are computed using the assigned vectors. In the following version of k-means algorithm, proposed by [Eldén, 2007], these steps are alternated until the changes in the *overall coherence* are lower than a certain tolerance previously defined.

The k-means algorithm

1. Start with an initial partitioning $\prod^{(0)}$ and compute the corresponding centroid vectors $m_j^{(0)}$ for $j = 1, \ldots, k$. Compute $Q(\prod^{(0)})$. Put $t = 1$.
2. For each vector x_i find the closest centroid. If the closest centroid is m_p^{t-1} assign x_i to $\pi_p^{(t)}$.
3. Compute the centroids $m_j^{(t)}$ for $j = 1, \ldots, k$ of the new partitioning $\prod^{(t)}$.
4. If $\left|Q(\prod^{(t)}) - Q(\prod^{(t-1)})\right| <$ tol, stop; Otherwise $t = t + 1$ and return to step 2.

Since it is an heuristic algorithm there is no guarantee that k-means will converge to the global minimum, and the result may depend on the initial partition $\prod^{(0)}$. To avoid this issue, it is common to run it multiple times, with different starting conditions choosing the solution with the smaller $Q\left(\prod\right)$.

3 Spectral clustering method

Let $x_1, \ldots, x_m$ be a m data points set in a n-dimensional euclidian space. We want to group these m points in k clusters in order to have better within-cluster affinities and weaker affinities across clusters. The affinity between two observations x_i and x_j is defined by [Ng et al, 2002] as:

$$A_{ij} = \exp\left(-\frac{\|x_i - x_j\|_2^2}{2\sigma^2}\right) \tag{4}$$

where σ is a scaling parameter that determines how fast the affinity decreases with the distance between x_i and x_j. The appropriate choice of this parameter is crucial [Matinez et al, 2011]. In [Ng et al, 2002] we can find a description of a method able to choose the scaling parameter automatically.

The spectral clustering algorithm proposed by [Ng et al, 2002] is based on the extraction of dominant eigenvalues and their corresponding eigenvectors from the normalized affinity matrix $A \in \mathbb{R}^{m\times m}$. The components A_{ij} of A are given by equation 4, if $i \neq j$, and by $A_{ii} = 0$, if $i = j$. The sequence of steps in the spectral clustering algorithm is presented as follows:

The spectral clustering algorithm

1. Form the affinity matrix A as indicated in equation 4.
2. Construct the normalized matrix $L = D^{-1/2}AD^{-1/2}$ with $D_{ii} = \sum_{j=1}^{m} A_{ii}$.
3. Construct the matrix $V = [v_1 v_2 \ldots v_k] \in \mathbb{R}^{m\times k}$ by stacking the eigenvectors associated with the k largest eigenvalues of L.
4. Form the matrix Y by normalizing each row in the $m \times k$ matrix V (i.e. $Y_{ij} = V_{ij}/\left(\sum_{j=1}^{k} V_{ij}^2\right)^{1/2}$).
5. Treat each row of Y as a point in $\mathbb{R}^k$ and group them in k clusters by using the k-means method.
6. Assign the original point x_i to cluster j if and only if row i of matrix Y was assigned to cluster j.

4 Data description

The 278 data points represents the Portuguese continental *concelhos*. Each data point have 11 coordinates representing characteristics of the unemployed register individuals. Indeed, the unemployed individuals registered in the Portuguese public employment services of the *Instituto de Emprego e Formação Profissional (IEFP)* present a given set of distinctive characteristics related with gender, age, formal education, unemployment spell (unemployment for less than a year or more than a year)and situation related with the unemployment situation (unemployed individual looking for a first employment or for another employment). These characteristics are important determinants of unemployment and are important economic vectors regarding the development of public employment policies. Well targeted policies are more efficient, in terms of expected results, and avoid the waste of scarce resources.

A complete study of regional similarities (or dissimilarities) in a particular labour market, as the Portuguese, should not be limited by a descriptive analysis of the associated economic phenomena. It should also try to establish spacial comparison patterns among geographic areas in order to develop both national and regional public policies to fight the problem. Indeed high unemployment indicators and regional inequalities are major concerns for European policy-makers since the creation of European Union. However, even if the problem is known the policies dealing with unemployment and regional inequalities have been few and weak [Overman and Puga, 2002]. In Portugal, in particular, there are some studies that try to define geographic, economic and social homogeneous groups [Soares et al, 2003]. To the best of our knowledge, there are no studies that offer an analysis of regional unemployment profiles. Other economies are starting to develop this kind of statistical analysis using as a policy tool the cluster analysis methodology [Arandarenko and Juvicic, 2007, López-Bazo et al, 2005, Nadiya, 2008].

The data concerning the above mentioned characteristics are openly available in a monthly period base in the website of *IEFP* (http://www.iefp.pt/estatisticas/Paginas/Home.aspx). Additionally, the month of December gives information about the stock of registered unemployed individuals at the end of the respective year. In the case of this research work, data from unemployment registers in 2012 have been used. The eleven variables available to characterise the individuals and that have been used here are divided in demographic variables and variables related with the labour market. These variables are dummy variables, measured in percentage of the total number of register individuals in a given *concelho*, and describe the register unemployed as follows: 1: Female, 2: Long duration unemployed (individual unemployed for more than 1 year), 3: Unemployed looking for a new employment, 4: Age lower than 25 years, 5: Age between 25 and 35 years, 6: Age between 35 and 54 years, 7: Age equal or higher than 55 years, 8: Less than 4 years of formal education (includes individuals with no formal education at all), 9: Between 4 and 6 years of formal education, 10: Between 6 and 12 years of formal education and 11: Higher education.

Women, individuals in a situation of long duration unemployment, younger or older unemployed individuals and the ones with lower formal education are the most fragile groups in the labour market and, consequently, are the most exposed to unemployment situations. They are also the most challenging groups regarding the development of public employment policies.

5 Toward the optimal number of clusters

We begin by applying the k-means method to partition in k clusters the data points set $x_1, \ldots, x_m$, with $m = 278$ Portuguese mainland *concelhos* regarding the 11 chosen unemployment characteristics. As the optimal number of targeted groups is unknown *a priori*, we repeat the partition for $k = 2,\ 3,\ 4$ and 5 clusters.

To evaluate the quality of the results from the cluster methodology and to estimate the correct number of groups in our data set we resort the silhouette statistic framework. The silhouette statistic introduced by [Kaufman and Rousseeuw, 1990] is a way to estimate the number of groups in a data set. Given observation x_i, the average dissimilarity to all other points in its own cluster is denoted as a_i. For any other cluster c, the average dissimilarity of x_i to all data points in cluster c is represented by $\bar{d}\,(x_i, c)$. Finally, b_i denote the minimum of these average dissimilarities $\bar{d}\,(x_i, c)$. The *silhouette width* for the observation x_i is:

$$s_i = \frac{(b_i - a_i)}{\max\{b_i,\ a_i\}}. \tag{5}$$

The *average silhouette width* is obtained by averaging the s_i over all observations:

$$\bar{s}_i = \frac{1}{m}\sum_{i=1}^{m} s_i. \tag{6}$$

If the *silhouette width* of an observation is large it tends to be well clustered. Observations with small *silhouette width* values tend to be those that are scattered between clusters. The *silhouette width* s_i in Equation 5 ranges from -1 to 1. If an observation has a value close to 1, then it is closer to its own cluster than it is to a neighbouring one. If it has a *silhouette width* close to -1, then it is a sign that it is not very well clustered. A *silhouette width* close to zero indicates that the observation could just as well belong to its current cluster or one that is near to it.

The *average silhouette width* (equation 6) can be used to estimate the number of clusters in the data set by using the partition with two or more clusters that yield the largest average silhouette width [Kaufman and Rousseeuw, 1990]. As a rule of thumb, it is considered that an *average silhouette width* greater than 0.5 indicates a reasonable partition of the data, and a value less than 0.2 would indicate that the data do not exhibit a cluster structure [Matinez et al, 2011].

Figure 1 presents the *silhouette width* corresponding to the case of four different partitions of the data points set, this is, $k = 2,\ 3,\ 4$ and 5 clusters resulting from the application of the k-means method.

As it is possible to observe, the worst cases occur, clearly, when $k = 3$ and $k = 5$. For these cases, some clusters present negative values and others appear with small (even if positive) silhouette indexes.

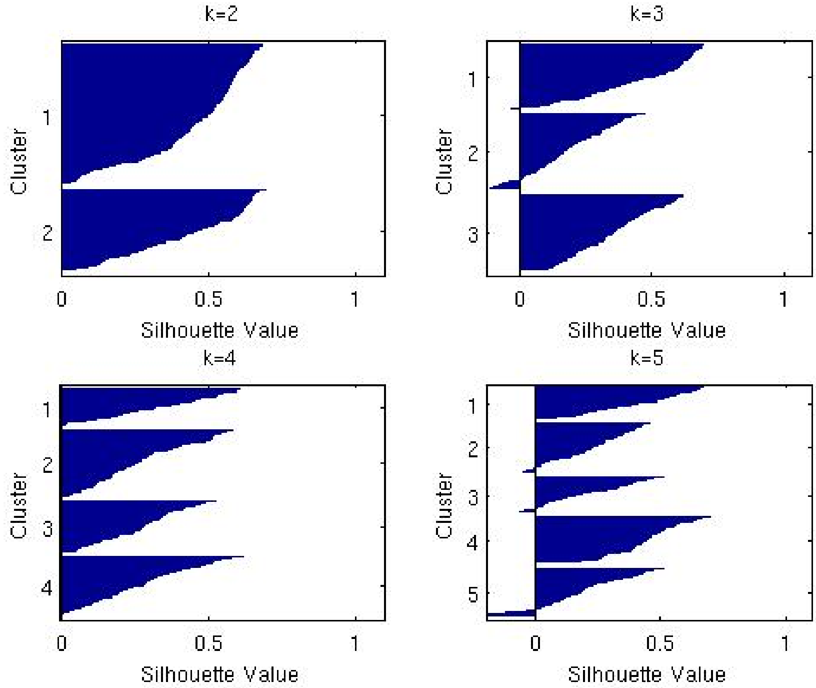


Figure 1: *Silhouette width* for $k = 2,\ 3,\ 4$ and 5 clusters resulting from the k-means method.

In the case of $k = 2$ and $k = 4$ clusters there are no negative values, however we find large silhouette values mostly in the case of the two clusters partition.

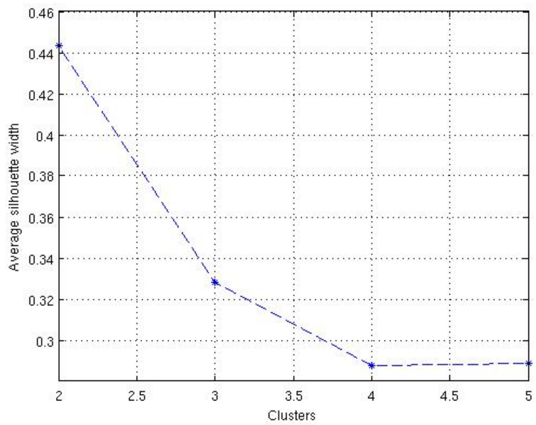


Figure 2: *Average silhouette width* for $k = 2,\ 3,\ 4$ and 5 clusters resulting from the k-means method.

To get a single number that is able to summary and describe each clustering process, we find the *average of the silhouette* values (equation 6) corresponding to $k = 2,\ 3,\ 4$ and 5. The results can be observed in figure 2.

The two cluster solution presents an average silhouette value near 0.44 and the four cluster solution presents an average silhouette value near 0.29. These results confirm the ones above. The best partition obtained with the application of the k-means method occurs with $k = 2$. Nonetheless, the *average of the silhouette* is close but smaller than 0.5 which reveals that the data set does not seem to present a strong trend to be partitioned in two clusters.

Figure 3 shows the *silhouette width* corresponding to each observation in the case of four different partitions of the data set points. This is, in $k = 2,\ 3,\ 4$ and 5 clusters, resulting from the application of the spectral clustering method.

In this case all the tested partitions present clusters where can be observed negative values. The worst

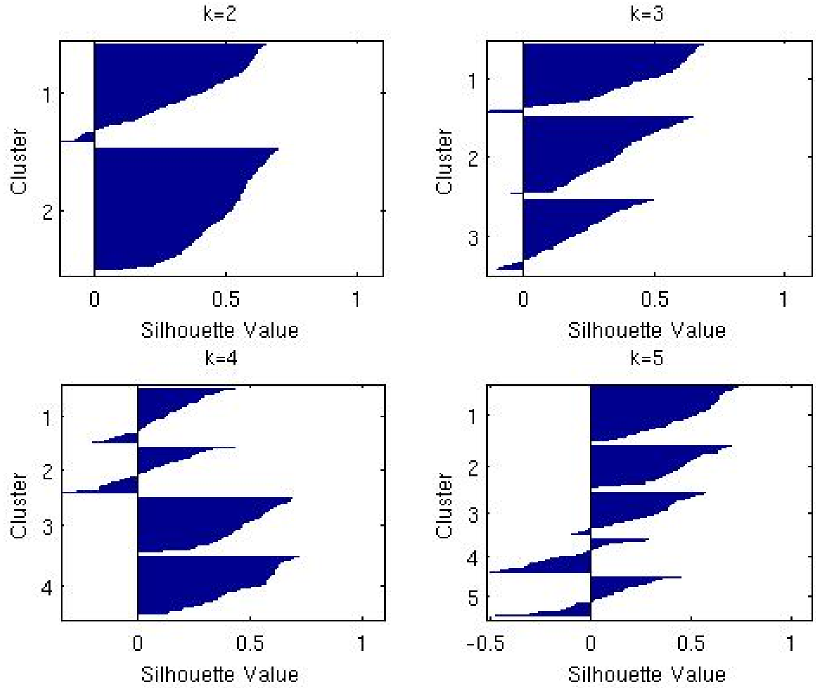


Figure 3: *Silhouette width* for $k = 2,\ 3,\ 4$ and 5 clusters resulting from the spectral method.

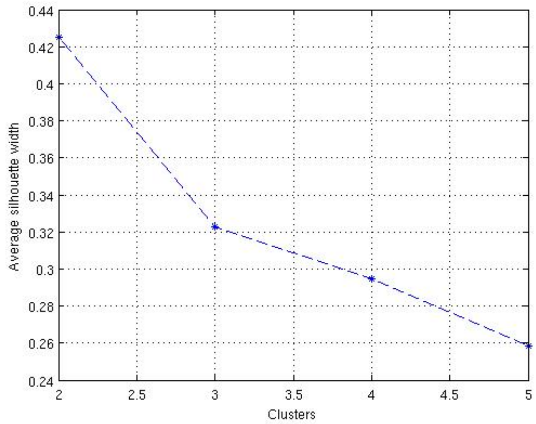


Figure 4: *Average silhouette width* for $k = 2,\ 3,\ 4$ and 5 clusters resulting from the spectral method.

cases occur, clearly, when $k = 4$ and $k = 5$. Here we get values close to -0.5. In the case $k = 3$ is possible to observe negative values in the three cluster obtained whereas in the case of $k = 2$ the negative values are just observed in one of the two clusters.

The trend observed with the *silhouette width* is confirmed by the *average of the silhouette* values corresponding to the spectral clustering process with $k = 2,\ 3,\ 4$ and 5 clusters (figure 4).

The two cluster solution has an average silhouette value near 0.43 and decrease as the number of clusters increases. The best partition obtained with the spectral clustering method occurs with $k = 2$. These results are in agreement with the partitioning found by using the k-means method. The *average of the silhouette* value (0.43) is very close to the one calculated with k-means method (0.44).

As mentioned before, the results obtained with the k-means method agree with the results obtained with the application of the spectral methods. The best partition of the data set is accomplished with two clusters. However, this trend is not completely crystal clear. Indeed, the *average of the silhouette* in the two cases is smaller than 0.5. The computed value indicates that the distance between the two considered clusters is not very large. This conclusion can be visually confirmed by the hierarchical distance between two data points. This distance is illustrated by the dendrogram presented on figure 5. In the dendrogram we can observe an initial bifurcation that divides the data in two main groups, but the distance between them is near 0.6. We can see also that one of the two main clusters can be partitioned into other two if we consider a distance near 0.5, very close to the distance between the two main clusters.

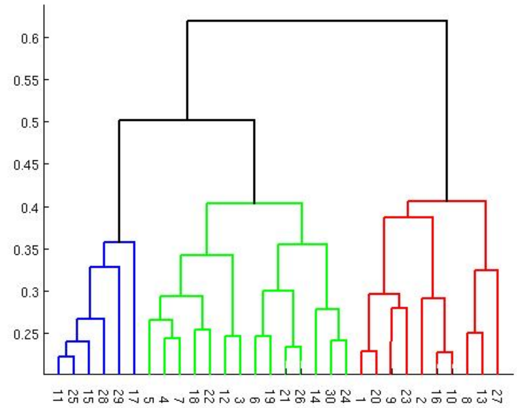


Figure 5: Dendrogram.

6 Mathematical and economic results′ analysis

Both spectral clustering method and k-means method indicate that the data are best partitioned into two clusters. The statistical properties of theses two clusters are presented in table 1.

method	j	n_j	q_j	Q
k-means	1	177	3.4161	5.3276
	2	101	1.9115	
spectral	1	154	2.7511	5.3536
	2	124	2.6026	

Table 1: Statistical properties of the two clusters resulting from k-means and spectral methods

Despite the number of observations in each cluster is not the same, it appears that for both methods the first cluster is the largest. This is, includes a bigger number of *concelhos*: $n_1 = 177$ for the k-means and $n_1 = 154$ for the spectral method. The difference of 23 observations for the first cluster is reflected in the computed local coherence q (equation 2) that is larger for the k-means methods ($q_1 = 3.4161$). The second cluster comprises $n_2 = 101$ observations and presents a local coherence of $q_2 = 1.9115$, for the k-means, and $n_2 = 124$ observations and a local coherence of $q_2 = 2.6026$ for the spectral method. Although the differences between the computed coherence for each cluster, we can observe that both methods achieve a very similar overall coherence (equation 3), $Q \approx 5.3$ for the k-means and $Q \approx 5.4$ for the spectral method.

For a more complete comparison analysis of the results obtained by k-means and spectral methods, it is also important to analyse two distribution measures: mean and standard deviation. The measures are presented for each one of the 11 variables used in the cluster analysis. In figure 6 we compare the mean value obtained for the 11 parameters that characterise the two clusters obtained by the two clusterisation methods. In figure 7 we compare the standard deviation value. Note that in these two figures the comparison analysis is done regarding the cluster methods applied.

It is visible that the computed mean values, regarding each one of the variables, are very similar in the two clusters independently of the cluster method used. For the computed standard deviation values we can observe a first cluster where the standard deviation, for the overall set of variables, are slightly higher for the k-means and a second cluster where the observed trend is reversed. In short, we can observe that the results for both methods are similar regarding the measure of central tendency of each one of the variables but the variability of values, regarding the central tendency, differ between cluster methods.

The mean and standard deviation measures can be compared regarding the values computed by cluster. From this point of view the analysis would have an economic focus. So, in figure 8 we compare

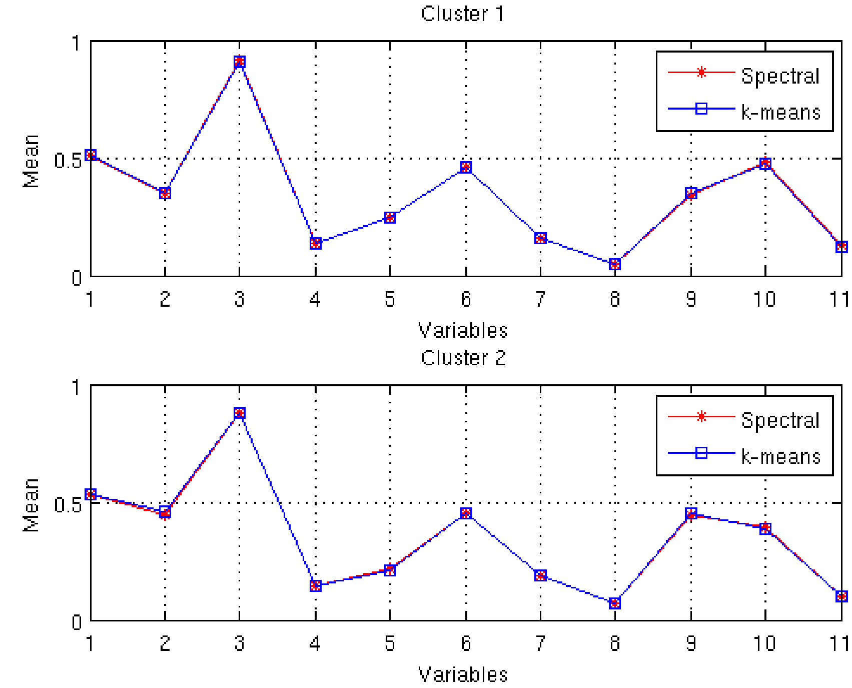


Figure 6: Mean values computed for the two clusters methods by cluster.

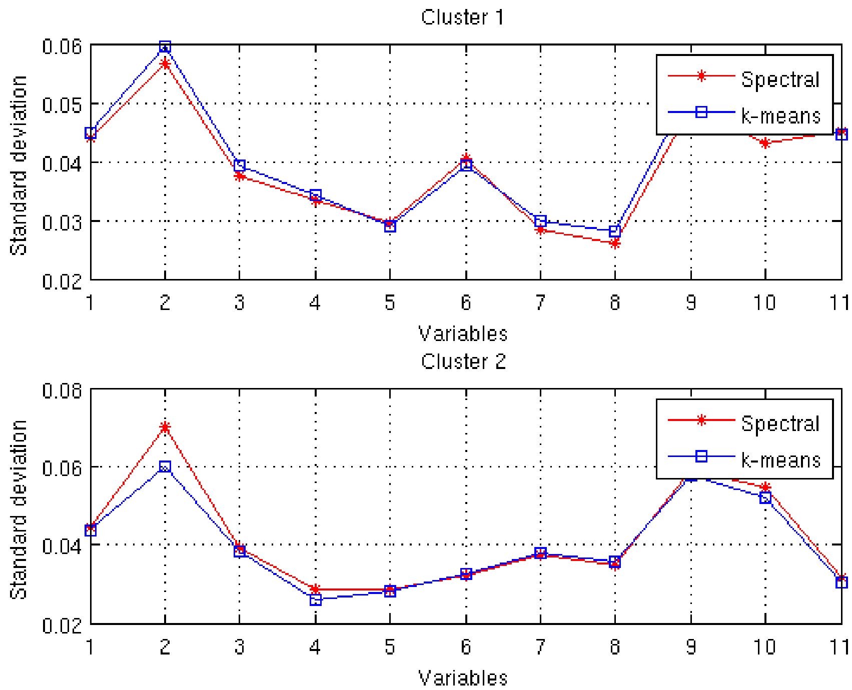


Figure 7: Standard deviation values computed for the two clusters methods by cluster.

the mean value obtained for the 11 parameters for each one of the clusters by cluster method. In figure 9 we compare the computed standard deviation value.

From the figure 8 and figure 9 it is possible to observe that both methods retrieve clusters that present the same pattern. In the second cluster (cluster 2) are gathered the Portuguese mainland *concelhos* that present a higher percentage of unemployed register individuals with more problematic characteristics - women, long duration unemployed individuals, individuals that are looking for a job for the first time

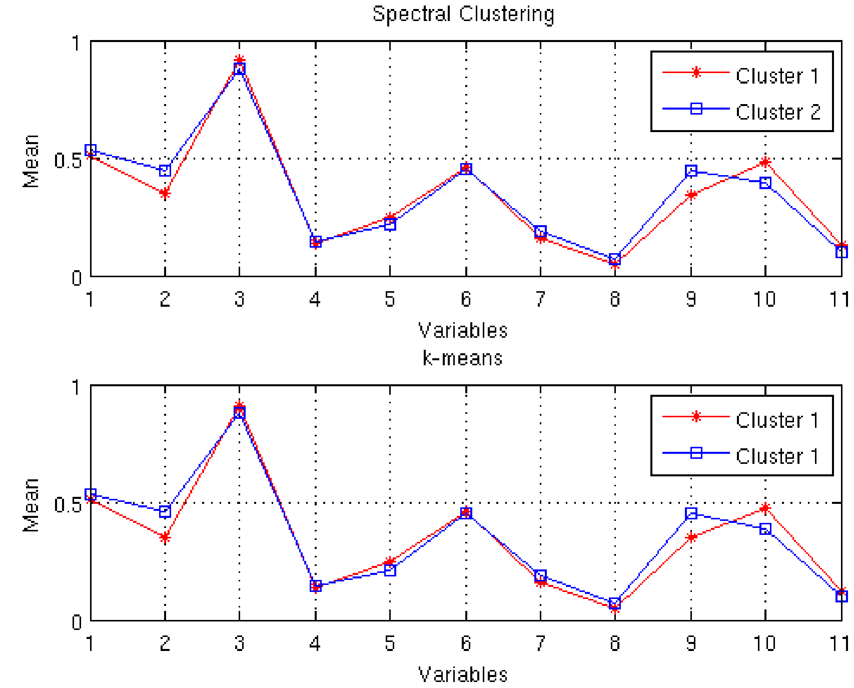


Figure 8: Mean values computed for the two clusters resulting from k-means and spectral method.

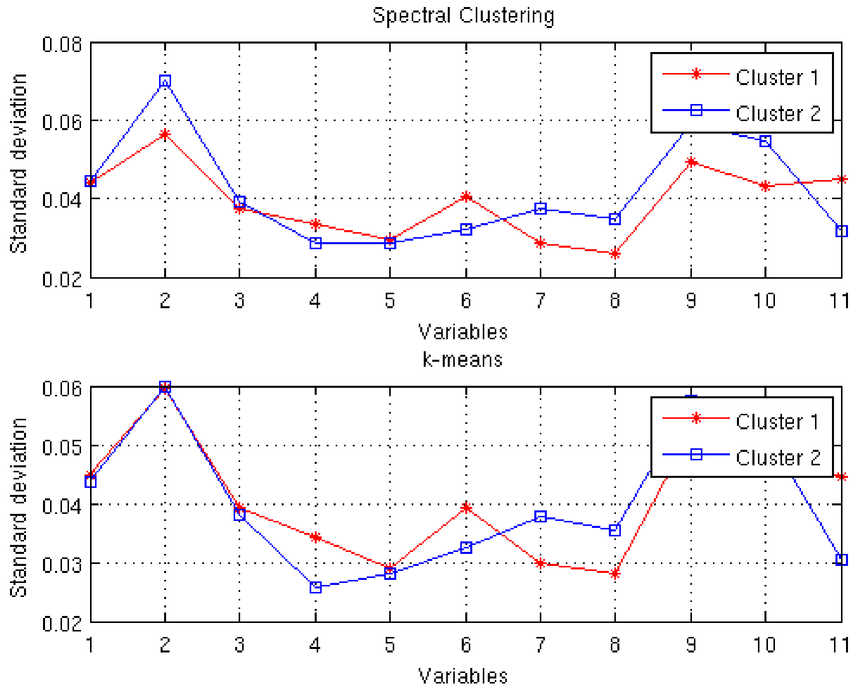


Figure 9: Standard deviation values computed for the two clusters resulting from k-means and spectral method.

(individuals with no connections with the labour market), individuals with more than 55 years and with lower number of years of formal education (for example, this cluster gathers the *concelhos* with a lower percentage of unemployed individuals with a higher education). As mentioned before these groups of individual are the most fragile labour market groups. Both cluster methods seem to divide the total number of *concelhos* in two economic meaningful clusters.

Regarding the standard deviation we observe that the k-means method retrieve clusters that present a lower variability among the observations in each cluster, by variable. The variability seems to be lower for the overall set of characteristics even if the k-means method divides the total number of observations in more uneven clusters.

7 Concluding remarks

In short, both methods denote the same data partition. Applying both methods, the data partition into two clusters minimises the dispersion of data values. The use of the spectral clustering method in an unusual economic application shows potential benefits. Without algorithm parameters refinement the method presented results that are consistent with the k-means results. From the economic point of view both methods show the importance of dividing Portuguese *concelhos* in well defined groups which could be object of distinct public policies. Well targeted labour market measures are, recognisable, more efficient with the cluster methodology helping the identification of different and well defined target regions.

References

[MacQueen, 1967] MacQueen, J. B. (1967). Some Methods for classification and Analysis of Multivariate Observations. Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, 1:281-297.

[Matinez et al, 2011] Martinez, W. L. and Martinez, A. R. and Solka J. L. (2011). Exploratory Data Analysis with MATLAB. *CRC Press.*

[Eldén, 2007] Eldén, L. (2007). Matrix Methods in Data Mining and Pattern Recognition. *SIAM.*

[Mouysset et al, 2008] Mouysset, S. and Noailles, J. and Ruiz, D. (2008). Using a Global Parameter for Gaussian Affinity Matrices in Spectral Clustering. *High Performance Computing for Computational Science - VECPAR 2008*, 378 - 390.

[Ng et al, 2002] Ng, A. Y. and Jordan, M. I. and Weiss Y. (2002). On spectral clustering: Analysis and an algorithm. *Advances in Neural Information Processing Systems (NIPS)*, 14:849-856.

[Kaufman and Rousseeuw, 1990] Kaufman, L. and Rousseeuw P. J. (1990). Finding Groups in Data: An Introduction to Cluster Analysis. *New York: John Wiley & Sons.*

[Overman and Puga, 2002] Overman, H.G. and Puga, D. (2002). Unemployment clusters across Europe's regions and countries. *Economic Policy*, 17(34):115-148.

[Soares et al, 2003] Soares, J. O., Marques, M. M. L. and Monteiro, C. M. F. (2003). A Multivariate Methodology to Uncover Regional Disparities: A Contribution to Improve European Union and Governmental Decisions. *European Journal of Operational Research*, 45:121-135.

[Arandarenko and Juvicic, 2007] Arandarenko, M. and Juvicic, M. (2007). Regional Labour Market Differences in Serbia: Assessment and Policy Recommendations. *The European Journal of Comparative Economics*, 4(2):299-317.

[López-Bazo et al, 2005] López-Bazo, E., Del Barrio, T. and Artís, M. (2005). Geographical Distribution of Unemployment in Spain. *Regional Studies*, 39(3):305-318.

[Nadiya, 2008] Nadiya, D. (2008). Econometric and Cluster Analysis of Potential and Regional Features of the Labor Market of Poland. *Ekonomia*, 21:28-44.

Otimização das visitas domiciliárias das equipas de profissionais de saúde nos Centros de Saúde

Bruno Bastos*, Tiago Heleno*, António Trigo*†, Pedro Martins*‡

* *ISCAC - Coimbra Business School, Instituto Politécnico de Coimbra, Portugal*

E-mail:`iscac10184@alumni.iscac.pt`

E-mail (segundo autor):`iscac10178@alumni.iscac.pt`

E-mail (terceiro autor):`aribeiro@iscac.pt`

E-mail (quarto autor):`pmartins@iscac.pt`

† *Centro ALGORITMI, Universidade do Minho, Portugal*

‡ *Centro de Investigação Operacional, Universidade de Lisboa, Portugal*

Resumo

Os Centros de Saúde têm entre uma das suas muitas tarefas, tarefa da prestação de cuidados de saúde ao domicílio. A organização das visitas é feita por um profissional de saúde que as agrupa em uma ou mais rotas, por forma a minimizar o tempo de saída das equipas. No entanto, não é utilizada nenhuma técnica ou aplicação informática que permita otimizar as visitas de forma sistematizada. Neste artigo apresenta-se a solução desenvolvida, a plataforma Web denominada, Saúde ao Domicílio, que utiliza a heurística de Clarke e Wright e que tem em consideração a existência de doentes prioritários (grau de assepsia).

Palavras chave: Tratamentos hospitalares, Centros de Saúde, Visitas domiciliárias, Otimização de rotas, Heurística de Clarke e Wright.

1 Introdução

A prestação de cuidados de saúde ao domicílio está-se a tornar numa das áreas mais importantes e de maior crescimento na Europa devido ao envelhecimento da população, sendo economicamente vantajoso ter as pessoas em casa em vez de as ter a ocupar camas no hospital [Nickel et al., 2012]. Os cuidados de saúde domiciliários prestados por entidades públicas ou privadas têm sido alvo de investigação recente, no âmbito da investigação operacional, no sentido da otimização das rotas e da composição das equipas que prestam esses serviços [Benzarti et al., 2012, Bertels and Fahle, 2006, Nickel et al., 2012, Rasmussen et al., 2012]

Na prestação de cuidados de saúde domiciliários os Centros de Saúde (CS) desempenham um papel importante pois estão mais próximos da população que os hospitais, mais focados em atender os doentes que a ele se lhe dirigem. Para cumprirem esta função os CS têm de organizar as equipas de profissionais de saúde (médicos e/ou enfermeiros) e as rotas das visitas a efetuar por essas equipas à casa dos doentes.

No sentido de otimizar estas equipas e rotas foi-nos proposto por uma enfermeira de um CS de uma zona urbana (Coimbra) o desenvolvimento de uma aplicação que pudesse responder de forma cabal a este desafio. Neste CS em particular são as equipas, que realizam as visitas domiciliárias, que determinam as rotas das visitas consoante a sua experiência no terreno, ou seja, utilizam os seus conhecimentos geográficos sobre a zona onde atuam definindo assim a rota que consideram ser a melhor, o que nem sempre é verdade. A probabilidade da rota obtida não ser a melhor aumenta quando é necessário atender a algumas restrições, como, por exemplo, a questão dos doentes assépticos ou prioritários.

A otimização de rotas de veículos, em inglês, Vehicle Routing Problem (VRP), é um problema conhecido da otimização combinatória. O objetivo neste tipo de problemas é o de determinar o menor caminho ou rota para, por exemplo, a distribuição ou recolha de mercadorias [Bräysy et al., 2009] ou pessoas [Nunes et al., 2011]. No caso concreto das equipas de profissionais de saúde, estas não entregam mercadorias mas procedimentos que têm de efetuar junto dos doentes, como, por exemplo, aplicação de injetável. Assume-se que os tempos das viagens entre os locais das visitas ao domicílio e que os tempos dos procedimentos a efetuar são conhecidos à partida e que as equipas regressam à origem, ou seja, ao CS.

A Figura 1 mostra os possíveis locais das visitas (casa dos doentes) efetuadas durante um dia pelas equipas de um CS de Coimbra. Os pontos a vermelho mostram os locais que as equipas têm de visitar. O ponto com um círculo vermelho à volta é o CS de onde as equipas de saúde partem e têm de obrigatoriamente regressar. Na secção relativa à implementação apresenta-se como caso de demonstração da plataforma e do algoritmo as rotas entre estes oito pontos de visita.

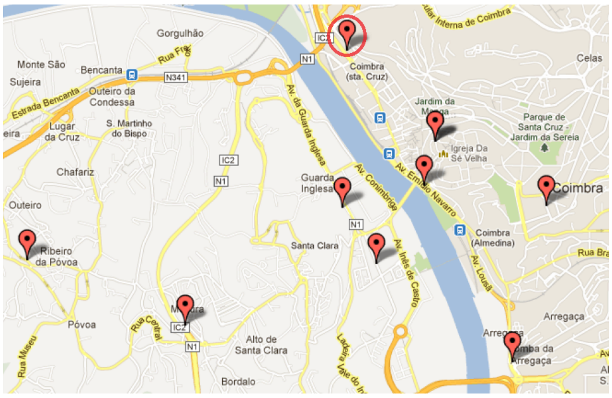


Figura 1: Locais das visitas ao domicílio.

Este artigo está estruturado da seguinte forma: a próxima secção apresenta o que são os cuidados de saúde domiciliários; a terceira secção uma formulação matemática para o problema; na quarta secção é apresentada a metodologia para resolução do problema, baseada na heurística de Clarke e Wright e uma variação através da utilização da heurística de segunda ordem; na quinta secção é apresentada a plataforma Web - Saúde ao Domicílio - desenvolvida com recurso à linguagem PHP e base de dados MySQL da Oracle; finalmente a última secção apresenta algumas considerações finais e propostas de trabalho futuro.

2 Cuidados de saúde domiciliários

Os cuidados de saúde domiciliários são um tipo de cuidado prestado de forma continuada, dirigido para a resolução de alguns problemas de saúde, cuja complexidade não requer o seu internamento hospitalar, mas que dada a sua situação de dependência global, transitória ou crónica, impede os doentes de se deslocarem ao centro de saúde [Moriya and Módena, 2008]. Em [Ministério da Saúde, 2007] é possível consultar alguns dos motivos das solicitações para a prestação deste tipo de cuidados de saúde, que incluem, entre outros, a doença pulmonar obstrutiva crónica e a insuficiência cardíaca congestiva.

Os procedimentos/tratamentos prestados ao domicílio são diversos [Ministério da Saúde, 2007], incluindo por exemplo, ao nível dos cuidados de enfermagem,a Inaloterapia, Aspiração de secreções, Fluidoterapia, Controlo de Feridas/Úlceras de pressão, Ventilação não invasiva, Oxigenoterapia, Alimentação Entérica, etc.

Embora as casas das pessoas não apresentem o risco de contaminação por microrganismos existente nos hospitais [Moriya and Módena, 2008] os profissionais de saúde têm a consciência de que é obrigatório prevenir este tipo de contágios, que podem ser prejudiciais para pessoas mais debilitadas. Neste sentido e por forma a reduzir as contaminações por microrganismos o trabalho aqui apresentado teve em conta o grau de assepsia na definição do tipo de doente.

2.1 Tipo de doente (ou grau de assepsia)

A assepsia é o conjunto de medidas que se tomam para impedir a penetração de microrganismos num ambiente que não os apresenta, logo um ambiente asséptico é aquele que está livre de infeção.

A antissepsia é o conjunto de medidas propostas para inibir o crescimento de microrganismos ou removê-los de um determinado ambiente, podendo ou não destruí-los, utilizando-se para tal fim antissépticos ou desinfetantes.

Os doentes assépticos (ou prioritários) são os que possuem um sistema imunitário mais debilitado e que necessitam, por isso, de serem os primeiros a serem visitados, numa tentativa de minimização de possíveis contaminações. Assim, no desenho da rota da visita, estes doentes não se poderão nunca seguir a doentes não prioritários, de forma a evitar ao máximo as contaminações por microrganismos.

3 Formulação matemática do problema

O problema em questão é uma variante do problema VRP que tem em conta a existência de nodos prioritários e não prioritários que, associados a doentes com elevado grau de assepsia e a doentes que não requerem cuidados especiais de assepsia, respetivamente.

Para o efeito considera-se o grafo orientado $G = (V \cup 0, A)$, com V o conjunto de nodos (doentes), 0 o nodo origem (Centro de Saúde) e A o conjunto de arcos (i, j) entre todos os pares de nodos $i, j \in V \cup 0$. Cada arco (i, j) tem associado um parâmetro t_{ij} que descreve o tempo de viagem de i para j; estando associado a cada nodo $i \in V$ um tempo de visita caracterizado pelo parâmetro r_i.

Cada rota é definida por uma sequência de arcos, com origem e terminus no nodo 0. O tempo total da rota corresponde à soma dos tempos de todos os seus arcos (t_{ij}) e à soma dos tempos dos seus nodos (r_i).

Neste problema, pretende-se calcular um conjunto de rotas que garantam a visita a todos os nodos (doentes), respeitando a restrição de ordem de assepsia, e com limitação no tempo total de cada rota, tendo como objetivo a minimização da soma dos tempos totais de todas as rotas envolvidas.

Apesar dos tempos entre locais serem simétricos, a natureza do problema beneficia por ser tratado como assimétrico, por dois motivos: i) por se conseguir uma formulação de fluxos mais forte em relaxação linear; e ii) por permitir que a restrição de ordem entre nodos dos conjuntos V^1 e V^2 seja feita de forma simples adaptando formulações orientadas conhecidas para o VRP.

De facto, o problema pode ser adaptado de um VRP com capacidades, bastando eliminar as variáveis associadas aos arcos (i, j) com $i \in V^2$ e $j \in V^1$, sendo V^1 e V^2 uma partição do conjunto de clientes V, em que V^1 representa os clientes prioritários e V^2 os não prioritários.

Enquanto que no VRP com capacidades clássico, as variáveis de fluxo traduzem a mercadoria acumulada/libertada nos nodos do grafo, nesta versão as variáveis de fluxo acumulam mercadoria (tempos) dos arcos e mercadoria (tempos) dos nodos.

Assim, propõe-se a seguinte formulação para o problema:

- Variáveis:
  - $x_{ij} = \begin{cases} 1, & \text{se o arco } (i,j) \text{ está na solução} \\ 0, & \text{caso contrário} \end{cases}$, para $i, j \in \{0, \ldots, 1\}$ e $i \neq j$
  - y_{ij}, tempo acumulado da equipa que faz a ligação de i para j, para $i, j \in \{0, \ldots, 1\}$ e $i \neq j$
- Parâmetros:
  - t_{ij}, tempo de viagem de i para j, para $i, j \in \{0, \ldots, 1\}$ e $i \neq j$
  - r_i, tempo de visita em i, , para $i \in \{0, \ldots, 1\}$
  - $V = V^1 \cup V^2$ e $V^1 \cap V^2 = \emptyset$, em que V^1 é o conjunto de nodos associados aos doentes prioritários e V^2 o conjunto de nodos associados aos doentes não prioritários
  - T é o tempo máximo de cada viagem
  - O nodo 0 é o ponto de partida/chegada das rotas
- Função objetivo:

$$min \;\; z = \sum_{i \in V} y_{i0} \tag{1}$$

- sujeito a,

$$\sum_{i \in V \cup \{0\}} x_{ij} = 1, \;\; j \in V \tag{2}$$

$$\sum_{j \in V \cup \{0\}} x_{ij} = 1, \;\; i \in V \tag{3}$$

$$\sum_{i \in V \cup \{0\}} y_{ji} - \sum_{i \in V \cup \{0\}} y_{ij} - \sum_{i \in V \cup \{0\}} t_{ji} x_{ji} = r_j, \;\; j \in V \tag{4}$$

$$(t_{ij} + r_i) x_{ij} \leq y_{ij} \leq (T - r_j) x_{ij}, i, j \in V \cup \{0\} \tag{5}$$

$$x_{ij} \in \{0, 1\}, y_{ij} \geq 0, i, j \in V \cup \{0\} \tag{6}$$

Devem ser eliminadas da formulação todas as variáveis x_{ij} e y_{ij} tais que $i \in V^2$ e $j \in V^1$, assim como, todas as variáveis x_{ii} e y_{ii}, para todo $i \in V \cup \{0\}$.

As restrições (2) e (3) garantem que a cada nodo (doente) chega uma única equipa e sai também uma única equipa. As restrições de conservação de fluxo (4) garantem o incremento temporal de cada rota, evitando a criação de subcircuítos entre nodos de V. As restrições (5) estabelecem a ligação entre as variáveis x e y, e garantem que o fluxo máximo acumulado em qualquer rota não pode ultrapassar o limite temporal T.

Na função objetivo pretendemos minimizar o tempo total contabilizado em todas as rotas, traduzido nas variáveis de fluxo que encerram cada rota.

4 Metodologia

Para a resolução deste problema optou-se, numa primeira abordagem, pela utilização da heurística de Clarke e Wright [Clarke and Wright, 1964]. Numa segunda fase decidiu-se melhorar o algoritmo de Clarke e Wright primeiro com a introdução de aleatoriedade e depois com a integração da heurística de Clarke e Wright num algoritmo melhorativo, nomeadamente um algoritmo de 2ª ordem [Karnaugh, 1976, Martins, 2007]. A combinação de heurísticas é uma prática comum na resolução de problemas deste tipo [Bertels and Fahle, 2006].

4.1 Heurística de Clarke e Wright

Esta heurística, também conhecida como o algoritmo das poupanças, foi proposta pela primeira vez em 1964 por Clarke e Wright para resolver problemas de roteamento de veículos com capacidade, designados em inglês por Capacitated Vehicle Routing Problem (CVRP), em que o número de veículos é livre [Clarke and Wright, 1964, Laporte, 1992].

- Algoritmo de Clarke e Wright [Laporte, 1992] adaptado ao problema:
  1. Calcular as poupanças
  $$s_{ij} = t_{i0} + t_{0j} - t_{ij}, \text{ para } i, j = 1, \ldots, n \text{ e } i \neq j \tag{7}$$
     – com exceção das poupanças s_{ij} tais que $i \in V^2$ e $j \in V^1$;
  2. Ordenar as poupanças por ordem decrescente;
  3. Considerar os percursos que contêm arcos $(i, 0)$ e $(0, j)$ respetivamente. Se $s_{ij} > 0$ fundir provisoriamente estas duas rotas mediante a introdução do arco (i, j) e eliminação dos arcos $(i, 0)$ e $(0, j)$. Verificar a viabilidade da rota obtida. Repetir este passo até que nenhuma melhoria seja possível. Parar.

4.2 Exemplo

Considere-se o exemplo com um CS (CS) e cinco doentes, D_1, D_2, D_3, D_4, D_5, dois dos quais prioritários (o D_1 e o D_3) e os restantes não prioritários.

A Tabela 1 mostra os tempos entre os diferentes nodos e a Tabela 2 as poupanças calculadas utilizando a equação (7), com exceção das poupanças s_{ij} tais que $i \in V^2$ e $j \in V^1$.

Tabela 1: Tempos (minutos)

t_{ij}	CS	D_1	D_2	D_3	D_4	D_5
CS		5	4	2	1	3
D_1			3	1	1	2
D_2				2	1	3
D_3					4	5
D_4						3
D_5						

Tabela 2: Poupanças ordenadas

s_{ij}	=	$t_{0i} + t_{0j} - t_{ij}$	s_{ij}	=	$t_{0i} + t_{0j} - t_{ij}$
s_{12}	=	6	s_{52}	=	4
s_{13}	=	6	s_{24}	=	3
s_{31}	=	6	s_{42}	=	3
s_{15}	=	6	s_{45}	=	1
s_{14}	=	5	s_{54}	=	1
s_{32}	=	4	s_{35}	=	0
s_{25}	=	4	s_{34}	=	-1

Note-se que embora a matriz das distâncias seja simétrica, o nosso algoritmo é assimétrico, pois tem em conta os sentidos dos fluxos, pelo que, no início todas as poupanças são tidas em conta.

O algoritmo inicia com 5 rotas, uma para cada cliente. Depois junta duas dessas rotas, e obtém uma solução com 4 rotas, uma rota com os doentes 1 e 2 e mais 3 rotas, uma para cada um dos restantes doentes, como se pode ver na Figura 2. Já a solução final é só uma rota que passa por todos os doentes. De notar que a restrição de não ir de um doente não prioritário (a azul na Figura 2) para um doente prioritário (a vermelho na Figura 2) não foi quebrada.

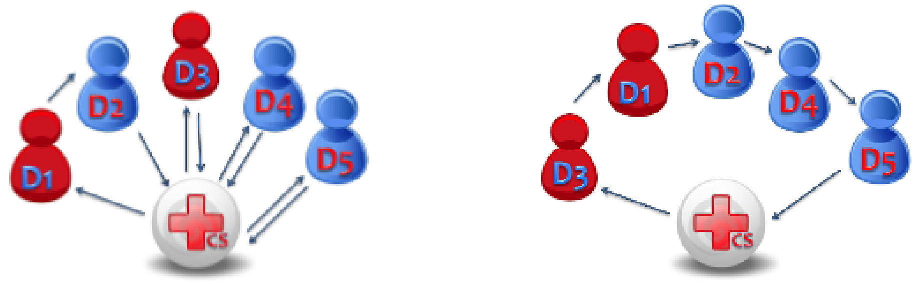


Figura 2: Solução da primeira iteração e solução final do algoritmo

O objetivo deste problema é o de minimizar o tempo total das rotas (em minutos), visitando todos os doentes. Na solução inicial com as cinco rotas, o tempo total da visita (Z) é de 30 minutos. Na solução obtida na primeira iteração, com duas rotas, o Z é de 24 minutos, uma melhoria de 6 minutos em relação à solução inicial, enquanto que na solução final, só com uma rota, o Z é de 13 minutos. A solução final permite uma redução no tempo total das rotas de 16 minutos, permitindo poupar mais de metade do tempo em relação à solução inicial (com 5 rotas).

4.3 Adaptação do algoritmo

Num primeiro teste ao algoritmo, com os doentes da Tabela 3 verificou-se que as soluções apresentadas, embora resolvessem o problema, não eram as melhores. O algoritmo sofre de alguma miopia pois a escolha inicial das maiores poupanças não conduz forçosamente à melhor solução. O tempo das rotas obtido para o problema apresentado na secção cinco foi de 2h28m. Este tempo obtido corresponde a cinco rotas: $CS \to D_1 \to CS$; $CS \to D_7 \to D_2 \to CS$; $CS \to D_3 \to D_8 \to CS$; $CS \to D_6 \to D_4 \to CS$; e $CS \to D_5 \to CS$.

No sentido de melhorar o algoritmo optou-se por introduzir alguma aleatoriedade através da imposição de saltos na escolha da ordem das poupanças, ou seja, em vez de escolher sequencialmente por ordem decrescente as poupanças, a nova versão do algoritmo salta algumas destas poupanças por forma a tentar encontrar melhores soluções. Esta solução permitiu obter resultados melhores que solução anterior. Para o problema anterior o melhor tempo obtido para a rota foi de 2h11m, com três rotas, a $CS \to D_1 \to D_7 \to D_2 \to D_4 \to CS$, a $CS \to D_3 \to D_8 \to CS$ e $CS \to D_5 \to D_6 \to CS$, o que se traduz numa poupança de 17 minutos. Dado que este algoritmo possui uma componente aleatória poderiam ser obtidos melhores resultados se ele fosse executado durante mais tempo. Não obstante, um dos critérios para a escolha do algoritmo a utilizar é a rapidez com que ele consegue obter boas soluções.

Finalmente decidiu-se testar uma última abordagem/adaptação ao algoritmo de Clarke e Wright que foi a utilização de um algoritmo de segunda ordem proposto por [Karnaugh, 1976] e melhorado por [Martins, 2007]. Nesta abordagem, depois de encontrada uma solução inicial, fixa-se uma ligação e testam-se as restantes poupanças à procura da melhor solução. Quando se encontra a melhor solução esta é guardada e as poupanças que lhe deram origem são inutilizadas, exceto a que originou a melhor solução, a qual é tornada permanente. O algoritmo volta a ser executado considerando o conjunto de ligações até ao momento tornadas permanentes, procurando fixar mais ligações. Com este algoritmo foi possível obter um tempo de 2h05. O tempo obtido corresponde à realização de uma só rota a $CS \to D_5 \to D_6 \to D_1 \to D_7 \to D_3 \to D_8 \to D_2 \to D_4 \to CS$ e permite uma poupança de 13 minutos.

5 Implementação

Dado a necessidade de ter uma plataforma central e acessível aos vários profissionais de saúde e que no futuro interaja como o sistema de informação do CS optou-se pelo formato Web para o desenvolvimento da aplicação de criação de rotas das equipas dos profissionais de saúde. A plataforma Web - Saúde ao Domicílio - foi desenvolvida utilizando tecnologias *Open Source* designadamente a linguagem PHP, o servidor de páginas Web Apache e o sistema de gestão de bases de dados MySQL da Oracle.

5.1 Funcionalidades

A plataforma Web - Saúde ao Domicílio - irá fornecer as funcionalidades que se apresentam no diagrama de casos de uso na Figura 3. O utilizador representa o profissional de saúde responsável pela gestão

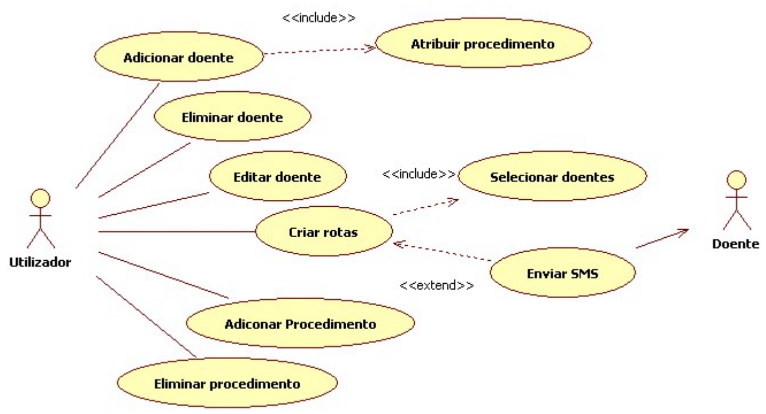


Figura 3: Diagrama de casos de uso do sistema (plataforma Web)

dos doentes e das rotas. Na funcionalidade de adição dos doentes, são adicionados dados como, nome, morada, se é ou não prioritário e o procedimento/tratamento a efetuar. A criação de rotas pressupõe a escolha dos doentes para os quais se quer criar a rota e possuirá a funcionalidade de envio de mensagens SMS aos doentes a informar da data e hora da visita.

5.2 Plataforma Web - Saúde ao Domicílio

A plataforma Web - Saúde ao Domicílio - desenvolvida possui três áreas de trabalho, uma relativa à gestão dos dados dos doentes, outra relativa à gestão dos procedimentos (tratamentos) dos doentes e uma terceira que é a criação das rotas das visitas domiciliárias, como se pode ver na Figura seguinte.

Na Figura 4 é possível visualizar a edição de dados do doente (fictício) prioritário José Correia.

Figura 4: Saúde ao Domicílio - edição dos dados do doente

5.3 Experimentação

Para demonstrar a utilização da plataforma Web Saúde ao Domicílio, apresenta-se de seguida o exemplo mostrado na Figura 1 com 8 doentes.

A Tabela 3 mostra os nomes, moradas, tipos de doentes e procedimentos. Os nomes são fictícios e as moradas, embora correspondam a locais reais, da cidade de Coimbra, não têm qualquer relação com possíveis doentes nesses locais. Só se utilizaram moradas reais para que se pudessem obter os tempos, que servem de base aos cálculos e fosse possível apresentar uma demonstração da execução do algoritmo.

Tabela 3: Dados dos doentes dos locais a visitar (fictícios)

Cód.	Nome	Morada	Prioritário	Procedimento
1	José Correia	Rua António Augusto Gonçalves	sim	Úlceras de pressão
2	Maria Correia	Av. Emídio Navarro 37	não	Oxigenoterapia
3	Rui Fernandes	Rua Coutinhos 26	sim	Úlceras de pressão
4	Fátima Neves	Estrada da Guarda Inglesa, nº 17	não	Ventilação
5	Ana Costa	Rua Caminho das Vinhas 12	sim	Controlo de Feridas
6	Fernando Esteves	Rua da Escola, 37	sim	Úlceras de pressão
7	Celeste Marques	Av. Cónego Urbano Duarte 92	sim	Alimentação Entérica
8	Mário Gouveia	Rua Augusto Filipe Simões	não	Oxigenoterapia

O primeiro passo é a autenticação do utilizador. Depois de validado, caso necessite criar algum doente acede à área de gestão de doentes, caso contrário, pode ir diretamente para a área de gestão de rotas. É nesta área que são otimizadas as rotas das equipas dos profissionais de saúde. Para isso o utilizador deverá selecionar todos os doentes que deseja que façam parte da simulação (no caso deste exemplo em particular são os oito doentes da Tabela 3) e correr o algoritmo, como se pode ver na Figura 5. Por forma a facilitar a identificação do tipo de doente, a plataforma Saúde ao Domicílio lista os doentes prioritários a vermelho e os não prioritários a verde.

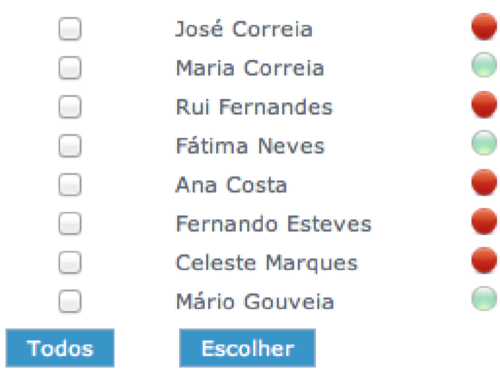


Figura 5: Saúde ao Domicílio - seleção dos doentes para criar as rotas

A Figura 6 e a Figura 7 mostram o resultado obtido, uma sob a forma numérica e outra sobre a forma de mapa.

Trajecto da 1 ª equipa:
Saída do Centro de Saúde:
5 - Rua Caminho das Vinhas 12
6 - Rua da Escola, 37, 3040-563 Coimbra
1 - Rua António Augusto Gonçalves, 3041-901 Coimbra
7 - Avenida Cónego Urbano Duarte 92, 3030 Coimbra
3 - Rua Coutinhos 26, 3000 Coimbra
8 - Rua Augusto Filipe Simões, 3000 Coimbra
2 - Avenida Emídio Navarro 37, 3000-150 Coimbra
4 - Estrada da Guarda Inglesa, N.º 17, 3040-193 Coimbra
Regresso ao Centro de Saúde!

O TEMPO FINAL DESTA SOLUÇÃO É: 02:05:00

Figura 6: Saúde ao Domicílio - rota proposta

Neste caso a aplicação devolveu uma só rota, a rota $CS \rightarrow 5 \rightarrow 6 \rightarrow 1 \rightarrow 7 \rightarrow 3 \rightarrow 8 \rightarrow 2 \rightarrow 4 \rightarrow CS$, que no mapa aparece por ordem alfabética.

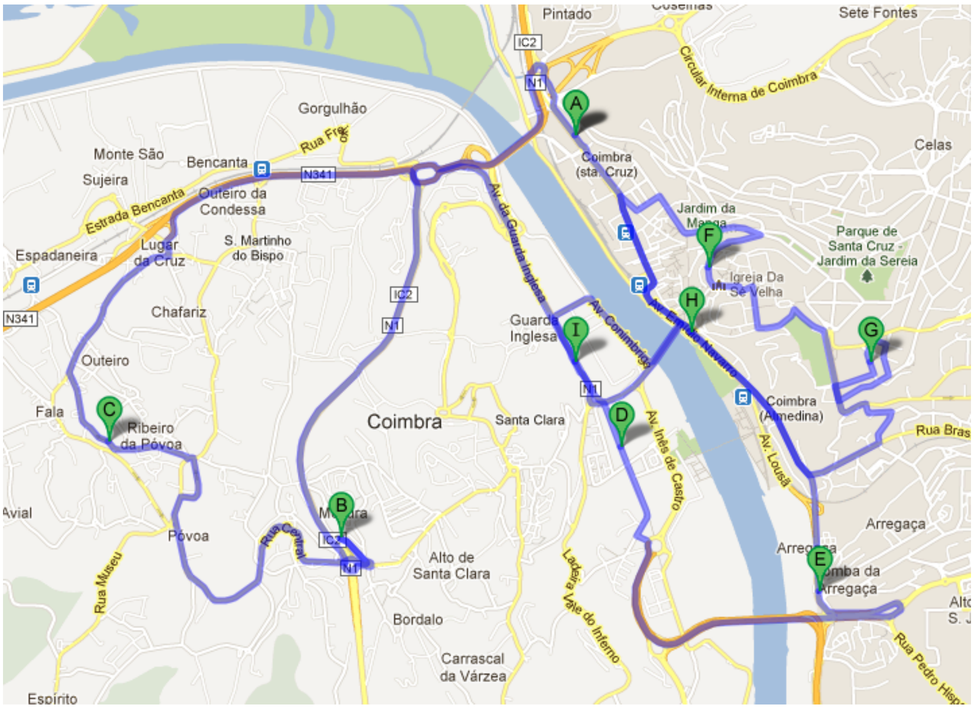


Figura 7: Saúde ao Domicílio - rota proposta sob a forma de mapa

As soluções apresentadas incluem o tempo das viagens entre as casas dos doentes (arcos) e o tempo despendido em casa dos doentes (nodos). Embora se pudessem ter considerado tempos diferentes para efeitos da demonstração considerou-se que as equipas médicas despendem 10 minutos em casa dos doentes a efetuar os respetivos procedimentos.

6 Considerações finais

A plataforma desenvolvida otimiza as rotas das visitas das equipas de enfermagem, que tem como consequência a redução de custos de combustível e o aumento da eficácia das equipas de enfermagem.

A inovação desta plataforma, que não se encontra na revisão da literatura efetuada, é otimizar as rotas das visitas das equipas de enfermagem em função do grau de assepsia dos doentes, permitindo a sua diferenciação em prioritários e não prioritários, contribuindo desta forma para minorar a transmissão de infeções entre os doentes.

Relativamente a trabalhos futuros ter-se-á que identificar mais restrições, como, por exemplo, as alturas em que os pacientes preferem ser visitados ou a preferência dos doentes por determinados enfermeiros [Bertels and Fahle, 2006], por forma a melhorar a satisfação dos doentes e enfermeiros no serviço prestado.

Outra situação não abordada, mas também não solicitada, foi a questão da otimização da composição das equipas dos profissionais de saúde referenciada noutros trabalhos similares [Bertels and Fahle, 2006, Nickel et al., 2012] e que poderá ser alvo de estudo em futuros desenvolvimentos. O algoritmo atualmente implementado vê as equipas como entidades únicas independentemente do número de elementos que a compõem.

O próximo passo será a criação de uma agenda na plataforma que permita listar os doentes que precisam de ser visitados num determinado dia.

Não obstante o trabalho desenvolvido e apresentado neste artigo ter sido realizado com base num CS da zona centro, este pode ser generalizado, para todo o país.

Agradecimentos

Agradecemos aos nossos colegas do curso em Informática de Gestão que contribuíram para o desenvolvimento deste projeto, em especial ao Daniel Fernandes que contribui com conselhos e críticas para a formulação de algumas questões do projeto. Gostaríamos de agradecer à enfermeira Mariana Costa, profissional de enfermagem, pela ideia que originou todo este projeto, bem como todo o seu apoio. Por último, gostaríamos de agradecer ao revisor 1 as correções e sugestões de melhoria do nosso artigo.

Referências

[Bräysy et al., 2009] Bräysy, O. Nakari, P., Dullaert, W. and Neittaanmäki, P. (2009). An optimization approach for communal home meal delivery service: A case study. *Journal of Computational and Applied Mathematics*, 232(1):46-53.

[Benzarti et al., 2012] Benzarti, E., Sahin E. and Dallery Y. (2012). Operations management applied to home care services: Analysis of the districting problem. *Decision Support Systems*.

[Bertels and Fahle, 2006] Bertels S. and Fahle T. (2012). A hybrid setup for a hybrid scenario: combining heuristics for the home health care problem. *Computers & Operations Research*, 33(10):2866-2890.

[Clarke and Wright, 1964] Clarke, G. and Wright, J., (1964). Scheduling of vehicles from a central depot to a number of delivery points. *Operations Research*, 12 (4): 568-581.

[Laporte, 1992] Laporte, G. (1992). The vehicle routing problem: An overview of exact and approximate algorithms. *European Journal of Operational Research*, 59(3):345-358.

[Karnaugh, 1976] Karnaugh M. (1976). A new class of algorithms for multipoint network optimization. *IEEE Transactions on Communications*, COM-24:500–505.

[Martins, 2007] Martins, P. (2007). Enhanced second order algorithm applied to the capacitated minimum spanning tree problem. *Computers & Operations Research*, 34(8):2495-2519.

[Ministério da Saúde, 2007] Ministério da Saúde (2007). Cuidados Continuados Integrados nos CSP – Carteira de Serviços, 2007-05-21.

[Moriya and Módena, 2008] Moriya, T. and Módena, J. L. P., (2008). Assepsia e antissepsia: técnicas de esterilização. *Simpósio: Fundamentos em Clínica Cirúrgica - 1ª Parte*, 41 (3): 265-273.

[Nickel et al., 2012] Nickel, S., Schröder, M. and Steeg, J. (2012). Mid-term and short-term planning support for home health care services. *European Journal of Operational Research*, 219(3):574-587.

[Nunes et al., 2011] Nunes, J., Matos, L. and Trigo, A. (2011). Taxi pick-ups route optimization using genetic algorithms. *Proceedings of the 10th international conference on Adaptive and natural computing algorithms (ICANNGA'11) - Volume Part I*, 410-419.

[Rasmussen et al., 2012] Rasmussen, M.S., Justesen, T., Dohn, A. and Larsen J. (2012). The Home Care Crew Scheduling Problem: Preference-based visit clustering and temporal dependencies. *European Journal of Operational Research*, 219(3):598-610.

Aproximação de cálculos iterativos por redes neuronais em sistemas de equações diferenciais ordinárias

Ana S. R. Brásio*†, Andrey Romanenko*, Natércia C. P. Fernandes†,[1]

* *Ciengis, SA, Coimbra, Portugal*

† *CIEPQPF, Departamento de Engenharia Química, Faculdade de Ciências e Tecnologia, Universidade de Coimbra, Portugal*

Resumo

Este trabalho descreve uma abordagem para melhorar o desempenho numérico de modelos dinâmicos mecanísticos, tornando-os viáveis no contexto de optimização dinâmica e controlo preditivo não linear. O método consiste em aproximar partes computacionalmente intensivas dos modelos mecanísticos por redes neuronais.

A sua aplicação é exemplificada no modelo da separação de fases numa linha de produção de biodiesel. Os resultados principais são a eliminação do cálculo iterativo *flash* sujeito a uma condição de paragem baseada na comparação de previsões adjacentes e a possibilidade de utilização de ferramentas de diferenciação automática para facilitar a resolução de problemas de optimização não-linear.

Palavras chave: Processo de produção de biodiesel, decantador, modelo, redes neuronais.

1 Introdução

A sociedade actual encontra-se largamente dependente do petróleo como fonte energética. Porém, tratando-se de um recurso finito, as alternativas renováveis têm vindo a assumir um papel cada vez mais importante. Em particular, o biodiesel — um biocombustível produzido a partir de óleo — apresenta nesta perspectiva um conjunto de características bastante atractivas. Em consonância, um considerável esforço de investigação devotado ao desenvolvimento e bom funcionamento da indústria de biodiesel tem vindo a ser levado a cabo nos últimos anos.

Uma das peças de mais relevo na linha produtiva do biodiesel é o reactor, onde o óleo reage com metanol sob certas condições operatórias produzindo uma mistura de biodiesel e do sub-produto glicerol. Após a reacção, a mistura é arrefecida e os seus componentes separados. A Figura 1 representa esquematicamente a linha produtiva. Convém realçar que o passo de separação é vulgarmente realizado na

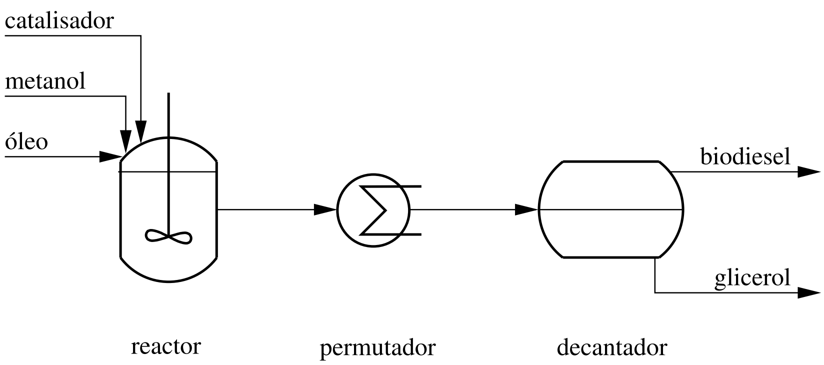


Figura 1: Representação esquemática simplificada do processo de produção descontínuo de biodiesel.

indústria de biodiesel num decantador gravítico. A decantação gravítica é um processo moroso, pelo que este passo representa uma fatia significativa do tempo total de produção, ultrapassando várias vezes o tempo de residência necessário no reactor.

[1]To whom all correspondence should be addressed. Email: natercia@eq.uc.pt.

A diminuição do tempo de decantação traduzir-se-ia numa melhoria económica do processo. Para tal, pode recorrer-se a ferramentas de optimização dinâmica [1] que permitem obter um compromisso entre os objectivos pretendidos e os custos que lhes estão associados. Estas técnicas são baseadas em modelos que descrevem a dinâmica dos processos e que podem atingir um grau de complexidade elevado. Por outro lado, a operação de uma linha produtiva de biodiesel pode ser francamente melhorada através de um sistema de controlo preditivo não-linear como descrito em [2]. O sistema requer igualmente um conjunto de modelos dinâmicos das unidades processuais existentes na linha produtiva. Estes modelos são usados para obter previsões temporais das trajectórias das variáveis de estado e das variáveis de saída assim como para determinar a sensibilidade das soluções relativamente às variáveis manipuladas.

No decantador coexistem duas fases líquidas (a fase leve e a fase pesada) que interagem entre si. É, então, necessário proceder a cálculos de equilíbrio líquido-líquido que permitam quantificar essa interacção para posterior incorporação deste fenómeno no modelo dinâmico do decantador. A quantificação de equilíbrios líquido-líquido efectua-se através de um cálculo *flash* [3], que é um método iterativo.

Porém, um modelo dinâmico que recorra a métodos iterativos não é passível de ser integrado de forma eficiente na plataforma computacional de controlo preditivo. O modelo é invocado dezenas de vezes por iteração. Apesar do sistema ter mecanismos que permitem acelerar a convergência do algoritmo, a utilização de ferramentas iterativas aumenta significativamente o tempo de cálculo e a memória necessária. Por um lado, o cálculo iterativo do equilíbrio de fases em cada chamada do modelo inviabiliza-o do ponto de vista de tempos de computação requeridos. Por outro lado, a forma iterativa prejudica a utilização de métodos de diferenciação automática, tal como ADOL-C [4] ou CppAD [5], pois aumenta significativamente a memória necessária para efectuar os respectivos cálculos.

No presente trabalho desenvolve-se uma abordagem alternativa para o cálculo do equilíbrio de fases por forma a evitar o método iterativo sem deteriorar a qualidade das previsões. Para tal, os resultados obtidos pelo cálculo *flash* são aproximados por um modelo baseado em redes neuronais cujo tipo, composição e características se detalham na Secção 3.

2 Equilíbrio líquido-líquido

A metodologia mais utilizada para quantificar o equilíbrio líquido-líquido é o cálculo *flash*, descrita em detalhe em [3] para situações de equilíbrio entre duas fases líquidas parcialmente miscíveis.

Considerando uma corrente de alimentação composta por n_c componentes de composição $x_{i,\text{in}}$, o equilíbrio à pressão P e à temperatura T é atingido formando duas fases distintas de composição $x_{i,\text{leve}}$ e $x_{i,\text{pesa}}$ (onde $i = 1, \ldots, n_c$). A alimentação é separada em duas fases: a fracção molar L_{leve} constituirá a fase leve e o restante, fracção $1 - L_{\text{leve}}$, constituirá a fase pesada. O equilíbrio de cada componente na mistura é estabelecido pela razão K_i traduzida pelo quociente entre as fracções molares da espécie química i nas duas fases líquidas, isto é,

$$K_i = \frac{x_{i,\text{leve}}}{x_{i,\text{pesa}}} = \frac{\gamma_{i,\text{leve}}}{\gamma_{i,\text{pesa}}} \ , \tag{1}$$

onde $\gamma_{i,\text{leve}}$ e $\gamma_{i,\text{pesa}}$ são os coeficientes de actividade do componente i nas fases leve e pesada, respectivamente.

Na Figura 2 representa-se, de forma esquemática, a quantificação mecanística do equilíbrio líquido-líquido, sendo evidente a sua natureza iterativa. Após especificação da alimentação, e já no interior do ciclo iterativo, recorre-se ao método UNIFAC (ou a variações deste método) para calcular os coeficientes de actividade, necessários, por sua vez, ao cálculo das constantes de equilíbrio. O método UNIFAC [6] estima os coeficientes com base no somatório das contribuições dos grupos funcionais presentes nos componentes da mistura: éster (biodiesel), metanol e glicerol.

O óleo que dá origem ao éster é composto por glicerídeos (essencialmente triglicerídeos) cujo esqueleto consiste numa molécula de glicerol a que se ligam ácidos gordos. A origem biológica do óleo confere-lhe variabilidade. Contudo, o ácido láurico é habitualmente o ácido gordo em maior quantidade nos diversos óleos vegetais. Por esta razão e no contexto deste trabalho, considera-se que o ácido gordo presente na matéria-prima é exclusivamente o ácido láurico (ou seja, o biodiesel é constituído exclusivamente pelo éster metillaureato).

Atingida a convergência do cálculo *flash*, é então possível quantificar o grau de separação do componente i pelas fases leve e pesada. Da quantidade presente inicialmente, a fracção de componente i que irá para a fase leve é dada por

$$\xi_i = L_{\text{leve}} \frac{x_{i,\text{leve}}}{x_{i,\text{in}}} \ . \tag{2}$$

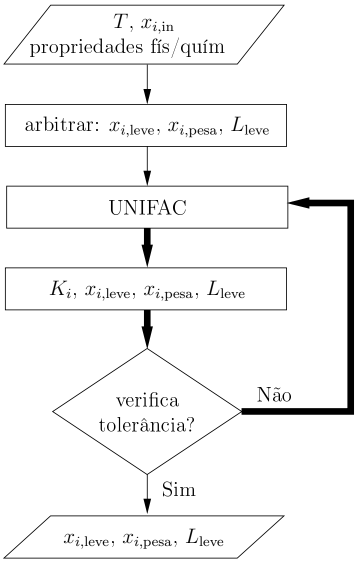


Figura 2: Fluxograma do cálculo *flash* para determinação do equilíbrio líquido-líquido.

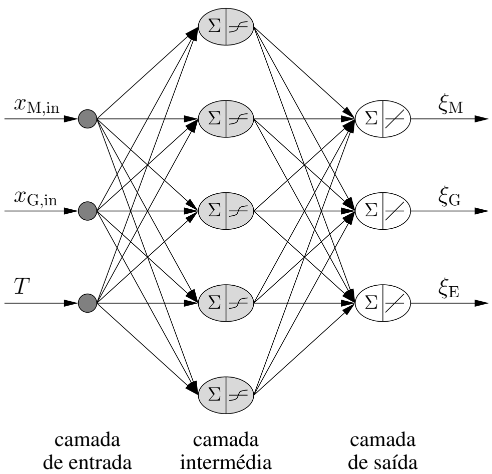


Figura 3: Rede neuronal usada para substituição do cálculo *flash*.

3 Rede neuronal artificial

As redes neuronais artificiais utilizam-se em inúmeras aplicações de engenharia para previsão de variáveis de sistemas complexos [7, 8]. A sua utilização permite a simulação de fenómenos físicos sem formulação mecanística explícita para descrever as relações existentes entre as variáveis [8].

As redes neuronais do tipo *feedforward back-propagation* são as mais utilizadas nestas aplicações. Este tipo de redes são consideradas estáticas porque dependem apenas das variáveis de entrada actuais e de constantes. A ausência de informação posterior (*feedback*) assegura a estabilidade do modelo [7].

A Figura 3 representa a arquitectura da rede neuronal do tipo *feedforward back-propagation* implementada. É composta por três camadas de nodos ou neurónios (a camada de entrada, a camada intermédia e a camada de saída), como é típico deste tipo de redes.

O número de camadas intermédias pode variar fazendo aumentar a capacidade de previsão da rede, o que se revela especialmente útil em problemas com um elevado número de variáveis de entrada. Contudo, o aumento do número destas camadas também contribui para o treino excessivo da rede devido ao elevado número de constantes a determinar, isto é, pode levar ao *overfitting* da rede [9, 7], além de aumentar exponencialmente o seu tempo de aprendizagem.

Os neurónios intermédios e de saída são estruturados por uma função de agregação e por uma função de activação. De acordo com a prática habitual, usa-se o somatório como função de agregação. As funções de activação mais comuns são a função linear, a função sigmóide e a tangente hiperbólica [7], tendo-se recorrido no presente trabalho à tangente hiperbólica e à função linear para as camadas intermédia e de saída, respectivamente, como representado no interior dos referidos neurónios na Figura 3.

Cada neurónio está directamente ligado aos neurónios das camadas adjacentes. A cada ligação é atribuído um peso que representa o grau de relação entre os dois neurónios envolvidos. Matematicamente, tem-se [9, 7]

$$\text{neurónio intermédio}_j = f_{\text{I}}\left(\sum_{i=1}^{n_X} w_{ij} \cdot X_i + \theta_j\right) \quad , \qquad j = 1, \ldots, n_I \tag{3}$$

e

$$Y_k = f_{\text{Y}}\left(\sum_{j=1}^{n_I} W_{jk} \cdot \text{neurónio intermédio}_j + \Gamma_k\right) \quad , \qquad k = 1, \ldots, n_Y \ , \tag{4}$$

onde n_X é o número de neurónios de entrada, n_I é o número de neurónios intermédios, n_Y é o número de neurónios de saída, X_i é o neurónio de entrada i, Y_k é o neurónio de saída k, w_{ij} é o peso do neurónio de entrada i relativamente ao neurónio intermédio j, W_{jk} é o peso do neurónio intermédio j relativamente ao neurónio de saída k, θ_j é a bias do neurónio intermédio j, Γ_k é a bias do neurónio de saída k, $f_{\text{I}}(\cdot)$ é a função de activação dos neurónios intermédios e $f_{\text{Y}}(\cdot)$ é a função de activação dos neurónios de saída.

A função de activação $f_{\mathrm{I}}(\cdot)$ foi definida através da tangente hiperbólica porque permite a convergência mais rápida do algoritmo de treino [9]. Já a função de activação para a camada de saída, $f_{\mathrm{Y}}(\cdot)$, é frequentemente traduzida por uma função linear [7]. Em linguagem matemática,

$$\text{Tangente hiperbólica :} \qquad f_{\mathrm{I}}(q) = \frac{\mathrm{e}^{q} - \mathrm{e}^{-q}}{\mathrm{e}^{q} + \mathrm{e}^{-q}} \ , \qquad -1 < \ f_{\mathrm{I}} \ < 1 \tag{5}$$

$$\text{Linear :} \qquad f_{\mathrm{Y}}(q) = q \ , \qquad -\infty < \ f_{\mathrm{Y}} \ < \infty \tag{6}$$

considerando q uma variável genérica.

A função contínua e diferenciável com capacidade de previsão que a rede neuronal gera é definida na forma vectorial por

$$\boldsymbol{Y}(\boldsymbol{P}, \boldsymbol{X}) = \boldsymbol{W} \cdot \tanh(\boldsymbol{w}^{\top} \cdot \boldsymbol{X} + \boldsymbol{\theta}) + \boldsymbol{\Gamma} \ , \tag{7}$$

onde $\boldsymbol{P}$ representa o conjunto de parâmetros matriciais $\boldsymbol{w} \in \mathbb{R}^{n_X \times n_I}$, $\boldsymbol{W} \in \mathbb{R}^{n_Y \times n_I}$, $\boldsymbol{\theta} \in \mathbb{R}^{n_I \times 1}$ e $\boldsymbol{\Gamma} \in \mathbb{R}^{n_Y \times 1}$; $\boldsymbol{X} \in \mathbb{R}^{n_X \times 1}$ representa o vector de neurónios de entrada e $\boldsymbol{Y} \in \mathbb{R}^{n_Y \times 1}$ o vector de neurónios de saída.

Considerando um conjunto de dados com m pontos $\{(\boldsymbol{X}_1, \ \boldsymbol{Y}_1), \ldots, (\boldsymbol{X}_i, \ \boldsymbol{Y}_i), \ldots, (\boldsymbol{X}_m, \ \boldsymbol{Y}_m)\}$, no treino da rede neuronal pretende-se determinar os parâmetros por forma a que, para a entrada $\boldsymbol{X}_i$, a estimativa das variáveis de saída $\hat{\boldsymbol{Y}}_i$ seja igual aos valores $\boldsymbol{Y}_i$ [10]. Este problema de optimização é traduzido pela minimização do erro quadrático médio (MSE), isto é,

$$\min_{\boldsymbol{P}} \ F(\boldsymbol{P}) = \frac{1}{m} \sum_{i=1}^{m} \boldsymbol{e}_i^{\top} \, \boldsymbol{e}_i \ ,$$

onde $\boldsymbol{e}_i = \boldsymbol{Y}_i - \hat{\boldsymbol{Y}}_i(\boldsymbol{P}, \boldsymbol{X})$. Para a sua resolução foi utilizado o algoritmo de Levenberg-Marquardt por ser um algoritmo eficiente mesmo em problemas muito mal-condicionados [11]. O algoritmo baseia-se nos métodos de Gauss-Newton e do Gradiente e determina $\boldsymbol{P}$ a cada iteração usando

$$\Delta \boldsymbol{P} = -(\boldsymbol{H} + \mu \boldsymbol{I})^{-1} \ \nabla F \ , \tag{8}$$

onde $\boldsymbol{H}$ é a matriz hessiana, $\boldsymbol{I}$ a matriz identidade, ∇F o vector gradiente de $F(\boldsymbol{P})$ e μ a taxa de aprendizagem que é actualizada por forma a minimizar MSE [12].

4 Resultados e Discussão

Com base na composição e na temperatura de uma mistura de éster, metanol e glicerol que entra no decantador, a rede neuronal deve indicar a forma como os três componentes se separam pelas fases leve e pesada, isto é, deve prever as fracções de separação dos componentes para a fase leve. Assim, do ponto de vista da rede neuronal, as variáveis de entrada são a temperatura (T) e a composição da mistura que se pretende separar. A composição da mistura exprime-se em termos das fracções molares do metanol e do glicerol[2], $x_{\mathrm{M,in}}$ e $x_{\mathrm{G,in}}$. As variáveis de saída são as fracções de separação dos três componentes ξ_{E}, ξ_{M} e ξ_{G}, indicando, para cada componente, a fracção da quantidade inicial presente que vai para a fase leve.

4.1 Geração e tratamento de dados

Os dados de equilíbrio líquido-líquido a ajustar com recurso a uma rede neuronal foram gerados a partir do cálculo *flash* descrito na Secção 2.

A caracterização da mistura inicial na unidade de separação é de especial importância, já que a rede deve ser treinada com um conjunto de dados relevantes tendo em conta as gamas habitualmente verificadas neste tipo de sistemas. Os autores de [13] realizaram experimentalmente a reacção de transesterificação do óleo de girassol a 60℃ usando uma razão molar entre o metanol e o óleo de 6:1, 0.50 % (m/m) de NaOH como catalisador e uma velocidade de agitação de 400 rpm. No referido trabalho, mostra-se a concentração molar dos componentes ao longo do tempo. Contudo, a informação sobre o metanol, um dos componentes em maior quantidade na mistura, é omitida. Por esta razão, foi necessário proceder à simulação da reacção de transesterificação (reactor) por forma a poder obter os perfis dinâmicos da composição das diversas espécies químicas e, em última instância, quantificar completamente a mistura no

[2]Note-se que a fracção molar do éster é linearmente dependente das fracções molares dos outros dois componentes.

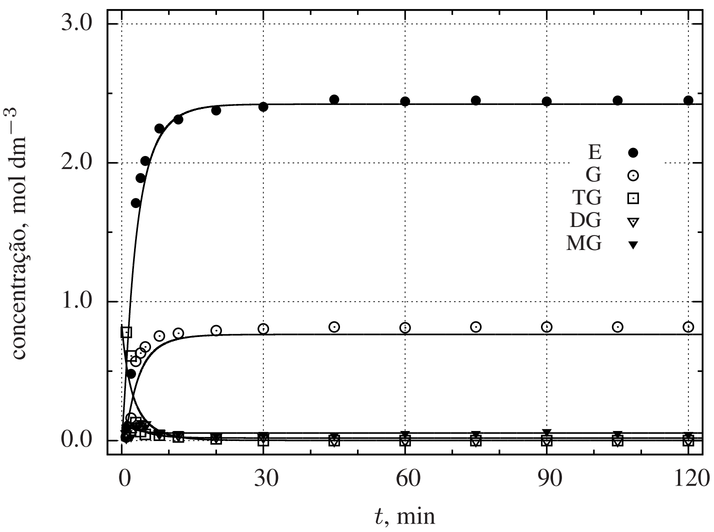


Figura 4: Simulação da cinética de reacção de transesterificação do óleo de girassol.

fim da reacção, nomeadamente no que respeita ao metanol. Como o equilíbrio e a velocidade do conjunto das reacções de transesterificação são condicionados pela agitação do meio reaccional, o trabalho [14] propõe uma metodologia para contemplar de forma sistemática esta variável no modelo do reactor.

O modelo e os parâmetros descritos no trabalho [13] foram usados na simulação do sistema e os resultados correspondentes apresentam-se na Figura 4. A comparação visual entre os pontos simulados e os experimentais assim como o valor do coeficiente de determinação ($R^2 = 0.99998$) permitem concluir sobre a boa reprodução do sistema apresentado em [13].

A mistura reaccional já esgotada que abandona o reactor é então direccionada para o decantador sem sofrer qualquer alteração da sua composição, pelo que os valores obtidos na simulação do reactor para o tempo final (t = 120 min) correspondem à concentração à entrada do decantador (mistura antes da separação). A concentração molar da mistura a separar é, então, $\boldsymbol{C} = [C_{\mathrm{TG}} \;\; C_{\mathrm{DG}} \;\; C_{\mathrm{MG}} \;\; C_{\mathrm{M}} \;\; C_{\mathrm{G}} \;\; C_{\mathrm{E}}] = [0.0018 \;\; 0.0188 \;\; 0.0550 \;\; 2.6181 \;\; 0.7644 \;\; 2.4219]$ mol/dm^3.

A fracção molar correspondente é dada por $\boldsymbol{x} = [x_{\mathrm{E}} \;\; x_{\mathrm{M}} \;\; x_{\mathrm{G}}] \approx [x_{\mathrm{E}} + x_{\mathrm{TG}} + x_{\mathrm{DG}} + x_{\mathrm{MG}} \;\; x_{\mathrm{M}} \;\; x_{\mathrm{G}}] = [0.42 \;\; 0.45 \;\; 0.13]$ (n/n). As fracções de tri-, di- e monoglicerídeos foram consideradas na fracção do componente éster por estarem em quantidades muito reduzidas e terem propriedades semelhantes.

Para gerar os dados experimentais referentes ao equilíbrio líquido-líquido, utilizaram-se diversas temperaturas da corrente de entrada. Para cada temperatura, construiu-se uma malha fazendo variar as fracções molares da mistura. Os intervalos considerados foram: $25 < T < 60$°C, $0.32 < x_{\mathrm{E,in}} < 0.52$ (n/n) e $0.35 < x_{\mathrm{M,in}} < 0.55$ (n/n). A fracção molar de glicerol à entrada do decantador, $x_{\mathrm{G,in}}$, foi calculada pela relação $\sum_{i=1}^{n_c} x_{i,\mathrm{in}} = 1$. O intervalo para a temperatura foi seleccionado tendo em conta as temperaturas de reacção típicas definidas por [13]. Fizeram-se incrementos de 1°C para a variável temperatura. O intervalo das composições foi definido tendo em conta uma gama de ± 0.10 (n/n) nas fracções $x_{\mathrm{E,in}}$ e $x_{\mathrm{M,in}}$ calculadas previamente. Percorreu-se a gama de composições definida com incrementos de 0.01 (n/n). No total, geraram-se 36 malhas de 405 pontos.

A normalização dos dados tem um papel fundamental no treino da rede neuronal. A utilização de dados de diferentes ordens de magnitude pode favorecer a atribuição de importâncias de ajuste diferentes durante a fase de treino da rede neuronal [7], pelo que se procedeu à normalização dos dados recorrendo aos valores indicados na Tabela 1.

Dados de Entrada	μ	σ
$x_{\mathrm{M,in}}$, (n/n)	0.44291	0.05769
$x_{\mathrm{G,in}}$, (n/n)	0.14301	0.07483
T, K	315.49	10.39

Tabela 1: Média (μ) e desvio padrão (σ) para a normalização dos dados de entrada.

Depois do pré-tratamento, os dados foram divididos aleatoriamente em três conjuntos: o conjunto de treino, o conjunto de validação e o conjunto de teste. Os conjuntos de treino e validação foram utilizados

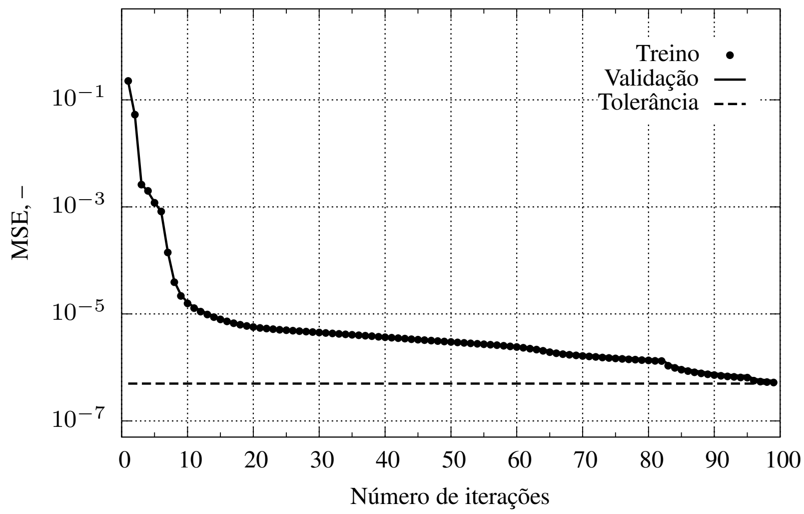


Figura 5: Evolução do erro médio quadrático durante o algoritmo de Levenberg-Marquardt.

no ajuste da rede neuronal. Já o conjunto de teste foi usado para simular a rede permitindo a comparação posterior entre os dados obtidos pelo método *flash* e os previstos pela rede neuronal.

4.2 Caracterização da rede neuronal

Após a geração dos dados, foi treinada uma rede neuronal para previsão das fracções de separação dos componentes da mistura ternária.

A rede neuronal desenvolvida para prever as fracções de separação dos componentes da mistura ternária foi estruturada em três camadas distintas. A camada de entrada possui três neurónios correspondentes às três variáveis de entrada na rede: $\boldsymbol{X} = [x_{\mathrm{M,in}} \ \ x_{\mathrm{G,in}} \ \ T]^\top$. Considera-se uma camada intermédia com cinco neurónios e uma camada de saída com três neurónios correspondendo às variáveis $\boldsymbol{Y} = [\xi_{\mathrm{E}} \ \ \xi_{\mathrm{M}} \ \ \xi_{\mathrm{G}}]^\top$. A Figura 3 mostra graficamente a estrutura da rede.

O algoritmo de treino descrito na Secção 3 encontra-se disponível no pacote de *software* `octave-nnet 0.1.13-2` para GNU Octave [15] e foi utilizado no treino da rede neural. Neste algoritmo, a inicialização dos pesos é feita através de elementos aleatórios uniformemente distribuídos no intervalo $[-1, 1]$. A taxa de aprendizagem inicial μ_0 é definida como 10^{-3} [12]. Outros parâmetros relacionados com o treino da rede neuronal foram especificados da seguinte forma: o número máximo de iterações foi 2×10^3, a tolerância foi 5×10^{-7} e o tempo máximo para treino foi 10^3 s.

4.3 Treino da rede neuronal

A rede neuronal com a estrutura descrita foi treinada, apresentando-se na Figura 5 (pontos) a evolução do erro quadrático médio ao longo das iterações do treino. O algoritmo Levenberg-Marquardt levou 99 iterações para atingir a tolerância especificada de 5×10^{-7} (linha a tracejado). O processo de treino demorou 20 s. A validação da rede é feita automaticamente pelo pacote de *software* utilizado. A Figura 5 compara o MSE resultante da validação ao longo das iterações (linha contínua) com o MSE do conjunto de pontos de treino, podendo constatar-se que são coincidentes.

Do treino da rede neuronal surgem as matrizes dos pesos $\boldsymbol{w}$ e $\boldsymbol{W}$ definidas por

$$\boldsymbol{w} = \begin{bmatrix} -0.280399 & -1.089354 & -0.085569 & 0.139296 & 0.074241 \\ -0.133304 & -0.569687 & -0.994122 & 0.322463 & 1.155133 \\ -0.477709 & 0.281168 & 0.030668 & -0.061436 & -0.017357 \end{bmatrix}$$

e

$$\boldsymbol{W} = \begin{bmatrix} 1.6266 \times 10^{-4} & 1.7624 \times 10^{-4} & 8.4364 \times 10^{-3} & 1.2702 \times 10^{-4} & 4.2978 \times 10^{-3} \\ 5.7984 \times 10^{-3} & 7.9335 \times 10^{-3} & 2.2665 \times 10^{-0} & -2.9864 \times 10^{-1} & 7.3118 \times 10^{-1} \\ -1.3643 \times 10^{-3} & 7.3683 \times 10^{-4} & 3.9128 \times 10^{-1} & 2.6466 \times 10^{-3} & 1.5635 \times 10^{-1} \end{bmatrix}$$

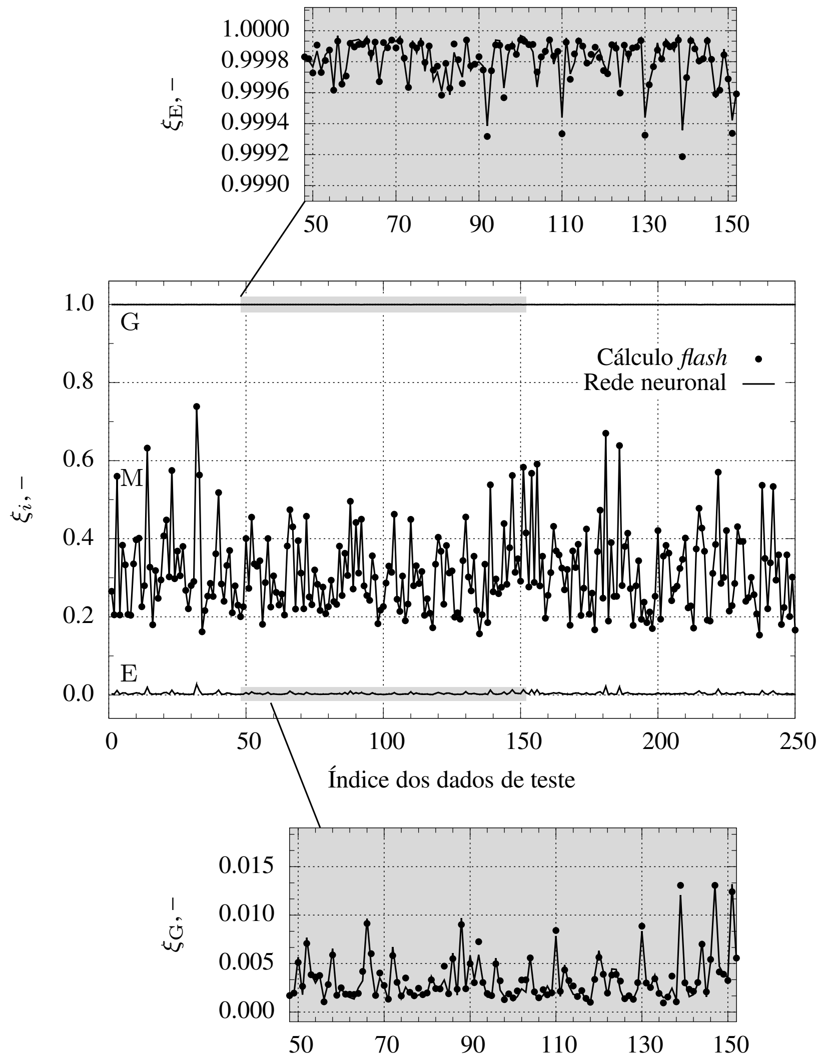


Figura 6: Previsão das fracções de separação através do modelo da rede neuronal.

e os vectores das bias $\boldsymbol{\theta}$ e $\boldsymbol{\Gamma}$ correspondendo a

$$\boldsymbol{\theta} = \begin{bmatrix} 0.62280 \\ 1.15884 \\ -2.46359 \\ 0.27300 \\ 2.24627 \end{bmatrix} \text{ e } \boldsymbol{\Gamma} = \begin{bmatrix} 1.00373 \\ 1.89116 \\ 0.23502 \end{bmatrix}.$$

Dada a motivação para o desenvolvimento de uma rede neuronal que substituísse de forma mais rápida mas igualmente eficaz o cálculo *flash* tradicional, procedeu-se a medidas dos tempos médios de computação necessários por cada um dos métodos. O tempo médio necessário à geração de um ponto pelo cálculo *flash* foi de cerca de 0.018 129 s, enquanto que utilizando a rede neuronal foi de aproximadamente 0.000 129 s. O tempo usado pela rede neural foi substancialmente menor que o tempo exigido pelo cálculo *flash* (cerca de 141 vezes menor), mostrando a adequação da opção feita à aplicação em questão.

4.4 Capacidade de previsão da rede

A Figura 6 mostra a previsão das fracções de separação usando a rede neuronal. Incluem-se igualmente os primeiros 250 pontos do conjunto de teste.

O metanol é o componente cuja fracção de separação apresenta maiores variações dentro da gama de composição e de temperatura cobertas pela malha. Por seu lado, o éster é o componente cuja fracção

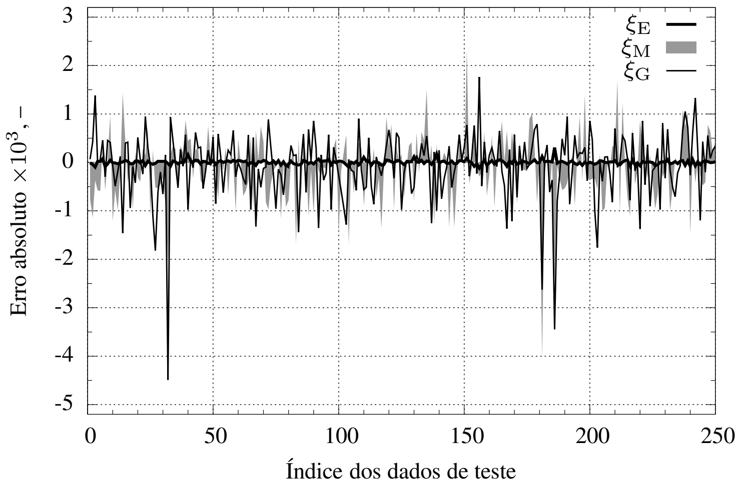


Figura 7: Erro absoluto entre a previsão da rede neuronal e a do cálculo *flash*.

de separação está menos dependente das condições da mistura inicial a tratar (isto é, da sua composição e temperatura). De facto, como se pode constatar na Figura 6, mais de 99.9% do éster vai sempre para a fase leve, independentemente das condições iniciais da mistura a separar. Finalmente, a fracção de separação do glicerol apresenta também uma variação mínima em função das condições iniciais de composição e temperatura da mistura que alimenta o decantador. O glicerol migra quase na totalidade para a fase pesada. Para permitir uma melhor apreensão dos dados, efectuaram-se duas ampliações do gráfico principal da Figura 6, uma relativa aos dados do componente glicerol e outra relativa aos dados do componente éster.

Os coeficiente de determinação correspondentes às previsões dos componentes metanol, éster e glicerol são, respectivamente: $R^2(\xi_M) = 0.9999$, $R^2(\xi_E) = 0.9137$ e $R^2(\xi_G) = 0.9676$. A previsão é especialmente boa no caso do metanol, já que este componente é mais sensível às condições de partida da mistura. Contudo, apesar de no caso do glicerol e do éster os coeficientes de determinação serem inferiores, repare-se que os erros absolutos entre as previsões e os valores experimentais são extremamente reduzidos (ver Figura 7).

O efeito da temperatura no equilíbrio líquido-líquido é muito importante. Para mostrar a capacidade de previsão deste efeito por parte da rede neuronal, estudaram-se os equilíbrios correspondentes a duas misturas A e B com composições distintas sujeitas a diferentes valores de temperatura.

A mistura A caracteriza-se por uma fracção molar $\boldsymbol{x}_{\text{in}} = [0.42 \;\; 0.45 \;\; 0.13]$, valores em consonância com os valores experimentais de [13]. O segundo estudo recai sobre uma mistura, designada por B, que traduz a situação de uma mistura resultante de uma eventual maior extensão da reacção química, isto é, uma mistura proveniente de uma reacção com maior rendimento do que aquela que deu origem à mistura A. Neste pressuposto, a mistura B foi definida como possuindo a composição $\boldsymbol{x}_{\text{in}} = [0.47 \;\; 0.43 \;\; 0.15]$.

Na Figura 8 representam-se as fracções de separação previstas pelo cálculo *flash* e pela rede neuronal em função da temperatura das duas misturas, sendo evidente a boa previsão da rede neuronal.

Como discutido acima, a fracção de separação do metanol varia bastante com a temperatura, ao invés das fracções de éster e de glicerol que se mantêm praticamente constantes. Uma ampliação considerável da representação gráfica destas duas fracções (ver Figura 8) revela o que, à primeira vista, poderia considerar-se como uma discrepância, nomeadamente no caso do éster. Contudo, tenha-se em mente que a "discrepância" é inferior a 0.006% (6 milésimas percentuais), sendo, portanto, desprezável.

Com o aumento da temperatura, as fracções de separação do metanol e do glicerol aumentam em ambas as misturas, embora de forma menos intensa no caso do glicerol. Na mistura B (mistura mais rica em éster), o metanol é mais solúvel na fase leve e, por isso, a fracção de separação é superior à obtida para a mistura A. Efeito semelhante é verificado para o glicerol. Por outro lado, ao aumentar a temperatura da mistura, o éster torna-se mais solúvel na fase pesada e, portanto, a fracção de separação de éster para a fase leve diminui.

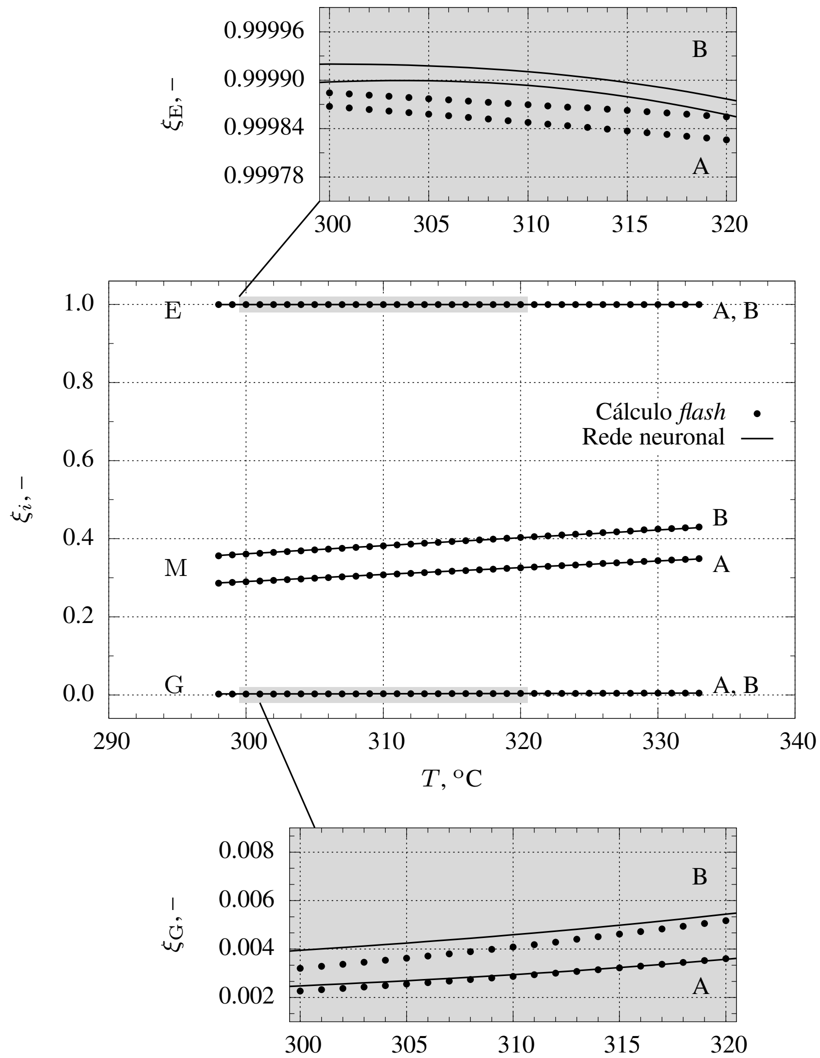


Figura 8: Previsão das fracções de separação em função da temperatura para duas misturas diferentes A e B (A: $\boldsymbol{x}_{\text{in}} = [0.42 \;\; 0.45 \;\; 0.13]$, B: $\boldsymbol{x}_{\text{in}} = [0.47 \;\; 0.43 \;\; 0.15]$).

5 Conclusões

Com o objectivo de obter uma ferramenta de optimização e controlo preditivo de um separador de fases na indústria de biodiesel, é necessário um modelo dinâmico mecanístico da unidade e, por conseguinte, uma forma de quantificar o equilíbrio líquido-líquido que aí se estabelece. O cálculo *flash* geralmente usado para descrever o equilíbrio líquido-líquido revela-se inadequado na presente situação por se tratar de um método iterativo. Assim sendo, optou-se por aproximar o cálculo *flash* por uma rede neuronal artificial do tipo *feedforward back-propagation*. A rede desenvolvida prevê a separação da mistura ternária inicial, composta por éster, metanol e glicerol, a partir da sua composição e temperatura.

Com esta aproximação eliminou-se um cálculo iterativo sujeito a uma condição de paragem baseada na comparação de previsões adjacentes e permitiu-se a utilização de ferramentas de diferenciação automática para facilitar a resolução de problemas de optimização não-linear. Estas vantagens foram conseguidas sem prejuízo da qualidade do modelo global uma vez que as previsões das fracções de separação obtidas pelo modelo da rede neuronal reproduzem bem os dados de equilíbrio obtidos pelo cálculo *flash*. Adicionalmente, o tempo de computação foi reduzido significativamente com a utilização da rede ao evitar o processo iterativo característico do cálculo *flash*.

Agradecimentos

Este trabalho teve apoio financeiro do QREN através do Programa Operacional Mais Centro e da União Europeia via FEDER no âmbito do projecto APCFAME com referência 3509/2009, em consórcio entre a Ciengis, SA e a Universidade de Coimbra. Os autores expressam ainda o seu agradecimento a Vitor Marques pelo acesso aos seus estudos preliminares nesta área.

Referências

[1] Biegler L. *Real-time PDE-constrained optimization.* Computational science and engineering, Society for Industrial and Applied Mathematics, 2007.

[2] Brásio AS, Romanenko A, Leal J, Santos LO, Fernandes NC. Nonlinear model predictive control of biodiesel production via transesterification of used vegetable oils. *Article submitted to Journal of Process Control.*

[3] Lobo LQ, Ferreira AGM. *Termodinâmica e Propriedades Termofísicas – Volume I: Termodinâmica das Fases.* Imprensa da Universidade de Coimbra: Coimbra, 2006.

[4] Walther A, Griewank A. Getting started with ADOL-C. *Combinatorial Scientific Computing*, Naumann U, Schenk O (eds.). chap. 7, Chapman-Hall CRC Computational Science, 2012; 181–202.

[5] Bell BM. CppAD: a package for C++ algorithmic differentiation. *Computational Infrastructure for Operations Research* 2012; **COIN-OR**(`http://www.coin-or.org/CppAD`).

[6] Fredenslund A, Jones RL, Prausnitz JM. Group-contribution estimation of activity coefficients in nonideal liquid mixtures. *AIChE Journal* 1975; **21**(6):1086–1099. DOI: 10.1002/aic.690210607.

[7] Chaturvedi D. *Soft Computing: Techniques and Its Applications in Electrical Engineering.* Studies in Computational Intelligence, Springer, 2008.

[8] Du X, Liu L, Xi X, Yang L, Yang Y, Liu Z, Zhang X, Yu C, Du J. Back pressure prediction of the direct air cooled power generating unit using the artificial neural network model. *Applied Thermal Engineering* 2011; **31**(14-15):3009 – 3014. DOI: 10.1016/j.applthermaleng.2011.05.034.

[9] Bishop C. *Neural Networks for Pattern Recognition.* Clarendon Press, 1995.

[10] Yegnanarayana B. *Artificial Neural Networks.* Prentice-Hall Of India Pvt. Limited, 2004.

[11] Sjöberg J. Neural Networks — Train and analyze neural networks to fit your data. *Technical Report*, Wolfram Research 2005. URL `http://media.wolfram.com/documents/NeuralNetworksDocumentation.pdf`, consulted in March 2013.

[12] Hagan M, Demuth H, Beale M. *Neural Network Design.* Electrical Engineering Series, Brooks/Cole, 1996.

[13] Bambase M, Nakamura N, Tanaka J, Matsumura M. Kinetics of hydroxide-catalyzed methanolysis of crude sunflower oil for the production of fuel-grade methyl esters. *Journal of Chemical Technology & Biotechnology* 2007; **82**(3):273–280. DOI: 10.1002/jctb.1666.

[14] Brásio AS, Romanenko A, Santos LO, Fernandes NC. Modeling the effect of mixing in biodiesel production. *Bioresource Technology* 2011; **102**(11):6508 – 6514. DOI: 10.1016/j.biortech.2011.03.090.

[15] Schmid MD. A neural network package for Octave – User's Guide 2009. URL `http://www.plexso.com/61_octave/neuralNetworkPackageForOctaveDevelop.pdf`.

Computational comparison of algorithms for a generalization of the node-weighted Steiner tree and forest problems

Raul Brás[†], J. Orestes Cerdeira[‡]

[†] *Instituto Superior de Economia e Gestão, Universidade Técnica de Lisboa, Portugal & Centro de Matemática Aplicada à Previsão e Decisão Económica*
E-mail:`rbras@iseg.utl.pt`

[‡] *Departamento de Matemática, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Portugal & Centro de Estudos Florestais, Instituto Superior de Agronomia, Universidade Técnica de Lisboa, Portugal*
E-mail:`orestes@isa.utl.pt`

Abstract

Habitat fragmentation is a serious threat for the sustainability of species. Thus, the identification of effective linkages to connect valuable ecological units is an important issue in conservation biology. The design of effective linkages should take into account that areas which are adequately permeable for some species' dispersal may act as obstructions for other species. The determination of minimum cost effective linkages is a generalization of both node-weighted Steiner tree and node-weighted Steiner forest problems. We compare the performance of different procedures for this problem using large real and simulated instances.

Keywords: Combinatorial optimization, Graphs, Heuristics, Minimum Steiner Trees.

1 Introduction

In conservation biology, habitat fragmentation is considered a key driver of biodiversity loss [Brooks et al.(2002), Hanski(2005)]. To mitigate the impacts of fragmentation on biodiversity, connectivity between otherwise isolated populations should be promoted [Merriam(1984)]. To effectively promote connectivity, there is need for procedures to identify linkages (i.e., areas to establish the connection) between habitats of each of several species (i) that take into account that linkage areas for a species might be barriers for others, and (ii) that are cost-efficient, since placing linkage areas under conservation compete with other land uses.

The problem can be formulated as follows. Consider a graph $G = (V, E)$ where the nodes of V identify the cells (usually grid squares) in which the study region has been divided, and which are considered suitable for conservation actions. The edges of E define adjacencies between pairs of cells (usually two cells are adjacent if they have a common border).

For each species (or group of "similar" species, i.e., sharing the same habitats and suitable areas to disperse) k, $k = 1, \ldots, m$, let T^k be the set of nodes representing the habitats of species k (*terminals* of type k), and V^k the set of nodes corresponding to cells which can be used as linkage passages for species k. We assume that $T^k \subseteq V^k$, and call k-barriers to the cells of $V \setminus V^k$.

A feasible solution of the problem is a subset of nodes $S \subseteq V$ that, for $k = 1, \ldots, m$, includes a path that only uses nodes of V^k between every pair of nodes in T^k.

Suppose there is a (non negative) cost associated to every node, quantifying the charge of allocating the corresponding cell to conservation purposes. The problem, which we will call multi-type linkage problem (MTLinkP for short), seeks for a minimum cost feasible solution (i.e., which minimizes the sum of the costs of the nodes).

MTLinkP, that was independently considered by [Lai et al.(2011)] and by [Alagador et al.(2012)], is a generalization of the node-weighted Steiner tree [Winter(1987)] and of the node-weighted Steiner forest [Duin and Volgenant(1987)] problems. If $V = V^k$ and $m = 1$ (i.e., only one species, no barriers) MTLinkP is the node-weighted Steiner tree problem. If $V = V^k$, for $k = 1, \cdots, m > 1$ (i.e., different species, no barriers) MTLinkP is the node-weighted Steiner forest problem.

[Lai et al.(2011)] and [Alagador et al.(2012)] proposed a heuristic procedure for MTLinkP by solving a sequence of node-weighted Steiner tree problems, one for each type k, and outcome the union of these

m Steiner solutions. We call this approach *type by type*. In the same paper [Lai et al.(2011)] presented a heuristic for MTLinkP that is a generalization of the primal-dual algorithm of [Demaine et al.(2009)] for the node-weighted Steiner forest problem, and report computational results on synthetic instances of small size (up to 15×15 grids) and m up to 5, and a real instance consisting of two species, 4514 cells and up to 17 terminals.

[Brás et al.(2012)] developed a computer application, MulTyLink, implementing a version of the type by type algorithm and a GRASP type heuristic. This software was announced in [Brás et al.(2013)], along with a brief description of the two algorithms used by the program.

In this paper we compare the performances of the primal-dual algorithm of [Lai et al.(2011)] and the two heuristics in MulTyLink on real and simulated data sets. We start giving in Section 2 a multiflow formulation for MTLinkP. In Section 3 we give some details on the two heuristics of MulTyLink, and summarize the primal-dual heuristic of [Lai et al.(2011)]. In Section 4 we report results of computational experiments to compare running times and quality of the solutions produced with the three heuristics. We finish with some remarks in Section 5.

2 Multiflow formulation

Flow formulations have been used to model Steiner tree problems (see, e.g., [Wong(1984)] for the standard edge-weighted Steiner tree problem; [Segev(1987)] for a special case of non negative costs on the edges and negative costs on the nodes and [Magnanti and Raghavan(2005)] for general network design problems with connectivity requirements including the edge-weighted Steiner forest problem).

Here we give a multi-commodity flow based formulation of MTLinkP.

Let $w_v \geq 0$ be a cost associated to each node v of graph G and let $w_v = 0$ if $v \in T^k$, for some $k = 1, \cdots, m$. Denote by A the set of arcs obtained by assigning opposite directions to every edge of G.

For $k = 1, 2 \ldots, m$, choose an arbitrary $t^k \in T^k$ as supply node for commodity k. Node t^k will supply every other node in T^k (the demanding nodes) with one unit of commodity k. Variables $f^k_{(u,v)}$ on arcs indicate the amount of commodity k (amount of flow of type k) along arc (u, v). Connectivity of the nodes of T^k in solutions is ensured by the mass balance constraints which state that the amount of commodity k out of a node v minus the commodity k into v must equal the supply/demand amount. In addition to flow variables $f^k_{(u,v)}$, binary variables x_v on nodes will be used to indicate whether node v is included ($x_v = 1$) or not ($x_v = 0$) in the solution. With these variables MTLinkP can be formulated as follows.

$$\min \sum_{v \in V} w_v x_v \tag{1}$$

subject to:

$$\sum_{\{u \in V^k : (t^k, u) \in A\}} f^k_{(t^k,u)} = |T^k| - 1, \ k = 1, \cdots, m \tag{2}$$

$$\sum\nolimits_{\{u \in V^k : (v,u) \in A\}} f^k_{(v,u)} - \sum\nolimits_{\{u \in V^k : (u,v) \in A\}} f^k_{(u,v)} = \begin{cases} 0 & \text{if } v \in V^k \setminus T^k \\ -1 & \text{if } v \in T^k \setminus \{t^k\} \end{cases}, \quad k = 1, \cdots, m \tag{3}$$

$$\sum_{\{u \in V^k : (u,v) \in A\}} f^k_{(u,v)} \leq \left(|T^k| - 1\right) x_v, \quad \begin{array}{l} v \in V^k \setminus \{t^k\}, \\ k = 1, \ldots, m \end{array} \tag{4}$$

$$x_v \in \{0, 1\}, \qquad v \in V \tag{5}$$

$$f^k_{(u,v)} \geq 0, \qquad u, v \in V^k, \text{with } (u, v) \in A, \ k = 1, \ldots, m. \tag{6}$$

The mass balance equations (2)-(3) dictate that one unit of flow of type k will be routed between the supply node t^k and each demanding node of $T^k \setminus \{t^k\}$. The "capacity" constraints (4) ensure there is no flow along the arcs entering nodes that are not included in the solution. Constraints (2)-(6) guarantee the existence of a directed path between t^k and every other node of T^k, thus ensuring that all nodes of T^k will be in the same connected component of the solution.

We will use the compact formulation (1)-(6) above to derive, from a mixed integer programming solver, MTLinkP optimal values for small size instances, to assess the quality of the solutions produced by the heuristic algorithms of Section 3.

3 Heuristics

3.1 Type by type heuristic

Given a permutation $i_1, i_2, \ldots, i_m$ of (integer) types $1, 2, \ldots m$, the *type by type* heuristic computes in step k a Steiner solution with respect to $< V^{i_k} >$, the subgraph of G induced by V^{i_k}, updates the costs of nodes v of that solution letting $w_v = 0$, and proceeds to the next step $k+1$. The final MTLinkP solution results from turning minimal feasible (with respect to inclusion) the union of nodes of the Steiner solutions obtained in each step.

The process can be repeated for a number of different permutations of integers $1, 2, \ldots, m$, and the best solution is returned (see Figure 1).

Figure 1: TbT heuristic

1. $Sol \leftarrow \emptyset$
2. $w(Sol) \leftarrow \infty$
3. $\mathcal{P} \leftarrow$ subset of permutations of $\{1, \ldots, m\}$.
4. For all $P \in \mathcal{P}$
   (a) $X \leftarrow \emptyset$
   (b) For all $k \in P$
      i. Build graph $< V^k >$ with weights: 0 if $v \in X$, w_v otherwise.
      ii. $X^k \leftarrow$ Steiner solution w.r.t. $< V^k >$ and T^k
      iii. $X \leftarrow X \cup X^k$
   (c) Turn X minimal. For each $v \in X \setminus \bigcup_k T^k$ (randomly ordered) remove v from X if $X \setminus \{v\}$ is MTLinkP feasible.
   (d) $w(X) \leftarrow \sum_{v \in X} w_v$
   (e) If $w(X) < w(Sol)$ then
      i. $Sol \leftarrow X$
      ii. $w(Sol) \leftarrow w(X)$
5. Return Sol.

To solve the node-weighted Steiner tree problem in each step, we use the following straightforward modification of the well known *distance network heuristic* suggested by [Kou et al.(1981)] for the edge-weighted Steiner tree. If H is a graph with costs w on the nodes and terminal set S, we define the distance network $D(S)$ which is the complete graph with S as its node set, and where the weight of every edge (u, v) is the cost of the minimum cost path connecting terminal u to v on H. Note that determining a node-weighted shortest path on undirected graph H between nodes u and v, with $w_u = w_v$, reduces to finding an edge-weighted shortest path from u to v in the digraph obtained assigning opposite directions to every edge of H, and defining the cost of every arc (i, j) as being equal to w_j.

A minimum spanning tree of $D(S)$ is determined and a (node-weighted) Steiner solution N is defined as the set of the nodes of the shortest paths corresponding to the edges of the spanning tree.

In the final step the nodes of N are considered randomly and node j is removed from N if all nodes of S belong to the same connected component of the subgraph of H induced by $N \setminus \{j\}$.

We use the above modification of the *distance network heuristic* since it is fast and does not use large data structures. Procedures such as [Klein and Ravi(1995)] heuristic, based on the Rayward-Smith [Rayward-Smith(1983)] algorithm, that perform well for node-weighted Steiner problems, would be impractical for the large size instances of MTLinkP we want to handle. [Klein and Ravi(1995)] heuristic needs to compute the minimum cost paths between all pairs of nodes, which is time consuming and requires large amounts of memory.

[Lai et al.(2011)] version of this heuristic uses the Dreyfus-Wagner (DW) algorithm [Dreyfus and Wagner(1971)] to solve the Steiner problem at each step. DW is an exact dynamic programming algorithm that runs in exponential time, not suitable to solve the instances that we present in this paper.

3.2 Primal Dual heuristic

[Lai et al.(2011)] gave a modified version of the [Demaine et al.(2009)] heuristic for node-weighted Steiner forest problems. The heuristic operates on the following cut-covering formulation of MTLinkP. Minimize

(1) subject to (5), and

$$\sum_{v \in \Gamma^k(S)} x_v \geq f^k(S), \qquad \begin{array}{l} S \subseteq V^k \\ k = 1, \ldots, m \end{array} \tag{7}$$

where $f^k(S) = 1$, if $\emptyset \neq S \cap T^k \neq T^k$ (i.e., if S includes at least one terminal of T^k, but not all), and $f^k(S) = 0$, otherwise, and $\Gamma^k(S)$ is the set of nodes $v \in V^k \setminus S$ adjacent to at least one node in S.

The dual of the linear relaxation of (1),(5),(7) is:

$$\max \sum_{k=1}^{m} \sum_{S \subseteq V^k} f^k(S) y^k(S)$$

subject to:

$$\sum_{k=1}^{m} \sum_{S \subseteq V^k : v \in \Gamma^k(S)} y^k(S) \leq w_v \quad v \in V$$

$$y^k(S) \geq 0 \qquad \begin{array}{l} S \subseteq V^k \\ k = 1, \ldots, m \end{array}$$

The heuristic maintains an infeasible primal solution X, and dual variables $y^k(S)$. The algorithm is described in Figure 2, where $\mathcal{C}(\langle X^k \rangle)$ are the connected components of the graph induced by $X^k = X \cap V^k$.

Figure 2: PD heuristic

1. $X \leftarrow \bigcup_k T^k$
2. For $k = 1, \ldots, m$ calculate $\mathcal{C}(\langle X^k \rangle)$. $y^k(C) \leftarrow 0$ for every $C \in \mathcal{C}(\langle X^k \rangle)$
3. While X is not MTLinkP feasible
   (a) Simultaneously increase $y^k(C)$ until, for some v, $\sum_{k=1}^{m} \sum_{C \subseteq V^k : v \in \Gamma^k(C)} y^k(C) = w_v$.
   (b) $X \leftarrow X \cup \{v\}$
   (c) For $k = 1, \ldots, m$ recalculate X^k and $\mathcal{C}(\langle X^k \rangle)$.
4. Turn X minimal. For each $v \in X \setminus \bigcup_k T^k$ (by reverse order of insertion) remove v from X if $X \setminus \{v\}$ is feasible
5. Return X.

3.3 GRASP heuristic

The type by type (TbT) heuristic and the primal-dual heuristic (PD) of [Lai et al.(2011)] define a feasible solution adding in each step nodes to a current unfeasible solution X. TbT heuristic adds to X a set of nodes that guarantee the connection of all terminals from a certain predetermined type, and assigns costs equal to zero to all the added nodes. PD heuristic adds to X one single node that belongs to V^k and is adjacent to X^k, for at least one not previously determined type k. We present a kind of greedy randomized adaptive search procedure (GRASP) [Feo and Resende(1995)] that hybridizes the two heuristics.

GRASP starts with set X consisting of all terminals of T^k, $k = 1 \cdots, m$, and in each step grows the current unfeasible solution X as follows. First, some type k, for which not all terminals of T^k are connected, is uniformly selected. Next, a connected component S of $< X^k >$, the subgraph induced by X^k, that includes terminals of type k, is uniformly selected, and a minimum cost path P, among the minimum cost paths connecting S with every other component of $< X^k >$ that includes terminals of type k, is determined. The nodes of P are added to X, and costs are updated letting $w_v = 0$ to every node v of P. Note that, since the costs of nodes of X are all equal to zero, P can be easily obtained with Dijkstra algorithm [Dijkstra(1959)], choosing an arbitrary node in S as the starting node and ending as soon as a node of a component of X^k, including terminals of T^k and different from S, is added to the path.

The final GRASP solution is obtained by turning minimal feasible the solution X obtained in the last step. Given its random behavior, repeating GRASP a number of times with the same input is likely to produce different solutions, and the best solution is outcome (see Figure 3).

Figure 3: GRASP heuristic

1. $w\,(Sol) \leftarrow \infty$
2. $r \leftarrow$ number of repetitions
3. For $i = 1, \ldots, r$
   (a) $X \leftarrow \bigcup_k T^k$
   (b) While X is not MTLinkP feasible
      i. $X^k = X \cap V^k$. Calculate $\mathcal{C}(\langle X^k \rangle),\ \ k = 1, \ldots, m$
      ii. $Q = \{k$: not all terminals of T^k belong to the same $S \in \mathcal{C}(\langle X^k \rangle),\ k = 1, \ldots, m\}$
      iii. If $Q = \emptyset$ then end the while cycle
      iv. $p \leftarrow$ member of Q uniformly selected
      v. $S \leftarrow$ member of $\mathcal{C}(\langle X^p \rangle) : S \cap T^p \neq \emptyset$ uniformly selected
      vi. $P \leftarrow$ minimum cost path connecting S to $U \in \mathcal{C}(\langle X^p \rangle) \setminus S : U \cap T^p \neq \emptyset$
      vii. $X \leftarrow X \cup P,\ \ w_v = 0,\ \forall v \in P$
   (c) Turn X minimal. For each $v \in X \setminus \bigcup_k T^k$ (randomly ordered) remove v from X if $X \setminus \{v\}$ is MTLinkP feasible.
   (d) $w(X) \leftarrow \sum_{v \in X} w_v$
   (e) If $w(X) < w(Sol)$ then
      i. $Sol \leftarrow X$
      ii. $w(Sol) \leftarrow w(X)$
4. Return Sol.

4 Computational experiments

We performed computational tests to evaluate the quality of the solutions produced by the heuristics, as well as the practicality of the flow formulation of Section 2.

4.1 General case

Here we report results for the case where not all V^k coincide.

4.1.1 Data

We used real and simulated instances to test the heuristics.

1) Real data. Real data concerns the linkage of climatically-similar protected areas (PA) in the Iberian Peninsula (IP). IP is represented as 580,696 1km $\times$ 1km cells, from which 80,871 cells intersecting the 681 existent PA in the IP were defined as terminals. Terminals were clustered in four groups sharing similar climates (with respect to four climatic variables which are considered important drivers of species' distributions). Adjacency was considered in terms of common edges or corners of the square cells.

Cells with considerable human intervention (values derived from Human Footprint Index data available from http://www.ciesin.columbia.edu/wild_areas greater than 60 in a range from 0 to 100) were excluded as they were considered poorly permeable to species' movements. This has reduced the number of cells to 438,948 (which includes every protected cell).

Figure 4 shows the location of protected cells from each class, and cells that were excluded due to presenting high levels of human intervention.

For $k = 1, \ldots, 4$, V^k was defined as the set of cells that do not significantly differ from the mean climatic conditions of PA of class k. This was delineated as follows. The centroid, in the climatic space, of the PA cells of each climatic class was defined, and the Euclidean distances from the climate conditions of each cell to the centroid of each class were computed. This retrieved four values $d^k(v)$, for each cell v, expressing the dissimilarity of cell v to every climatic class k. Cell $v \in V^k$ (i.e., v was not considered k-barrier) if $d^k(v)$ is below a certain threshold value B^k. Two scenarios were considered. In scenario 1, B^k was defined as the largest dissimilarity $d^k(v)$, among the protected cells v in every PA from class k. In scenario 2, B^k was set as the third quartile of the $d^k(v)$ values for protected cells v of class k. Cell u was included in V^k (i.e., u was not considered k-barrier) if u belongs to some PA of class k, or $d^k(u) < B^k$.

The rational for the identification of linkages between climatically-similar protected areas, free from climatic barriers, stands on the assumption advocated by [Alagador et al.(2012)] that species with similar ecological requirements occupy the same environments. Thus, linking climatically-similar protected areas is an effective way to promote the dispersal of species, counteracting in part the negative effects of fragmentation.

A cost was assigned to every non protected cell that is proportional to the cell's fraction not covered by Natura 2000 Network (w_v = (100−percentage of Natura 2000 Network covered by v)/100). The Natura 2000 network is a European scaled conservation scheme designed to complement nationally-defined protected areas. We assigned cost equal to zero to every protected cell.

Details on the IP data can be found in [Alagador et al.(2012)].

We denote by $IP1$ and $IP2$ the IP instances under scenarios 1 and 2, respectively.

2) Simulated data. Simulated data were generated as follows. Each node of V is a cell from a $n \times n$ grid. Two cells are adjacent if they have a common edge or corner.

To define V_k we start by uniformly selecting an integer $s \in [0, m]$ and assume that species $1, \ldots, s$ are "specialist" (can only thrive in a narrow range of environmental conditions) and species $s + 1, \ldots, m$ are "generalist" (are able to thrive in a wide variety of environmental conditions). Each node v of V is included in V_k with probability 1/4 for each "specialist" species $k \leq s$, and with probability 3/4 for each "generalist" species $k > s$. The number of terminals of each type was obtained from a discrete uniform distribution in interval $[2, \max\{|V|/1000, 5\}]$, and terminals chosen uniformly among the nodes of V^k.

We assigned to every node in $V \setminus (\cup_k T^k)$ a cost from an uniform distribution in $[0, 1]$, and cost zero to every node of T^k.

We generated small instances with $n = 10, 20, 30, 40, 50$ and $m = 2, 3, 4, 6, 8, 10$ and large instances with $n = 100, 200, 300, 400, 500$ and $m = 4, 6, 8, 10$.

For the same values of n and m we generated 10 instances. This gave a total of 500 instances.

4.1.2 Results

Here we report the main results of the computational tests that we carried out.

Heuristics were implemented in C++, using the Boost Graph Library [Siek et al.(2002)] to calculate spanning trees, shortest paths and connected components. Parallel programming was not used and so they ran in a single thread. All times refer to elapsed times. The computers were dedicated to running the instances, so that elapsed time is close to CPU time. Solutions for the Iberian Peninsula data were obtained with a Intel Core2 Quad CPU Q9450 @2.66GHz and 4GB of memory machine, while for simulated data the solutions were obtained in a machine with 2 AMD Opteron 6172 processors (24 cores) @2.1GHz and 64GB of memory.

1) Real data. Table 1 displays results obtained for the Iberian Peninsula's data with each of the three heuristics. The first column identifies the problem instance (scenarios 1 and 2). Each of the three pairs of the remaining columns contains the value of the solution obtained with a heuristic: GRASP, type by type (TbT) and primal-dual (PD), and the corresponding running time in seconds. The TbT heuristic ran for every permutation of the $m = 4$ types, while GRASP was limited to two hours of execution. We let the program finish the current repetition i, if it has started before the time expired, thus computation times can exceed 7200 seconds. PD was not time-limited in order to produce a solution.

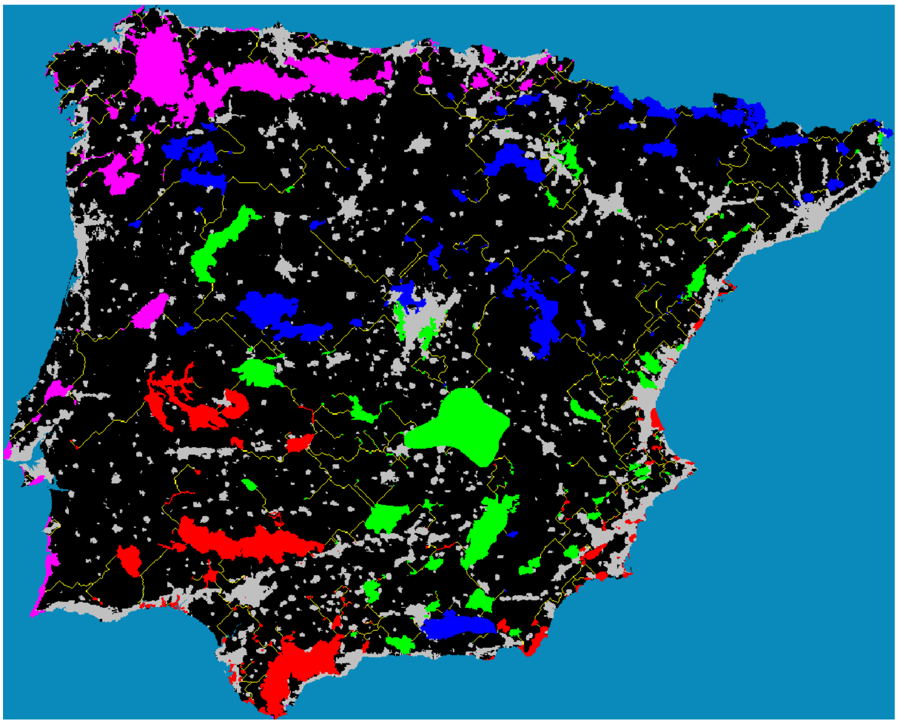
Table 1: Results for the Iberian Peninsula.

Instance	GRASP		TbT		PD	
	cost	time	cost	time	cost	time
$IP1$	2024.67	7782.55	2035.73	5012.39	2162.11	544490.00
$IP2$	2121.03	7525.25	2148.49	7075.90	2167.62	347003.00

GRASP obtained the best solutions. The costs of the solutions produced by TbT were slightly higher, but the times to run the 24 permutations of the four types were lower than the two hours that limited the execution of GRASP. PD had a poor performance. Long computation times were necessary to obtain solutions with costs that are greater than those of the solutions obtained with GRASP and with TbT. This negative behavior of PD can be explained by the specific structure of these graphs. Nodes which are far apart on the grid are connected by long paths. Thus, it is likely that PD includes a large number of redundant nodes until a feasible solution is reached. Solutions with many redundant nodes are difficult to turn minimal. The process is time consuming and produces poor solutions. GRASP and TbT, in each step, add to the solution that is being constructed the nodes of a minimum cost path connecting a pair of terminals. Thus, the number of redundant nodes is likely to be much less than that generated by PD.

Figure 4 shows a solution, obtained with the TbT heuristic, for the IP2 instance. Protected cells are colored red, green, blue and magenta and the cells of the solution are yellow. Grey areas represent cells not in V (human footprint over 60). For a detailed interpretation of the solution, knowledge of the location of the barriers from each type would be needed.

Figure 4: Iberian Peninsula data (scenario 2). Protected cells are colored red, green, blue and magenta. Grey areas represent cells not in V. Cells of the solution obtained with TbT heuristic are colored yellow.



2) Simulated data. The main results derived with small and large instances for simulated data are given in Tables 2 and 3, respectively. Recall that 10 instances with the same values of $|V|$ and m were considered and, therefore, each row of Tables 2 and 3 summarizes the results of 10 instances.

We established common elapsed time limit values for the heuristics. One minute for small instances and 30 minutes for large instances, but we allowed GRASP to finish the current repetition i, and TbT to finish the current permutation P.

For most of small instances we were able to obtain optimal solutions from the flow formulation (1)-(6), using CPLEX 12.4, with parallel mode set to opportunistic and 24 parallel threads (all other options used default values). For each instance, CPLEX execution time-limit was set to 1 hour of elapsed time, meaning up to 24 hours of CPU time since the machine has 24 cores.

In Table 2 column *#Opts* indicates the number of instances for which optimal solutions were found. The two columns labeled *% dev. from* indicate, for each heuristic, the mean of the relative deviations (in percentage) from the optimal values (*opt*) and from the best values of the heuristic solutions (*bestH*). The relative deviation is calculated by the expression $100(h - w^*)/w^*$, where h is the value of the heuristic solution, and w^* is the optimal value (*opt*), or the minimum of the values of the three heuristic solutions (*bestH*), respectively. The number of optimal values with respect to which averages were computed is given in column *#Opts*. Columns *best* report the number of instances for which the heuristic found the best value among the values of the three solutions obtained for the same instance with the three heuristics. Columns *time* indicate the mean computation times (in seconds) for TbT and PD. The computation times are not reported for GRASP since it uses all the amount of time allowed.

Table 3 is similar except that there are no columns regarding optimal values, since CPLEX was unable to handle the large instances. Thus, columns *% dev.* and *best* refer to comparisons with the best values of heuristic solutions.

Table 2: Results for small instances.

$\|V\|$	m	#Opts	GRASP			TbT				PD			
			% dev. from			% dev. from				% dev. from			
			opt	bestH	best	opt	bestH	best	time	opt	bestH	best	time
100	2	10	1.67	1.67	9	1.67	1.67	9	0.00	2.15	2.15	9	0.00
	3	10	1.66	1.66	9	1.66	1.66	9	0.01	0.00	0.00	10	0.00
	4	10	4.02	1.94	8	3.87	1.79	9	0.06	6.49	4.20	6	0.00
	6	10	2.32	2.17	9	3.31	3.16	8	5.19	1.61	1.48	7	0.00
	8	10	1.21	0.00	10	1.21	0.00	10	44.43	2.06	0.86	5	0.01
	10	10	0.31	0.31	9	1.27	1.27	7	60.51	0.75	0.75	8	0.01
400	2	10	0.34	0.00	10	2.51	2.18	5	0.01	7.53	7.17	5	0.02
	3	10	1.80	0.25	9	4.55	2.90	5	0.01	11.47	9.71	4	0.02
	4	10	1.60	0.00	10	4.48	2.79	5	0.06	9.02	7.27	3	0.03
	6	10	2.08	0.95	8	3.56	2.38	7	1.49	9.06	7.87	1	0.04
	8	10	0.90	0.10	9	3.70	2.87	6	18.21	9.50	8.62	2	0.06
	10	10	2.24	0.00	10	6.33	3.95	3	43.46	11.20	8.69	0	0.15
900	2	10	0.85	0.00	10	2.60	1.72	6	0.01	5.95	4.96	6	0.07
	3	10	0.54	0.00	10	1.82	1.28	8	0.02	8.15	7.50	5	0.09
	4	10	1.34	0.00	10	4.66	3.26	4	0.08	15.41	13.82	3	0.14
	6	8	2.71	0.19	9	6.54	4.55	3	3.57	19.06	15.32	2	0.27
	8	7	2.86	0.00	10	4.70	2.05	2	14.31	13.09	13.72	0	0.37
	10	7	1.82	0.39	8	3.25	2.75	5	42.45	10.26	11.07	2	0.56
1600	2	7	1.49	0.22	9	1.92	1.34	4	0.02	7.93	7.66	3	0.22
	3	5	0.09	0.00	10	0.49	4.01	4	0.05	0.65	8.51	4	0.36
	4	6	0.69	0.20	8	1.17	2.38	5	0.21	5.92	10.07	5	0.53
	6	5	1.52	0.43	8	1.93	1.13	5	4.33	4.39	16.84	3	0.74
	8	5	0.29	0.00	10	1.49	2.80	3	12.44	10.56	11.81	2	0.77
	10	1	0.00	0.06	9	1.47	4.27	1	36.13	3.32	15.09	0	1.96
2500	2	4	0.77	0.00	10	0.77	3.35	5	0.04	0.77	5.62	4	0.75
	3	6	0.00	0.08	9	0.98	1.11	6	0.07	0.98	6.91	4	0.67
	4	6	1.52	0.51	9	0.64	2.39	6	0.24	1.96	6.48	4	0.78
	6	4	5.05	0.15	8	5.57	2.41	3	6.01	4.65	7.39	4	1.71
	8	2	1.24	0.00	10	3.35	3.83	2	17.60	22.30	15.73	2	2.60
	10	0		0.04	8		1.59	6	16.20		16.19	3	2.60

GRASP was clearly superior for the instances considered, while PD had a poor performance.

For small instances the average over the 30 values of column *% dev. from opt* in Table 2 was 1.48 for GRASP, 2.81 for TbT and 7.11 for PD. Only four of these 30 values exceeded 2.5% for GRASP, while six values exceeded 4.5% for TbT. GRASP was a best heuristic in at least eight instances out of the 10 with the same $|V|$ and m. Considering all the 300 small instances, GRASP was a best heuristic in 275, TbT in 161 and PD in 116.

For the large instances the superiority of GRASP was even more evident. It has obtained the best results in 198 out of 200 instances. The mean relative deviations between TbT results and the best heuristic values were always below 5.3%, but it attained the best result only on 24 instances.

Results on simulated data confirmed the bad behavior of PD with this kind of instances. For $|V| \geq 90000$, we were unable to find solutions within the time limit of 30 minutes, except for a few instances. These were not considered in order to not bias the analysis of the results. The corresponding entries are blank on Table 3. In general, solutions were of poor quality. It seems that PD has difficulties dealing with instances where graphs have the structures here considered. An explanation was previously given when analyzing the results on the IP instances.

A fact that should be mentioned is that, several times, TbT succeeded to complete its computations within the time limits established, despite the relative high values of m ($m = 8, 10$). This is justified by the way instances were generated. Each cell of the $n \times n$ grid belongs to V^k with probability $1/4$ for "specialized" species k and $3/4$ for "generalist" species k. Thus, it may happen that all components of the subgraph induced by V^k, particularly for "specialized" species k, include at most one terminal of T^k, i.e., every path connecting any two terminals of T^k include some k-barrier. In this case there is no need to consider species k, as no two terminals of T^k can be linked in V^k. Since "specialist" species were uniformly chosen among the m species, the number of species that needs to be considered might be much smaller than m.

4.2 Case where all V^k coincide

MTLinkP is a generalization of the node-weighted Steiner tree and of the node-weighted Steiner forest problems. Therefore, GRASP, TbT and PD can be used, with no modification, to solve those problems.

Table 3: Results for large instances.

$\|V\|$	m	GRASP		TbT			PD		
		%dev.	best	%dev.	best	time	%dev.	best	time
10000	4	0.00	10	5.27	1	1.56	18.17	1	43.71
	6	0.00	10	4.55	1	49.93	19.91	1	89.57
	8	0.00	10	3.43	2	214.82	11.90	2	67.37
	10	0.00	10	4.86	0	766.98	21.31	0	148.57
40000	4	0.00	10	3.13	2	52.95	14.74	2	1292.83
	6	0.00	10	3.50	0	732.16	7.55	0	1116.29
	8	0.00	10	3.44	1	982.21	14.08	1	1709.82
	10	0.00	10	3.59	1	1320.94	13.60	1	1998.23
90000	4	0.00	10	2.46	3	185.65			
	6	0.00	10	2.21	2	1056.98			
	8	0.00	10	3.66	0	1654.99			
	10	0.13	9	3.73	1	1823.81			
160000	4	0.00	10	2.24	4	819.54			
	6	0.00	10	2.42	2	1325.13			
	8	0.03	9	2.34	2	1931.62			
	10	0.00	10	2.14	0	1947.12			
250000	4	0.00	10	3.26	0	1378.50			
	6	0.00	10	2.45	0	1821.26			
	8	0.00	10	2.76	0	2053.90			
	10	0.00	10	2.09	2	2297.54			

We carried out some computational tests to assess how the heuristics perform on solving node-weighted Steiner forest problem.

For node-weighted forest problem, the PD heuristic is the [Demaine et al.(2009)] algorithm. Another heuristic, based on the Rayward-Smith algorithm [Rayward-Smith(1983), Rayward-Smith and Clare(1986)] that performs well in practice for node-weighted Steiner forest is the [Klein and Ravi(1995)] heuristic.

Klein and Ravi heuristic (KR) begins by computing the matrix M of the costs of minimum cost paths between every pair of nodes in V. Then, starting with X consisting of all terminals of T^k, $k = 1, \cdots, m$, in each step, KR adds to X the nodes of certain paths that connect a number of connected components of $< X >$, the subgraph induced by X. The connected components to merge are selected from the values of a function f that is calculated as follows, for every node $v \in V$. Let $\mathcal{S}$ be the set components of $< X >$ that, for some k, includes at least one node of T^k but not all, and let $\mathcal{S}_r$ be the family of all r sets of $\mathcal{S}$ (i.e., if $S_r \in \mathcal{S}_r, |S_r| = r$). For every $v \in V$ and $S_r \in \mathcal{S}_r$, let $w(v, S_r)$ be the sum of the costs of minimum cost paths connecting v with each of the r components in S_r. For every $v \in V$, define $f(v,r) = \min_{S_r} w(v, S_r) - (r-1)w_v$. The value $f(v,r)$ is the minimum cost of merging r components of $\mathcal{S}$ with r paths rooted at v. Note that the computation of $f(v,r)$ can be quickly achieved from matrix M. Finally, $f(v) = \min_{2 \le r \le |\mathcal{S}|} f(v,r)/r$, which is the minimum of the mean values of $f(v,r)$ with respect to r. In each step, KR adds to X the nodes of the paths from v which minimizes $f(v)$, while $\mathcal{S}$ is not the empty set. When there are no more components to merge, the heuristic proceeds turning solution X minimal.

We compared the performances of GRASP, TbT, [Demaine et al.(2009)] (PD) and [Klein and Ravi(1995)] (KR) heuristics on instances generated as above for simulated data, except that $V^k = V$, for $k = 1, \ldots, m$. We considered $n \times n$ grid graphs with $n = 50, 100, 200$ and $m = 2, 4, 6, 8, 10$ types of terminals. For each n and m two instances were generated. Table 4 reports costs and times (in seconds) on each instance. GRASP and TbT heuristics were restricted to 30 minutes of execution time. Computations were processed with the same machine that was used for the simulated data.

Results for GRASP and KR were very similar. KR obtained the best result in 56.7% of the cases, while GRASP was the best heuristic in 40.0% of the cases and PD in one case (3.3%). TbT never obtained the best result. The mean relative gap between the value vH obtained by the heuristic H and the value vKR obtained with KR, given by $(vH - vKR)/vKR$, was 0.5% for H=GRASP, 4.4% for H=TbT and 3.5% for H=PD. Considering only the cases for which KR performed better than heuristic H ($vKR < vH$), the mean of the relative gap was 3.1% for H=GRASP, 5.2% for H=TbT and 3.9% for H=PD.

Results showed that GRASP performed better than KR in smaller instances, while in general KR outperforms GRASP for larger ones. However, the relative gap did not exceed 8.1% (for an instance where $n = 200$ and $m = 8$). The values obtained by TbT were slightly greater than those produced by GRASP. This was more evident for larger instances ($n = 200$). PD obtains good results in the larger instances. For $n = 200$ and $m \geq 4$ it obtains better results than GRASP with relatively small times of execution. KR heuristic maintains in memory matrix M of the costs of minimum cost paths between every pair of nodes in V. For $n = 200$ 6GB of memory are needed, and for $n = 300$ 30 GB are needed.

Table 4: Results for Steiner forest.

$\|V\|$	m	GRASP		TbT		PD		KR	
		cost	time	cost	time	cost	time	cost	time
2500	2	19.32	1800.28	20.12	0.11	20.09	1.15	19.20	2.04
		12.61	1800.37	13.01	0.08	15.92	2.34	14.67	1.92
	4	24.45	1800.18	25.31	2.64	26.36	1.57	25.04	2.68
		22.78	1800.48	23.23	2.46	25.53	1.87	22.98	2.33
	6	24.39	1800.65	24.83	104.46	25.38	0.82	24.64	2.73
		30.03	1800.47	30.74	98.04	31.56	1.47	30.44	2.85
	8	36.56	1801.00	38.66	1800.60	38.32	1.67	36.86	4.58
		29.65	1800.23	30.27	1800.37	30.50	0.81	29.97	3.57
	10	34.30	1800.63	37.12	1801.13	36.90	1.82	35.51	3.82
		33.46	1801.23	34.80	1800.23	33.51	1.01	33.30	4.71
10000	2	46.64	1800.54	48.39	0.73	48.02	17.43	46.83	39.35
		42.19	1800.91	43.96	0.78	43.16	25.60	41.97	48.66
	4	62.06	1800.88	65.60	18.54	65.36	26.49	61.77	58.09
		47.44	1801.16	50.20	12.94	49.05	22.79	47.35	48.61
	6	78.49	1800.33	79.95	956.15	81.00	42.66	77.16	64.57
		67.07	1801.34	68.97	655.02	68.02	18.02	67.31	60.66
	8	74.07	1800.78	75.25	1801.68	75.34	17.31	71.36	72.82
		79.73	1800.39	81.12	1801.94	82.17	22.02	78.56	82.25
	10	100.02	1801.10	100.70	1802.77	99.83	59.72	98.07	96.04
		90.77	1801.33	95.66	1802.23	92.31	32.39	89.00	96.32
40000	2	178.80	1801.51	192.27	9.94	189.42	501.73	189.39	1541.14
		83.15	1801.27	87.51	3.47	89.15	328.08	84.28	1154.73
	4	257.27	1806.26	265.71	190.69	253.92	1155.15	273.20	2280.98
		188.07	1800.93	197.61	151.75	187.33	349.34	182.06	1666.43
	6	337.71	1813.55	348.49	1814.00	327.30	882.97	317.66	4884.35
		225.93	1803.42	236.66	1806.29	228.74	829.09	218.27	2187.45
	8	316.43	1815.54	327.89	1821.25	303.15	686.42	292.69	4239.43
		305.71	1802.64	322.56	1812.95	300.84	515.96	291.56	3982.47
	10	297.26	1813.47	310.58	1813.78	290.70	885.06	283.34	3827.51
		372.83	1804.53	386.94	1817.54	360.27	597.45	345.79	6483.80

Thus, KR heuristic could not be used for the real IP instances.

Given the above limitations of KR, GRASP and PD appear to be good options to solve large node-weighted Steiner forest problems for the type of graphs here considered.

5 Conclusions

In this paper we considered a mixed integer flow formulation and three heuristics for MTLinkP. The flow based formulation only permited to solve instances up to 2500 nodes, which is far below the size of the instances that occur in the context of conservation biology. For the specific structure of graphs of the instances that occur in conservation, GRASP seems to be a good option. Producing different solutions from different runs, on reasonable times, is relevant since, rather than a single solution, decision making needs to consider different options before proceeding negotiations with stakeholders. There are many issues (e.g., socioeconomic) involved in the analysis of conservation actions which are not easily quantifiable, thus having different alternatives to choose is an important feature.

Acknowledgements

We are grateful to Maria João Martins and Diogo Alagador for discussion and assistance, and to two referees for many suggestions on a previous version of the manuscript. Both authors were supported by the Portuguese Foundation for Science and Technology (FCT). R. Brás was funded by the project PEst-OE/EGE/UI0491/2011 and the *Centro de Matemática Aplicada à Previsão e Decisão Económica* (CEMAPRE) under the FEDER/POCI Programme. J. O. Cerdeira was funded through the project PEst-OE/AGR/UI0239/2011 and the *Centro de Estudos Florestais* under FEDER/POCI.

References

[Alagador et al.(2012)] D. Alagador, M. Triviño, J. O. Cerdeira, R. Brás, M. Cabeza, and M. B. Araújo. Linking like with like: optimizing connectivity between environmentally-similar habitats. *Landscape Ecology*, 27(2):291–301, 2012.

[Brooks et al.(2002)] T.M. Brooks, R.A. Mittermeier, C.G. Mittermeier, A.B. Rylands G.A.B. da Fonseca and, W.R. Konstant, P. Flick, J. Pilgrim, S. Oldfield, G. Magin, and C. Hilton-Taylor. Habitat loss and extinction in the hotspots of biodiversity. *Conservation Biology*, 16:909–923, 2002.

[Brás et al.(2012)] R. Brás, J. O. Cerdeira, D. Alagador, and M. B. Araújo. Multylink, version 2.0.2 (computer software http://purl.oclc.org/multylink), 2012.

[Brás et al.(2013)] R. Brás, J. O. Cerdeira, D. Alagador, and M. B. Araújo. Linking habitats for multiple species. *Environmental Modelling & Software*, 40:336 – 339, 2013.

[Demaine et al.(2009)] E. Demaine, M. Hajiaghayi, and P. Klein. Node-weighted steiner tree and group steiner tree in planar graphs. In S. Albers et al, editors, *Automata, Languages and Programming*, volume 5555 of *Lecture Notes in Computer Science*, pages 328–340. Springer Berlin / Heidelberg, 2009.

[Dijkstra(1959)] E. W. Dijkstra. A note on two problems in connexion with graphs. *Numerische Mathematik*, 1:269 – 271, 1959.

[Dreyfus and Wagner(1971)] S. E. Dreyfus and R. A. Wagner. The Steiner problem in graphs. *Networks*, 1(3):195–207, 1971.

[Duin and Volgenant(1987)] C. W. Duin and A. Volgenant. Some generalizations of the Steiner problem in graphs. *Networks*, 17(3):353–364, 1987.

[Feo and Resende(1995)] T. A. Feo and M. G. C. Resende. Greedy randomized adaptive search procedures. *Journal of Global Optimization*, 6(2):109 – 133, 1995.

[Hanski(2005)] I. Hanski. *The shrinking world: Ecological consequences of habitat loss.* Excellence in Ecology, 14. International Ecology Institute, Oldendorf/Luhe, Germany, 2005.

[Klein and Ravi(1995)] P. Klein and R. Ravi. A nearly best-possible approximation algorithm for node-weighted Steiner trees. *Journal of Algorithms*, 19(1):104–115, 1995.

[Kou et al.(1981)] L. Kou, G. Markowsky, and L. Berman. A fast algorithm for Steiner trees. *Acta Informatica*, 15:141–145, 1981.

[Lai et al.(2011)] K. J. Lai, C. P. Gomes, M. K. Schwartz, K. S. McKelvey, D. E. Calkin, and C. A. Montgomery. The Steiner multigraph problem: Wildlife corridor design for multiple species. In *Association for the Advancement of Artificial Intelligence*, pages 1357–1364, 2011.

[Magnanti and Raghavan(2005)] T. L. Magnanti and S. Raghavan. Strong formulations for network design problems with connectivity requirements. *Networks*, 45(2):61–79, 2005.

[Merriam(1984)] G. Merriam. Connectivity: a fundamental ecological characteristic of landscape pattern. In J. Brandt and P. Agger, editors, *Proceedings of the 1st international seminar on methodology in landscape ecological research and planning*, pages 5–15. Roskilde Univ, Denmark, 1984.

[Rayward-Smith(1983)] V. J. Rayward-Smith. The computation of nearly minimal Steiner trees in graphs. *International Journal of Mathematical Education in Science and Technology*, 14(1):15–23, 1983.

[Rayward-Smith and Clare(1986)] V. J. Rayward-Smith and A. Clare. On finding Steiner vertices. *Networks*, 16(3):283–294, 1986.

[Segev(1987)] A. Segev. The node-weighted steiner tree problem. *Networks*, 17:1–17, February 1987.

[Siek et al.(2002)] J. G. Siek, L. Lee, and A. Lumsdaine. *The boost graph library: user guide and reference manual.* Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2002.

[Winter(1987)] P. Winter. Steiner problem in networks: a survey. *Networks*, 17:129 – 167, 1987.

[Wong(1984)] R. Wong. A dual ascent approach for Steiner tree problems on a directed graph. *Mathematical Programming*, 28:271–287, 1984.

A multi-objective and multi-period approach for planning the delivery of long-term care services

Teresa Cardoso[a] *, Mónica Oliveira[a], Ana Barbosa-Póvoa[a], Stefan Nickel[b]

[a] *Centre for Management Studies of Instituto Superior Técnico, Technical University of Lisbon, Lisbon, Portugal*

[b] *Institute of Operations Research, Karlsruhe Institute of Technology, Karlsruhe, Germany*

* *E-mail:*`teresacardoso@ist.utl.pt`

Abstract

European countries are currently facing an increasing demand for Long-Term Care (LTC). Satisfying this increasing demand requires an adequate supply of services, which is still low in many countries. Accordingly, LTC planning ranks highly on the health policy agenda of many European countries. This study develops a multi-objective and multi-period mathematical programming model to inform on how to organize institutional LTC provision when three equity objectives are pursued - access, socioeconomic equity and geographical equity. The multi-objective function is structured and uses weights built with the MACBETH approach. The applicability of the model is illustrated through a case study in Portugal.

Keywords: LTC planning; equity; multi-objective; multi-period; MACBETH; optimization

1 Introduction

Long-term care (LTC) includes both nonmedical and medical services that are delivered to individuals who have lost some capacity for self-care due to chronic illness and/or disability [Knapp and Somani, 2009]. The organization and provision of these services varies across countries [European Commission, 2008]. In particular, there are different levels of services organization (with services ranging from short-term institutional services to longer institutionalizations) and different divisions of responsibility in the provision of care (shared between the family and the public and private sectors).

Across Europe, the LTC sector has been facing several challenges. First, it has been widely recognized that the population need for LTC is increasing at a global scale, mainly due to the ageing phenomenon and to the increase in the prevalence of chronic diseases [World Health Organization, 2000]. Satisfying this increasing demand requires an adequate supply of services, which is still low in many countries [World Health Organization, 2000]. In this context, planning the delivery of LTC represents a health policy priority in many European countries. Particularly, in countries based on a National Health Service (NHS) structure, planning of LTC services must consider the budget limitations that are currently in place (since many European countries are currently facing severe budget cuts), as well as the promotion of equity in the provision of services, with several equity objectives being pursued [Baker, 2000].

Different methodologies have been used for tackling these planning problems in the health care sector, with special emphasis on mathematical programming models [Brailsford and Vissers, 2011]. But the developed approaches have not yet comprehensively addressed important aspects often present in real planning problems [Stummer et al., 2004]. This is the case of modeling the multi-objective nature of the problem in hand, a feature that has only been explored in a few studies [Rahman and Smith, 2000]. Also, although equity represents one of the objectives most refereed in the health policy literature [Powell and Exworthy, 2003], no multi-objective model has been proposed to the joint attainment of different equity objectives in this sector. Furthermore, the consideration of a set of time periods in which the provision of services may change is also relevant for planning, and this aspect has not yet been often explored in strategic and tactical health care planning studies. Additionally and when it comes to the case of planning a LTC network, few studies have considered the above identified model features (as it can be seen from the few existing studies in the area [Shroff et al., 1998, Kim and Kim, 2010, Lin et al., 2012]).

The present study proposes a multi-objective and multi-period mathematical programming model to support the planning of a LTC network in the medium-term (in terms of capacity planning and location selection) in the context of NHS-based countries, when different equity related objectives are pursued -

access, socioeconomic equity and geographical equity [Marsh and Schilling, 1994]. Different weights are assigned to each equity objective, and a weighting procedure supported by the *Measuring Attractiveness by a Category-Based Evaluation Technique* (MACBETH) [Bana e Costa et al., 2012] is proposed to assist the process of building weights with the decision-maker (DM) and to explain how those weights should be interpreted. The applicability of the model is illustrated through the analysis of a case study in Portugal.

In this paper we start by presenting a brief review on the methods used to plan the delivery of health care services. Afterwards, the proposed methodology is described, and then the case study under analysis is presented. Some conclusions and lines for further research are included in the final section.

2 Literature review

Key issues commonly considered when planning the delivery of health care services include capacity planning and location selection [Rais and Viana, 2010]. Both issues have been addressed in the literature during the last decade [Santibáñez et al., 2009, Mestre et al., 2011, Smith et al., 2012], with mathematical programming models being commonly used in this area [Brailsford and Vissers, 2011]. Several aspects should be addressed when developing this type of models in the health care sector, such as i) analyzing the effect of multiple objectives, ii) taking equity as one of the fundamental objectives; and iii) considering a planning horizon divided into a set of time periods in which the provision of care might change. These aspects are relevant for planning health care services in general, and LTC services in particular.

Most mathematical programming models applied for health care planning make use of a single objective [Rahman and Smith, 2000]. Yet, planning the delivery of health care services typically depends on several objectives that often stand in conflict with each other [Stummer et al., 2004]. Accordingly, there seems to be an increasing interest in developing multi-objective approaches in recent years. Within this research line, Stummer et al. [2004] proposed a multi-objective approach to support decisions on the location and size of medical departments within a network of hospitals, using as objectives the minimization of travel time, costs, number of patients rejected and number of unit moves required to restructure the current allocation; and Abdelaziz and Masmoudi [2012] proposed a capacity planning mathematical programming model that minimizes the cost of creating and managing new beds and the number of nurses and physicians working in the hospitals.

Within the objectives most widely used in health policy and planning, equity plays a key role [Powell and Exworthy, 2003]. In particular, geographical and socioeconomic equity represent the two key areas of equity research for NHS based systems [Marsh and Schilling, 1994]. Various alternative definitions have been proposed in the literature for both geographical and socioeconomic equity - e.g., Oliveira and Bevan [2006] used two different definitions of geographical equity while redistributing hospital supply, namely, equity of utilization and equity of geographic access; Mestre et al. [2011] used a specific definition of geographical equity, by minimizing the distance traveled by patients and imposing that no patient should take more than a maximum amount of time to access hospital services; and Drezner and Drezner [2011] noted the relevance of using a socioeconomic equity objective so as to ensure that customers with different levels of wealth are treated equitably.

Different approaches have been used to deal with multiple objectives in mathematical programming models in the health care sector, with most of them aggregating different objectives into an overall objective function through the use of a set of weights [Santibáñez et al., 2009, Smith et al., 2012]. The weights used in these studies are usually arbitrarily chosen, there being no discussion on the meaning of those weights and on which preference information from DMs is required to build them.

Most mathematical programming models used for strategic and tactical planning in health do not explicitly model a planning horizon, being an exception the works developed by Santibáñez et al. [2009] and Ghaderi and Jabalameli [2013]. Also, few studies exist in the area of LTC, and with the exception of Shroff et al. [1998], they consider a single objective [Kim and Kim, 2010, Lin et al., 2012].

In summary, although existing studies already addressed several model features that are considered to be relevant when developing methods for planning the delivery of health care services in general, and LTC services in particular, no study exists proposing a mathematical programming model that comprehensively considers all these features. This study aims at filling this gap by developing a multi-objective and multi-period mathematical programming model for planning the delivery of LTC services that accounts for three different equity objectives (equity of access, socioeconomic equity and geographical equity). Considering this joint effect is relevant because different definitions of equity might clash [Oliveira, 2003], and up to our knowledge, no multi-objective model has considered this effect. The model objective function is structured and uses weights built with the MACBETH approach, which assists in explaining on how to build the weights for each objective and to clarify how one should interpret those weights.

3 Methodology

3.1 Planning context

This study aims at developing methods to inform on how to reorganize the LTC network in the medium-term (i.e., at a tactical level), so as to provide information about i) where and when to locate institutional LTC services and with which bed capacity, ii) how to distribute this capacity across services and patient groups, iii) which changes are needed in this network over time (e.g., capacity increasing or reduction; and closure and opening of services), and iv) how these changes impact on the achievement of the three equity objectives and on costs. Four basic types of institutional LTC services are considered in the model: convalescence care (CC); medium-term and rehabilitation care (MTRC); long-term and maintenance care (LTMC); and palliative care (PC) services. These are the four types of institutional LTC services currently provided in Portugal, as well as in other countries [European Commission, 2008]. Three key equity objectives to be accounted for when reorganizing a network of LTC services are: equity of access (EA), socioeconomic equity (SE) and geographical equity (GE). Considering these multiple objectives requires the definition of weights for each objective, with the choice and interpretation of those weights being assisted by the MACBETH approach.

3.2 Modeling equity objectives

3.2.1 Structuring equity objectives

The three equity related objectives considered in the present work are operationalized in the objective function through three different equity measures, denoted by f_1, f_2 and f_3:

1. Min f_1: Minimization of the total travelling time for individuals accessing institutional LTC services (Eqs. (1-4)), with t^{tot}, t^{pen} and t^{max} representing the total travelling time, the penalty attributed to individuals in need not receiving institutional care and the maximum total travelling time, respectively, in minutes. These variables depend on the travel time between demand point d and service location l (t_{dl}), on the maximum travelling time allowed for patients accessing institutional LTC services (mT) and on the number of individuals from demand point d and socioeconomic group g receiving institutional care s in location l at t (R_{dgslt}) and requiring institutional service s at t (D_{dgst}). This objective's structure is in line with the equity objective proposed by Mestre et al. [2011]. A penalty t^{pen} (Eq. (3)) is modeled in this objective so as to ensure that services will be provided for as many individuals in need as possible during the entire planning horizon. The penalty assumes that if individuals do not receive the care they need, in terms of access it is similar as incurring the maximum travel time (note that alternative assumptions could be chosen by the DM). As a result, this objective encapsulates two equity dimensions within a single objective - minimizing total travel time and motivating the provision of care for as many individuals in need as possible;

$$Min f_1 = \frac{t^{tot} + t^{pen}}{t^{max}} \tag{1}$$

$$t^{tot} = \sum_{d\epsilon D}\sum_{g\epsilon G}\sum_{s\epsilon S}\sum_{l\epsilon L}\sum_{t\epsilon T} t_{dl} R_{dgslt} \tag{2}$$

$$t^{pen} = \sum_{d\epsilon D}\sum_{g\epsilon G}\sum_{s\epsilon S}\sum_{t\epsilon T} mT(D_{dgst} - \sum_{l\epsilon L} R_{dgslt}) \tag{3}$$

$$t^{max} = \sum_{d\epsilon D}\sum_{g\epsilon G}\sum_{s\epsilon S}\sum_{t\epsilon T} D_{dgst} mT \tag{4}$$

2. Min f_2: Minimization of unmet need for the lowest income group (Eqs. (5-6)), with pR and pD representing the number of individuals with low income receiving and requiring institutional care, respectively. This objective ensures the maximum provision of care for those with the lower capacity to pay for it, thus avoiding situations of poverty or financial dependency;

$$Min f_2 = 1 - \frac{pR}{pD} \tag{5}$$

$$pR = \sum_{d\epsilon D}\sum_{s\epsilon S}\sum_{l\epsilon L}\sum_{t\epsilon T} R_{d(g=1)slt} \tag{6}$$

3. Min f_3: Minimization of unmet need in the geographical area with the highest level of unmet need (Eqs. (7-8)), with rR_d and rD_d representing the number of individuals from demand point d receiving and requiring institutional LTC, respectively. This objective ensures the maximum provision of care in the worst off areas, i.e., in the area with the lowest (relative) provision. The structure of this objective is similar to the minimax objective presented by Eiselt and Laporte [1995], although the distance is replaced by the proportion of unmet need.

$$Minf_3 \geq 1 - \frac{rR_d}{rD_d} \qquad \forall d \epsilon D \tag{7}$$

$$rR_d = \sum_{g \epsilon G} \sum_{s \epsilon S} \sum_{l \epsilon L} \sum_{t \epsilon T} R_{dgslt} \qquad \forall d \epsilon D \tag{8}$$

When analyzing a network of LTC, it is relevant to consider the *status quo* (SQ) in the provision of care (corresponding to the current level of achievement in each equity objective), as well as the medium-term targets set by DMs as the level of achievement desired for each equity objective. These levels are shown in Fig. 1, where the SQ is denoted by L_1^0, L_2^0 and L_3^0 and the target is denoted by T_1, T_2 and T_3 for each objective. These levels are allowed to take values between the maximum and minimum attainable levels for each objective, corresponding to the values of 1 and 0, respectively. The substantive meaning of these levels is shown in Fig. 1. The arrows shown in this figure indicate the desirable direction of change for each equity objective (from the SQ to the target in the medium term).

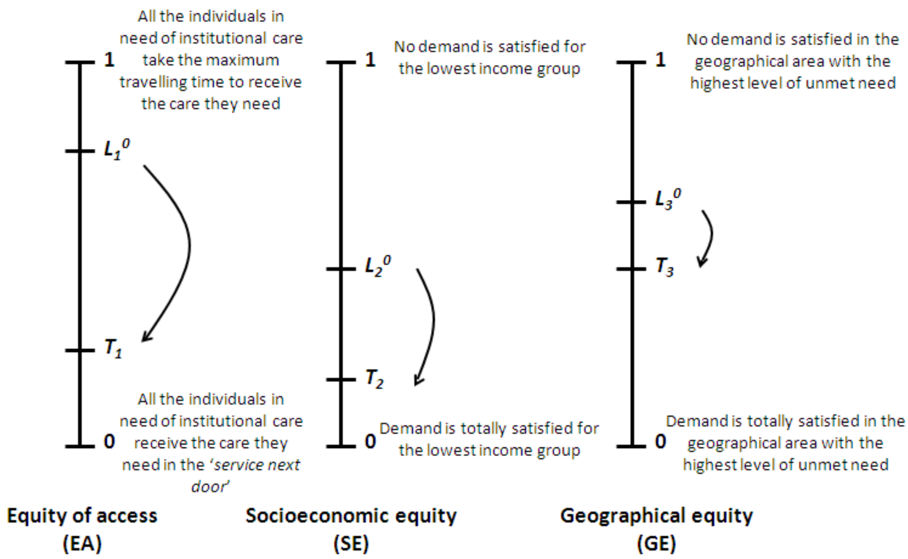


Figure 1: Levels defined for each equity objective

3.2.2 Evaluating the space for equity improvements

To deal with the before mentioned multiple objectives we consider an objective function that can be written as the weighted sum of the three equity objectives (Eq. (9)), in which W_1, W_2 and W_3 represent the weights assigned to each objective. These weights are scaling constants that allow the additive aggregation of the different objectives.

$$Min(W_1 f_1 + W_2 f_2 + W_3 f_3) \tag{9}$$

This study considers that these weights should be defined by DMs, thus reflecting their system of values. One should note that while many multi-objective approaches rely on the use of weights, it is not common to define how these weights can be obtained and what their substantive meaning is. For determining these weights one should not rely on the intuitive notion of importance of each objective, as this is the most common critical mistake in the decision analysis literature [Keeney, 1992]. In view of that, proper weighting procedures should be used, with weights being determined with reference to objective impact scales [Bana e Costa et al., 2012]. Accordingly, so as to properly build weights with the DM and to correctly explain their substantive meaning, we use the MACBETH approach (being assisted by the M-MACBETH decision support system [Bana e Costa et al., 2003]). MACBETH has been widely used in different contexts, showing that it can be used for the proper building of weights with a DM.

The first step is to define a descriptor of performance, which we name here as an equity measure that operationalizes each equity objective. This step is performed in section 3.2.1, where three equity measures are defined and four meaningful reference levels are made explicit for each measure: the SQ and the medium-term target; and the maximum and minimum attainable levels.

Once the impact scales and reference points to be used within MACBETH are defined, a MACBETH protocol of questioning is followed to build the weights - note that the references SQ and the targets are used for building weights in this protocol, avoiding that unrealistic questions are done to the DM when the extreme reference points are used in the weighting procedure. Within this context, the DM is first asked to rank by decreasing order of relative importance the swings from the SQ to the target level on each equity objective (while maintaining the other objectives in the SQ), so as to order the objectives in the MACBETH weighting matrix (see the matrix in Fig. 2b). Then the DM is asked to qualitatively judge the importance of swinging from the SQ to the target in each equity objective (while maintaining all the remaining objectives in the SQ) using one of the seven MACBETH semantic categories (null, very weak, weak, moderate, strong, very strong and extreme). These judgments are needed to complete the last column of the matrix. Afterwards, the DM is asked to qualitatively judge the difference of importance between each swing to the target and the swing to the target of the most important objective (which is the EA objective, as shown in Fig. 2). These judgments are needed to fill the first row of the weighting matrix. To complete the remaining cells of the matrix, the DM is then asked to judge the remaining pairs of swings.

After completing the MACBETH weighting matrix, the M-MACBETH decision support system suggests a first cardinal weighting scale that respects the qualitative judgments introduced in the matrix. This first set of weights should then be presented to the DM and adjusted until the DM validates that set of weights. Fig. 2c shows an example of a weighting scale (in percentage) obtained after following the previous steps ($W_1^* = 0.48$, $W_2^* = 0.36$ and $W_3^* = 0.16$).

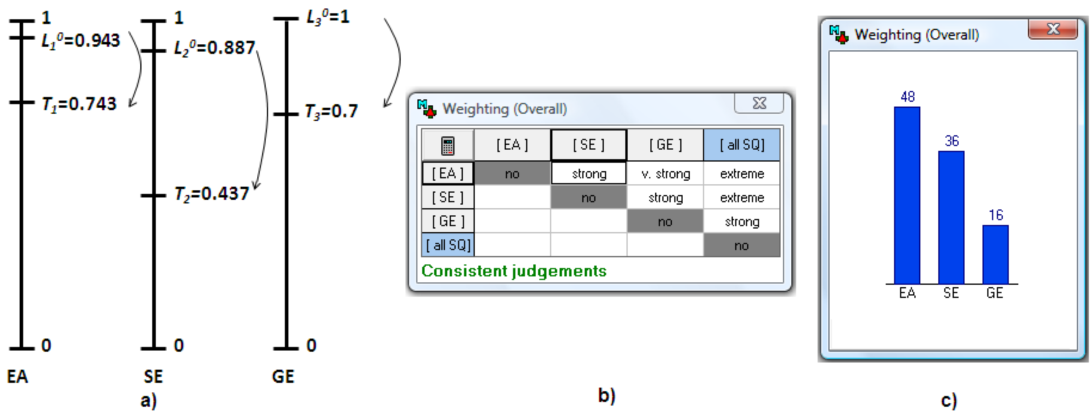


Figure 2: a) SQ and the target reference levels; b) judgments displayed in the MACBETH weighting matrix; c) weighting scale (in percentage) validated with the DM

Nevertheless, since these weights are obtained using as reference levels the SQ and the target levels, one needs to rescale these weights so as to allow its use in the multi-objective function from Eq. (9) - these new weights should have as reference levels the maximum and minimum attainable levels (i.e., 1 and 0 for each equity objective). The weighting scale shown in Fig. 2c is thus rescaled to $W_1 = 0.66$, $W_2 = 0.21$ and $W_3 = 0.13$.

3.3 Mathematical formulation: defining the constraints

The proposed mathematical programming model makes use of a set of constraints that, due to the space limitations, are only described in a summarized way:

- Resource requirement constraints - constraints used to calculate the number of beds that should be made available for institutional LTC provision, and to establish a balance between existing beds and additional beds in which one needs to invest over time;

- Reallocation constraints - constraints used to define the maximum number of beds allowed to be reallocated from and to a given LTC service, as well as to ensure that the beds should be preferentially kept in their original service whenever a lack of resources exists in that service;
- Minimum and maximum capacity constraints - constraints used to impose the minimum and maximum number of beds allowed per service;
- Opening and closure of services constraints - these constraints impose that opening/closing a service is not allowed after deciding upon closing/opening it in a previous time period;
- Patients' assignment constraints - constraints stating that individuals will only receive institutional care in locations where the required service is available;
- Single and closest assignment constraints - constraints used to ensure that individuals will receive the care they need in a single and closest service location, as well as that individuals cannot receive LTC in locations that are not within a maximum travel time;
- Budget limitation constraints - constraints used to guarantee that operations and investments in the LTC network are associated with total costs that are within the available operational and investment budgets.

4 Case study

The developed model was implemented in the General Algebraic Modeling System (GAMS) 23.7 using CPLEX 12.0 on a Two Intel Xeon X5680, 3.33GHz computer with 12GB RAM. The applicability of the model is illustrated through the solution of a case study in the Great Lisbon region. The planning period considered goes from 2012 to 2016 and two different planning contexts (hereafter called Case A and Case B) with different questions are explored, as described in detail in Table 1.

Table 1: Planning questions under study

Cases	Planning questions	Model parameters/constraints
A	To what extent can the current provision of institutional LTC be improved using the available budget? And which are the consequences of these improvements in each equity measure?	The budget is limited; and the weights attributed to each equity objective are W_1=0.66, W_2=0.21 and W_3=0.13.
B	How much would it cost to improve the provision of institutional LTC so as to ensure the achievement of pre-defined targets for each equity objective?	The budget is unlimited; the travelling time measure should be at most 0.743 ($f_1 \leq 0.743$); and the unmet need for the low income group and in the geographical area with the highest level of unmet need should be at most 0.437 ($f_2 \leq 0.437$)and 0.7 ($f_3 \leq 0.7$), respectively.

One should note that the weights shown in table 1 for Case A are illustrative weights obtained by following the protocol described in Section 3 (and after rescaling the weights shown in Fig. 2c) and do not reflect the views of individuals working on the LTC sector.

A wide range of data was used, including information on the i) demand for institutional LTC in each county in the Great Lisbon during the period considered, ii) current supply of institutional LTC, iii) average length of stay (LOS) for each type of institutional LTC service, iv) travel times between each county and each institutional LTC service, and v) operational and investment costs and budgets. Demand estimates for institutional LTC were predicted using the simulation model developed for that purpose by Cardoso et al. [2012].

4.1 Case A

As shown in Table 1, this case answers to the question: which changes to the current network of LTC should be carried out so as to improve equity while considering the available budget? Key results from Case A are shown in Table 2 and Fig. 3. These provide information regarding: i) the evolution of bed capacity for institutional LTC provision, for the different services considered, in the Great Lisbon region (Table 2 and Fig. 3), ii) where and in which years to open or close services (Fig. 3), and iii) which services should be used by individuals in each county (Fig. 3). For simplification purposes, institutions in each county are numbered - e.g., there are two institutions in Lisbon, and these are named Lisbon (1) and Lisbon (2).

Table 2: Evolution of total bed capacity for institutional LTC provision in the Great Lisbon region under Case A

LTC services	2011	2012	2013	2014	2015	2016
CC	136	415	493	519	545	567
MTRC	28	5	73	283	316	287
LTMC	60	0	0	0	0	0
PC	61	103	203	213	220	227

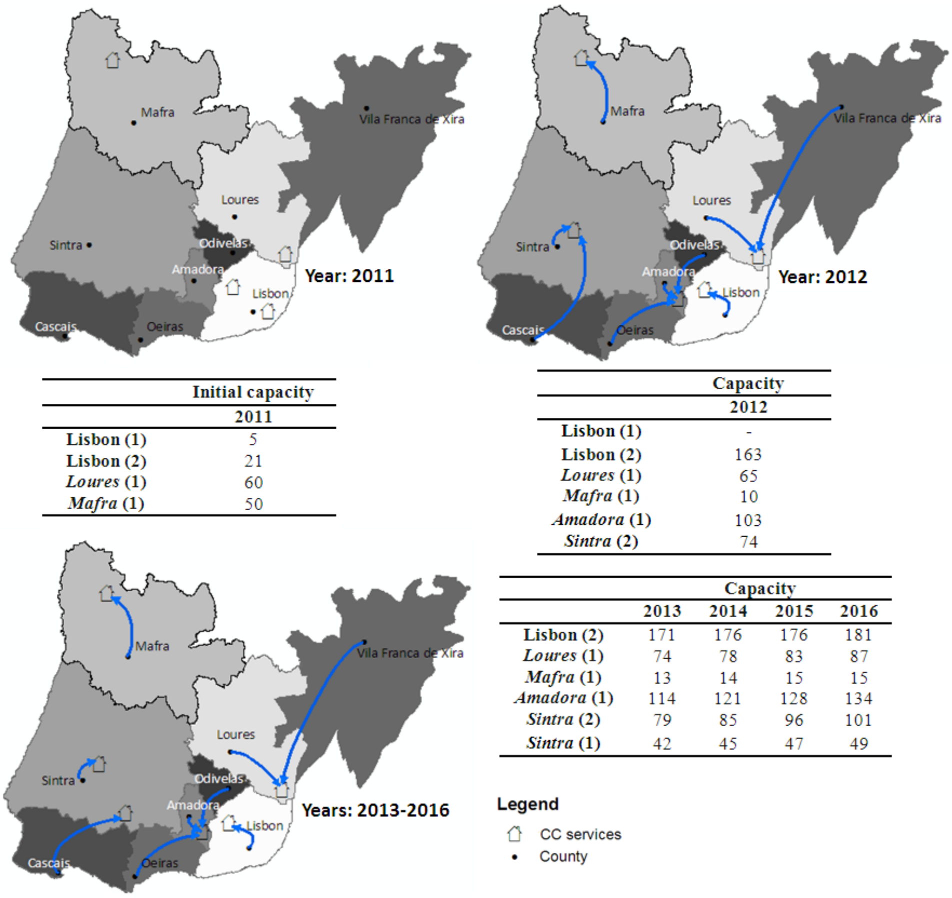


Figure 3: Changes in the provision of CC services within the LTC network over the 2012-2016 period under Case A (in brackets is the number of institutions providing care in that county)

According to Table 2, no LTMC service is opened during the period in analysis and the highest level of provision is observed for CC services, with the total provision of beds increasing from 136 (in 2011) to 567. Furthermore, Fig. 3 summarizes the results obtained for the future provision of CC (institutional service with the highest capacity requirements throughout the period in analysis). For this service it can be seen that: i) in 2012 one of the CC services currently provided in Lisbon should be closed; ii) also in 2012, *Amadora* (1) and *Sintra* (2) should expand their offer to CC provision; and iii) from 2013 onwards, *Sintra* (1) should ensure the provision of CC to individuals from *Cascais*.

The reported changes are associated with improvements in each equity measure as depicted in Fig. 4. This figure depicts three levels for each objective: the SQ, the target and the equity level obtained under Case A (denoted by E_{A1}, E_{A2} e E_{A3}). Taken as example the EA, it can be seen that this is improved from 0.943 (the SQ) to 0.771. This corresponds to an improvement of access from 17.5 to 9.5 minutes.

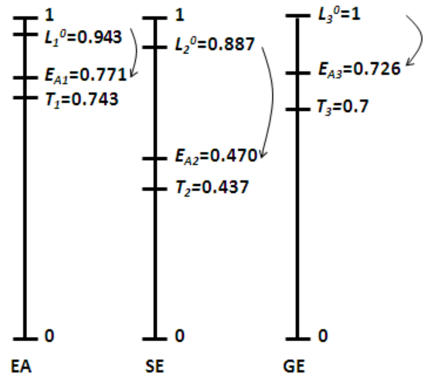


Figure 4: Equity improvements obtained when applying the model under Case A

4.2 Case B

This case answers to the question: how much would it cost to improve the provision of institutional LTC so as to ensure the achievement of all equity targets? Accordingly, the model is run imposing the targets defined in Table 1 and the results obtained are shown in Fig. 5. In this case, the objective function described in Eq. (9) is replaced by the minimization of total costs over the entire planning horizon.

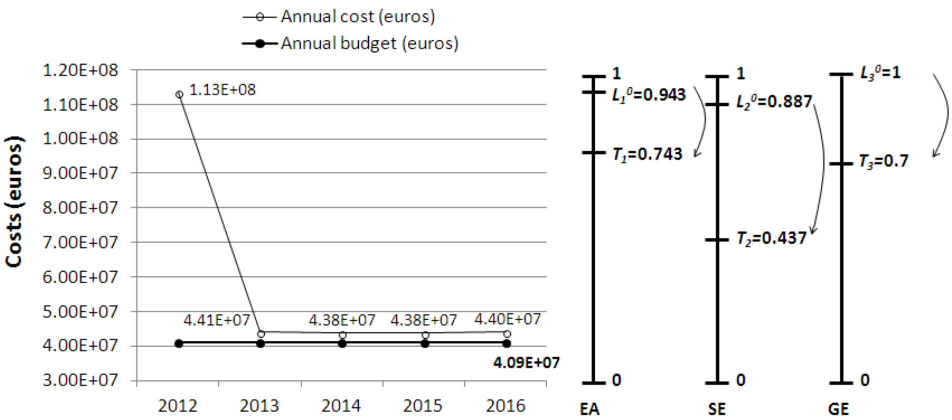


Figure 5: Total costs associated with achieving the targets for all the equity objectives when no budget constraints are applied

Results depicted in Fig. 5 show that the available budget is insufficient to achieve the targets established. Reaching these targets in the entire planning period costs around 290 million euros, which is 1.5 times higher than the total budget available for that period. Results also show that higher investments take place in 2012, and these are essentially related to the beds capacity expansion. This behavior clearly indicates that nowadays a poor provision of institutional LTC exists in the Great Lisbon region.

Table 3 shows the computational results obtained when applying the model for cases A and B. Given the tactical nature of the planning decisions under consideration, these results show that the model is solved in an acceptable time frame for all the cases.

Table 3: Computational results

Cases	Total variables	Discrete variables	Total constraints	Iterations	CPU (min.)	Gap (%)	Objective
A	14,199	11,655	23,037	914,576	5	0.00	0.7018
B	14,199	11,655	23,030	7,908,472	60	0.04	288,998,852 euros

5 Conclusions

Planning in the LTC sector is nowadays a policy priority in many European countries, mainly due to the current ageing phenomenon and to the increasing prevalence of chronic diseases. Moreover, the inadequate supply of LTC services along with the severe budget cuts that many European countries are currently facing also have been increasing the need for a proper planning of services.

Within this context, a multi-objective and multi-period mathematical programming model is proposed to support the planning of LTC networks in the context of NHS-based countries. The model accounts for three different equity-related policy objectives pursued by DMs, namely, equity of access, socioeconomic equity and geographical equity. These equity objectives are operationalized through three measures and the respective weights are built and interpreted with the help of the MACBETH approach. The model was developed in a generic way and can be easily adapted to inform planning decisions in the LTC sector under different contexts - e.g., when there are budget constraints or when the DM needs to compute the total cost that ensures the achievement of all the equity targets.

As main results, the model provides essential information for planning services in the medium-term, namely, on i) where and when to locate different types of institutional LTC services and with which bed capacity, ii) how to distribute this capacity across services and patient groups, iii) which changes are needed over time, and iv) which impact these changes have on equity objectives and on costs.

When applying the model to the Great Lisbon region it was shown that the budget currently available for increasing the provision of institutional LTC in that region is far from the one required to meet the demand for care.

Although the model developed is generic, further work on it should be pursued. In particular, the developed model should be extended to consider the entire range of LTC services (institutional, home-base and ambulatory care services). Also, given the uncertain nature of LTC demand [Cardoso et al., 2012], it is relevant to explore the presence of uncertainty in the LTC demand.

Acknowledgements

The first author acknowledges financing from the *Fundação para a Ciência e Tecnologia* (Portugal) (SFRH/BD/63966/2009) and thanks the support from the Centre for Management Studies of *Instituto Superior Técnico* (CEG-IST, Portugal). The authors acknowledge financing from *Conselho de Reitores das Universidades Portuguesas* (CRUP) and *Deutscher Akademischer Austauschdienst* (DAAD), and are grateful to the team from the Institute of Operations Research, Karlsruhe Institute of Technology (KIT, Germany).

References

[Knapp and Somani, 2009] Knapp, M. and Somani, A. (2009). Health financing: Long Term Care, Organization and Financing. In: Carrin, G., Buse, K., Heggenhougen, K. and Quah, S.R. (eds) *Health systems policy, finance, and organization.*, 1st edn. Academic Press, 250-259.

[European Comission, 2008] European Commission (2008). Long-Term Care in the European Union. Brussels: Directorate-General for Employment, Social Affairs and Equal Opportunities.

[World Health Organization, 2000] World Health Organization (2000). Home-based long-term care: report of a WHO study group. Technical Report Series, vol 898. World Health Organization, Geneva.

[Baker, 2000] Baker, M. (2000). Making sense of the NHS White Papers. Second Edition, Radcliffe Medical Press, Oxon.

[Brailsford and Vissers, 2011] Brailsford, S. and Vissers, J. (2011). OR in healthcare: A European perspective. *European Journal of Operational Research*, 212(2):223-234.

[Stummer et al., 2004] Stummer, C., Doerner, K., Focke, A. and Heidenberger, K. (2004). Determining location and size of medical departments in a hospital network: A multiobjective decision support approach. *Health Care Management Science*, 7(1):63-71.

[Rahman and Smith, 2000] Rahman, S. U. and Smith, D. K. (2000). Use of location-allocation models in health service development planning in developing nations. *European Journal of Operational Research*, 123(3):437-452.

[Powell and Exworthy, 2003] Powell, M. and Exworthy, M. (2003). Equal access to health care and the British National Health Service. *Policy Studies*, 24(1):51-64.

[Shroff et al., 1998] Shroff, H. F. E., Gulledge, T. R., Haynes, K. E. and O'Neill, M. K. (1998). Siting efficiency of long-term health care facilities. *Socio-Economic Planning Sciences*, 32(1):25-43.

[Kim and Kim, 2010] Kim, D. G. and Kim, Y. D. (2010). A branch and bound algorithm for determining locations of long-term care facilities. *European Journal of Operational Research*, 206(1):168-177.

[Lin et al., 2012] Lin, F., Kong, N. and Lawley, M. (2012). Capacity Planning for Publicly Funded Community Based Long-Term Care Services. In: Johnson, M. P. (ed). *Community-Based Operations Research*, vol 167. *International Series in Operations Research & Management Science*, Springer, New York, 297-315

[Marsh and Schilling, 1994] Marsh, M. T. and Schilling, D. A. (1994). Equity measurement in facility location analysis: a review and framework. *European Journal of Operational Research*, 74(1):1-17.

[Bana e Costa et al., 2012] Bana e Costa, C. A., De Corte, J. M. and Vansnick, J. C. (2012). MACBETH. *International Journal of Information Technology and Decision Making*, 11(2):359-387.

[Rais and Viana, 2010] Rais, A. and Viana, A. (2010). Operations Research in Healthcare: a survey. *International Transactions in Operational Research*, 18(1):1-31.

[Santibáñez et al., 2009] Santibáñez, P., Bekiou, G. and Yip, K. (2009). Fraser Health uses mathematical programming to plan its inpatient hospital network. *Interfaces*, 30(3):196-208.

[Mestre et al., 2011] Mestre, A., Oliveira, M. D. and Barbosa-Póvoa, A. (2011). Organizing Hospitals into Networks: A Hierarchical and Multiservice Model to Define Location, Supply and Referrals in Planned Hospital Systems. *OR Spectrum*, 34(2):319-348.

[Smith et al., 2012] Smith, H., Harper, P. and Potts, C. N. (2012). Bicriteria efficiency/equity hierarchical location models for public service application. *Journal of the Operational Research Society*, 64(4):500-512.

[Abdelaziz and Masmoudi, 2012] Abdelaziz, F. B. and Masmoudi, M. (2012). A multiobjective stochastic program for hospital bed planning. *Journal of the Operational Research Society*, 63(4):530-538.

[Oliveira and Bevan, 2006] Oliveira, M. D. and Bevan, G. (2006). Modelling the redistribution of hospital supply to achieve equity taking account of patient's behaviour. *Health Care Management Science*, 9(1):19-30.

[Drezner and Drezner, 2011] Drezner, T. and Drezner, Z. (2011). A note on equity across groups in facility location. *Naval Research Logistics*, 58(7):705-711.

[Ghaderi and Jabalameli, 2013] Ghaderi, A. and Jabalameli, M. S. (2013). Modeling the budget-constrained dynamic uncapacitated facility location-network design problem and solving it via two efficient heuristics: A case study of health care. *Mathematical and Computer Modelling*, 57(3-4):382-400.

[Oliveira, 2003] Oliveira, M. D. (2003). Achieving geographic equity in the Portuguese hospital financing system. PhD Thesis. Operational Research Department, London School of Economics, University of London, London.

[Eiselt and Laporte, 1995] Eiselt, H. A. and Laporte, G. (1995). Objectives in location problems. In: Drezner, Z. (ed). *Facility Location: A Survey of Applications and Methods*, Springer, New York, 151-180

[Keeney, 1992] Keeney, R. L. (1992). Value-Focused Thinking: A Path to Creative Decisionmaking. Harvard University Press, Cambridge.

[Bana e Costa et al., 2003] Bana e Costa, C. A., De Corte, J. M. and Vansnick, J. C. (2003). MACBETH. Working Paper LSE OR 03.56. London: London School of Economics of Political Science.

[Cardoso et al., 2012] Cardoso, T., Oliveira, M., Barbosa-Póvoa, A. and Nickel, S. (2012). Modeling the demand for long-term care services under uncertain information. *Health Care Management Science*, 15(4):385-412.

Design and planning of resilient closed-loop supply chains

Sónia R. Cardoso*, Ana Paula F. D. Barbosa-Póvoa*, Susana Relvas*

* *CEG-IST, Instituto Superior Técnico, Universidade Técnica de Lisboa, Av. Rovisco Pais, 1049-001 Lisboa, Portugal*

*E-mail:*sonia.cardoso@ist.utl.pt

apovoa@ist.utl.pt

susanaicr@ist.utl.pt

Abstract

In this paper, we study the design and planning of resilient closed-loop supply chains under products' demand uncertainty. This study relies on the development of a mixed integer linear programming (MILP) model that maximizes both the expected net present value (ENPV) and resilience, being the latter measured through costumer service level. The resulting bi-objective problem is solved through the augmented epsilon constraint method, which allows the generation of an approximation to the Pareto optimal curve. Network structures with different levels of logistics flexibility are studied and compared when a disruption affects them. The model applicability is shown through the solution of a European supply chain case.

Keywords: Supply chain management, Resilience, Design, Planning, Uncertainty, MILP.

1 Introduction

Traditionally, supply chains were seen as a direct flow of products and information from suppliers to end consumers. Nowadays, this concept has changed essentially due to the economic crisis, growing environmental concerns and the consequent need of reducing the consumption of natural resources and energy. In this way, supply chains are moving to the establishment of circular structures the so called closed-loop supply chains (CLSCs) (Guide et al., 2002, Salema et al., 2010, Özkir and Basligil, 2012). Furthermore, organizations are currently more vulnerable to risks due to their increasing geographical coverage and use of lean processes (Pfohl et al., 2010). According to Tang (2006) there are two types of risks within a supply chain: operational risks and disruption risks. Operational risks are related to inherent uncertainties such as in the supply, delivery lead times and demand. Disruption risks have a low probability of happening but can cause a significant business impact (Sawik, 2013), such as earthquakes, fires, equipment breakdowns, labour strikes and terrorist attacks. Thus the need to study the stochastic nature of CLSCs is of utmost importance. Some works have been recently published in the literature addressing this issue. Amaro and Barbosa-Póvoa (2009) presented a MILP planning model considering uncertainty on product portfolios demand and prices. Amin and Zhang (2012) presented a bi-objective facility location model where the stochastic programming is used to analyse the effects of demand and returns uncertainty on the network configuration. Zeballos et al. (2012) considered the design and planning problem under uncertainty on the quantity and quality of the reverse flow of products. More recently, Cardoso et al. (2013) developed a MILP model for a generic four echelon CLSC with demand uncertainty that integrates new features such as modelling in detail different processes, the possibility of capacity expansion of existent infrastructures and processes installed, the inclusion of different types of reverse flows (non-conforming and end of life products). Furthermore, being supply chains exposed to disruptions, decision makers should incorporate the concept of supply chain resilience when designing and planning these systems. Supply chain resilience can be defined as the ability of a supply chain to return to its original state or move to a new one, more desirable state after being disturbed (Christopher and Peck, 2004). Despite the increasing interest on the area, supply chain resilience is a new and still largely unexplored area of research and few studies have attempted to measure resilience on the supply chain performance (Spiegler et al., 2012). Vugrin et al. (2011) developed a framework that includes a quantitative approach to measure resilience in terms of resilience costs that result from a disruption. Carvalho et al. (2012) used simulation to compare the supply chain response to a disruption using two design strategies: one based on flexibility and another on redundancy. Two performance measures were defined, lead time ratio and total cost. Schmitt and Singh (2012) implemented a simulation model to analyze inventory placement and back-up methodologies in a multi-echelon supply chain in order to minimize the impact of disruptions. Customer fill rate was the performance metric used.

From the above literature review it can be concluded that there is still a large space to explore the resilience concept within supply chains. In an attempt to overcome some of the identified literature limitations, in the present work a design and planning model is developed for a generic resilient closed-loop supply chain where uncertainty in products' demand is taken into account through a scenario tree approach. Resilience is measured through the costumers service level attained. Two resilience strategies are implemented and their performances in terms of expected net present value (ENPV) and resilience are compared when disruptive occurrences happen. The paper is structured as follows: in section 2, the main problem characteristics are detailed. The mathematical formulation is characterized in section 3. The case study is presented in section 4 and the results obtained are discussed. Finally, in section 5 some conclusions are presented and some directions for future developments are identified.

2 Problem Definition

The problem under study consists on determining the design and planning of CLSCs under demand uncertainty in order to maximize its economical performance and its resilience to disruptions. Economical performance is expressed in terms of ENPV, while resilience is measured through customer service level. Design decisions include number, location and capacity of each network entity, as well as of the manufacturing, assembling and dismantling processes that have to be installed in the entities. The planning decisions are related to the determination of production rates, inventory levels, forward and reverse flows amounts and transportation links.

The supply chain comprises five echelons: suppliers that provide raw-materials to the plants where production processes can be installed with the necessary capacities, warehouses which in turn are responsible for assembling and store final orders before sending them to the markets and there is also the possibility of outsourcing the production. The reverse flow has origin at the markets and is composed by two types of products: i) non-conforming products that are sent to warehouses to be repacked before being introduced in the supply chain as new products and ii) end of life products that are collected in period t and were sold in period $T_n = t - n$, being n the products' estimated lifetime. These products are sent to warehouses and then to the disassembling centers (located in the plants) where they can be disassembled and remanufactured so as to be then reintegrated in the forward supply chain. Products that are too damaged are sent to disposal, representing also a percentage of the total returned products. This flow may occur at warehouses where products' sorting is carried out. Figure 1 represents in a schematic form the described network. Each supply chain echelon may have several entities. New entities can be installed or if they already exist their capacity can be expanded.

In order to better understand how a supply chain can be resilient, two different resilient strategies are implemented over a base case. In the base case we have a CLSC (figure 1) where each plant is supplied by its own set of suppliers, no transshipment is allowed and products are sent to warehouses to final assembly and then are delivered to markets.

Then we have case B, which is the same supply chain generic structure as in case A but where multiple sourcing of raw materials is available, thus flexibility was introduced in the network. Finally, we have case C, which is an expansion of case A and is represented in figure 2. In this case, transportation links are even more flexible, meaning that plants can send products directly to the markets. It is also considered that plants and warehouses can exchange all type of products between them and multiple sourcing of raw materials is also available. In the reverse flow, markets can send the end of life products directly to disassembling centers. These strategies will be later on applied to a case-study.

3 Resilience Modelling

The design and planning of the resilient CLSC is made using as basis a model previously developed by the authors (Cardoso et al., 2013). The novelty of the present work consists on the modeling of the supply chain resilience, which is measured through the achieved customer service level. This metric was chosen since in order to be resilient a supply chain after being disrupted should recover its normal operation, meaning that it should continue to satisfy its customers. The customer service level (SC) is then calculated using equations (1) and (2).

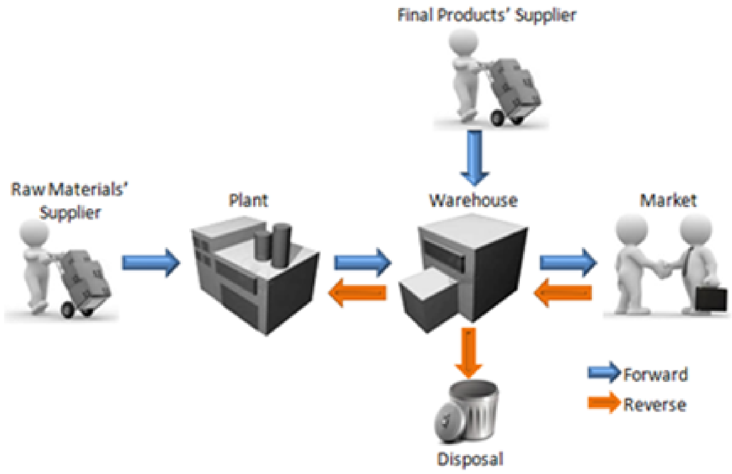


Figure 1: Generic supply chain network.

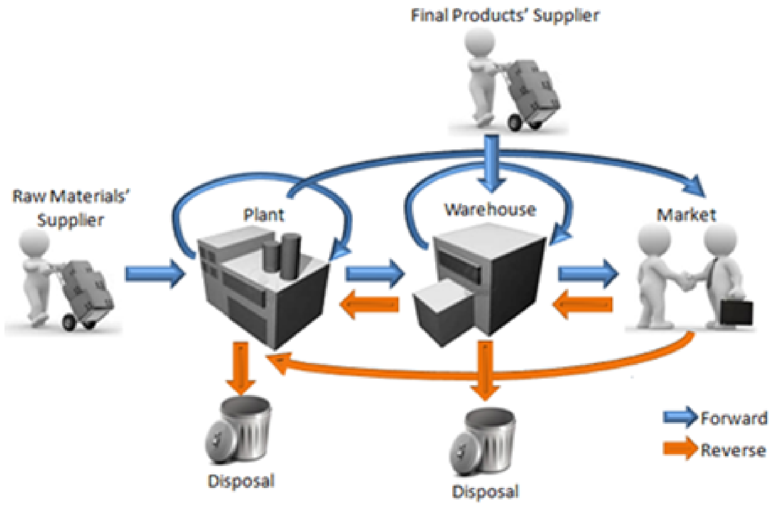


Figure 2: Case C supply chain network.

$$maxSC = \sum_t (1 - \sum_s pb_s (\frac{\sum_{v \in V_m} \sum_p ID_{pvst}}{\sum_{v \in V_m} \sum_p D_{pvst}})) \quad (1)$$

$$ID_{pvst} = D_{pvst} - \sum_{w:(w,v) \in F} Q_{wvpst}, \forall p \in G, v \in V_m, (s,t) \in S \quad (2)$$

Equation (1) represents the objective function, which aims to maximize the customer service level for all time periods. This is calculated through the summation of the service level per time period, which in turn is obtained by subtracting the division of the unsatisfied demand (ID_{pvst}) for all products and in all markets and in each scenario by all the demand in that scenario (D_{pvst}) multiplied by the probability of occurrence of each scenario (pb_s) to value 1. The unsatisfied demand of each product in each market is calculated in equation (2) as the difference between the demand of the respective product in that market and the flow of that product that was sent to that market (Q_{wvpst}).

Apart from this objective function also the maximization of the expected net present value of the supply chains is considered, since we are considering uncertainty at the costumer's demands. The net present value for each scenario is calculated as the summation of the discounted cash flows obtained in each time period, through the usage of the interest rate. The expected net present value is then obtained by the multiplication of each scenario NPV and its probability of occurrence. Using this two objective functions a bi-objective approach is then built for the design of resilient supply chains where different types of constraints are considered. In particular the model involves different sets of constraints namely: mass balance constraints, demand constraints, capacity constraints, reverse flow constraints, transportation link constraints and variables definition constraints.

4 Case Study

The developed model was applied to a European supply chain. The existent supply chain is formed by one plant in Hamburg (F1) with twelve production technologies (i1 to i12) that can be used to manufacture twelve products using fifteen raw materials and one intermediate product and six disassembling technologies (i19 to i24), one warehouse in Munich (W1) with six assembly lines (i13 to i18) and a storage capacity for 500 units. All the existing technologies have an initial capacity of 600 ton/5 years and can suffer capacities expansions or/and be installed in the new infrastructures with the necessary capacities. There are four raw materials suppliers in Frankfurt, Prague, Birmingham and Copenhagen (RS1 toRS4) and three final products suppliers located in Riga, Minsk and Warsaw (FPS1 to FPS3). The supply chain supplies eighteen markets located in different European countries.

Since the company aims to expand its business, a reconfiguration of the existent supply chain is under evaluation where different possibilities are in study. These involve the possibility of installing new plants in Bilbao (F2) and Milan (F3), where the first has raw materials suppliers located in Badajoz (RS5), Toledo (RS6), Barcelona (RS7) and Marseille (RS8) and the other has suppliers in Ljubljana (RS9), Lausanne (RS10), Linz (RS11) and Florence (RS12). New warehouses can also be installed in Portsmouth (W2), Lyon (W3), Bologna (W4) and Salamanca (W5). Such options are coupled with the ones of expanding the existing facilities. A set of three time periods each one representing 5 years is considered, resulting in a time horizon of 15 years. Figure 3 represents in a schematic way the location of all possible entities of the supply chain.

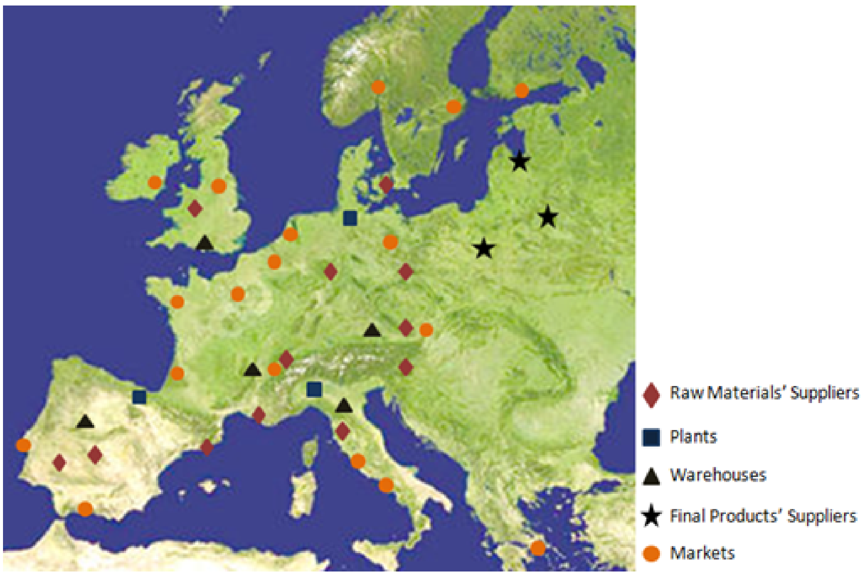


Figure 3: Supply chain network.

The capacity expansions have to be lower than 4000 ton per each entity. Due to transportation restrictions, the flow of materials has to be within 10 and 5000 ton. The warehouses turnover ratio is equal to 20 and the inventory costs are 0.30 €/ton.period. It is considered that there are no initial inventories in any infrastructure. Transportation costs are divided into a fixed and a variable part. The fixed cost is 200€for each transportation link established in the forward flow, except the link between the final products suppliers and the warehouses that costs 300€, and 150€it in the reverse flow. Regarding the variable part it is assumed a cost of 0.1€/ton.km for the forward flow and 0.2€/ton.km for the reverse.

Based on historic reasons the company has set 10% as the minimum quantity of products that has to be collected from markets. It is also assumed that 20% of the total products sold in each time period is non-conforming and 20% of the collected products are sent to disposal. When the final products production is outsourced, each ton of P29 to P34 costs, respectively, 250€, 270€, 240€, 290€, 255€and

265€.

The initial investment made in the existing technologies of the existing supply chain is 1€/ton for the plants, 0.5€/ton for warehouses and 0.1€/ton for the disassembling centers. For the warehouses (W1 to W5), the investment made in the storage capacities is respectively, 2.1€/ton, 2.15€/ton, 2.13€/ton, 1.95€/ton and 1.98€/ton.

The interest rate, salvage value and tax rate are equal to 10%, 20% and 30%, respectively. All costs have an increase of 5% in each period of time. All the remaining data is given in tables 1 to 3.

The three cases defined generically in section 3 are here applied to the supply chain under study. The economical and resilience performances of each case are measured and compared when a disruption occurs caused for instance by an earthquake and causing a 100% decrease in the production capacity of the most important plant in time period 2. This comparison is made considering two different situations: 1) there is no upper limit for the budget to invest in capacity and new infrastructures after a disruption and 2) there is no budget to invest. Demand uncertainty is also integrated in the model through a scenario tree approach. In order to build the scenario tree, it is assumed that demand is known for the first time period and then, for the second, three possible branches are considered: the optimistic that implies an increase of 10% with a probability of 0.25, the realistic with an increase of 3% and a probability of 0.5 and the pessimistic that involves a reduction of 2% in the demand with a probability of 0.25. For the last time period and for each branch of the previous period, another three possibilities are considered: an optimistic which implies an increase of 5% in the demand, the realistic with an increase of 2% and the pessimistic with a decrease of 2%.

The model developed was implemented in GAMS 23.6, and the case study problem was solved using CPLEX 12.0, in a Two Intel Xeon X5680, 3.33 GHz computer with 12 GB RAM.

Table 1: Plants operational and investment costs of production and disassembling processes, for t=1, in €/ton

Tech(i)/	F1		F2		F3	
Plant(F)	Operational	Investment	Operational	Investment	Operational	Investment
i1	8.90	5.24	11.05	7.68	10.68	6.28
i2	9.76	15.30	9.39	15.47	9.26	16.94
i3	5.20	12.44	6.03	11.72	5.96	11.43
i4	7.78	17.39	7.16	17.96	8.49	16.70
i5	4.59	15.87	5.02	15.69	4.12	16.44
i6	11.16	10.41	10.75	11.12	11.88	9.93
i7	9.95	11.81	8.73	12.26	9.93	12.96
i8	8.13	11.38	8.67	11.72	7.93	9.62
i9	9.06	7.68	9.28	7.06	9.51	8.00
i10	8.30	11.06	8.51	11.84	8.17	10.52
i11	7.17	8.94	7.95	9.02	7.29	9.48
i12	9.12	9.59	9.88	11.96	9.97	9.62
i19 to i25	0.3	0.5	0.3	0.5	0.3	0.5

Table 2: Warehouses operational and investment costs of assembling processes, for t=1, in €/ton

Tech(i)/	W1		W2		W3		W4		W5	
Warehouse(W)	Operat.	Inv.	Operat.	Inv.	Operat.	Inv.	Operat.	Inv.	Operat.	Inv.
i13	4.25	2	4.55	2	4.98	2	4.93	2	4.85	2
i14	4.06	2	3.08	2	3.76	2	4.85	2	4.12	2
i15	5.28	2	4.95	2	4.83	2	4.25	2	4.55	2
i16	4.93	2	4.06	2	5.28	2	4.86	2	3.75	2
i17	3.25	2	4.01	2	3.97	2	4.23	2	4.55	2
i18	5.01	2	5.16	2	5.23	2	5.12	2	5.12	2

4.1 Results without disruption

In this subsection we compare the three types of supply chains (cases A, B and C) when no disruption affects the normal operations. The augmented epsilon-constraint method is used to generate the

Table 3: Raw materials price for t=1, in €/ton

Prod.(P)/Sup.(RS)	RS1	RS2	RS3	RS4	RS5	RS6	RS7	RS8	RS9	RS10	RS11	RS12
P1	3.82	-	-	-	3.78	-	-	-	3.72	-	-	-
P2	6.37	-	-	-	6.27	-	-	-	6.18	-	-	-
P3	5.04	-	-	-	5.11	-	-	-	4.99	-	-	-
P4	4.34	-	-	-	4.3	-	-	-	4.33	-	-	-
P5	-	4.84	-	-	-	4.82	-	-	- 4.78	-	-	
P6	-	3.04	-	-	-	3.0	-	-	-	3.26	-	-
P7	-	4.94	-	-	-	4.23	-	-	- 5.0	-	-	
P8	-	1.73	-	-	-	2.40	-	-	- 2.80	-	-	
P9	-	-	2.0	-	-	-	1.88	-	-	- 1.99	-	
P10	-	-	4.11	-	-	-	3.88	-	-	-	3.96	-
P11	-	-	3.10	-	-	-	3.17	-	-	-	3.08	-
P12	-	-	4.45	-	-	-	4.18	-	-	-	4.38	-
P13	-	-	- 4.33	-	-	-	4.49	-		-	-	4.28
P14	-	-	- 1.26	-	-	-	1.48	-		-	-	1.18
P15	-	-	- 2.16	-	-	-	2.10	-		-	-	2.33

approximations to the Pareto-optimal curves that are presented in figure 4.

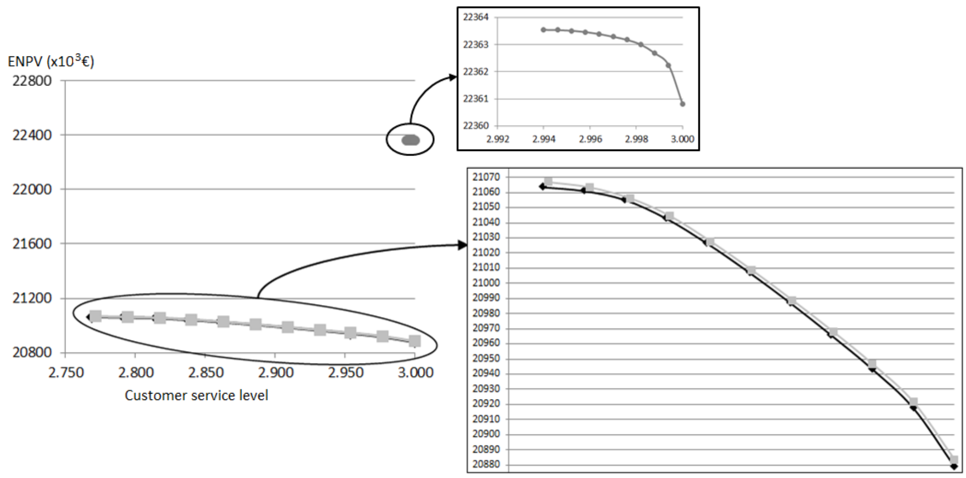


Figure 4: Approximation to the Pareto-optimal curve with no disruptions.

As can be observed, case A presents always worse ENPVs than the other two cases, although not really significant. This difference is justified by the fact that case A is characterized by a very straight network structure with no flexibility in terms of transportation links. Case B allows better results, but case C is the one that presents better results since all points of the curve dominate the points of cases A and B. It is important to notice that the length of curve C is much smaller than the others, because when the ENPV is maximized the customer service level obtained is of 100% in time period 1, 99.8% for time period 2 and 99.6% for time period 3, that corresponds to an accumulative value of 2.994. When the service level function is maximized, the maximum value of 3 is obtained, which represents a service level of 100% in all time periods. Although these two points are closer in terms of customer service level they represent slightly different supply chains structures. When the ENPV is maximized, all possible processes and infrastructures are installed, with an average of 1520 units for production processes, 1542 units for assembling processes, 466 for disassembling processes and 444 units for the capacity storage of the warehouses. Figure 5 represents the capacities of the processes and the storage entities for the maximum service level (case C). The processes inside dashed lines represent the disassembling processes.

In this case all new possible infrastructures were installed in the first time period and did not suffered any change during the planning horizon. The existing plant (Hamburg) remains with the initial processing capacity (600 units) for half of the processes. The other two plants (Bilbao and Milan) have also all processes installed. Regarding the warehouses, all processes suffered capacities expansions and present values closer but slightly higher than when the ENPV is maximized. These results are explained by the

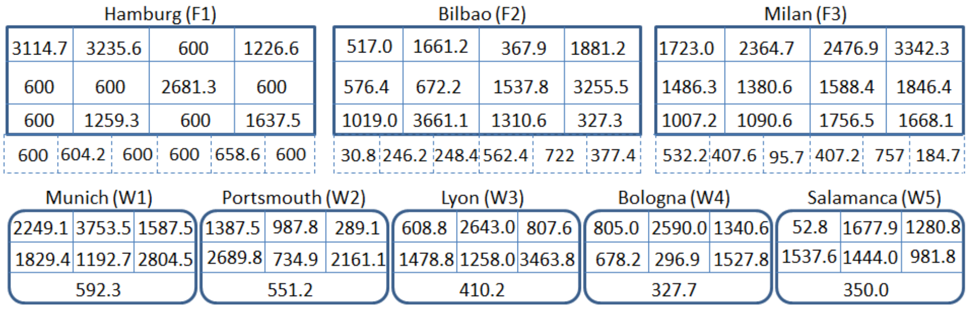


Figure 5: Capacities of the processes and storage entities for customer service maximization (case C).

fact that when maximizing the customer service level, using the augmented epsilon-constraint method, the results are the maximum ENPV that can be obtained for the maximum value of the costumer service. In this way, since the customer service level is closer in the extreme cases, the results are also similar. In order to better compare the two extreme points (maximize ENPV and maximize service level), figure 6 presents some operational results.

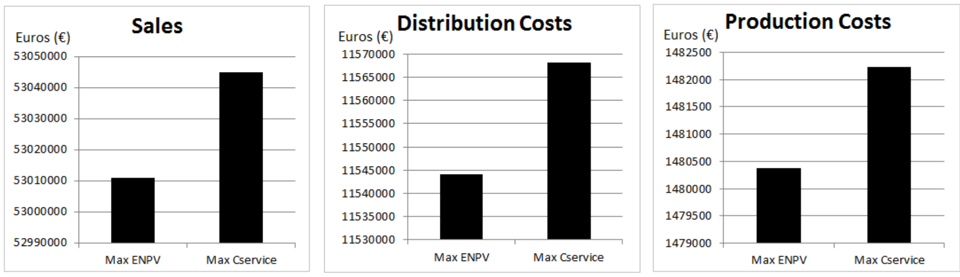


Figure 6: Operational results for the extreme points of the Pareto curve for case C.

As expected, as the extreme points are closer to each other, the operational results are similar. However, there are some differences. When the customer service level is maximized, in order to satisfy all the demand, the level of sales is higher, resulting in higher distribution and production costs. In none of the cases was verified outsourcing of final products, since this option implies higher costs.

4.2 Results with disruption

Disruption is responsible for a 100% decrease in the production capacity of the most important plant F3 (Milan) in time period 2 due to an earthquake occurrence. Two different analyses are made for each type of network: (1) when there is no upper limit on the investment budget; and (2) the situation where after the supply chain establishment (in time period 1) no budget is available to new investments. Figure 7 presents the Pareto-optimal curves for the three networks under disruption, with and without investment budgets limitations.

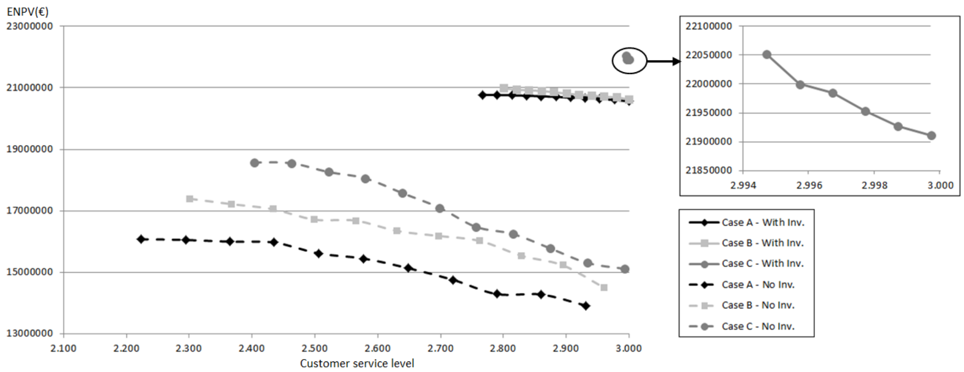


Figure 7: Approximation to the Pareto-optimal curves with disruption.

Once again, the supply chain that has more flexibility in terms of transportation links (case C) is the one that achieves better results in both cases. As expected, when there are no budget restrictions, after the disruption occurs there is an investment in more capacity, as can be observed in figure 8 for case C.

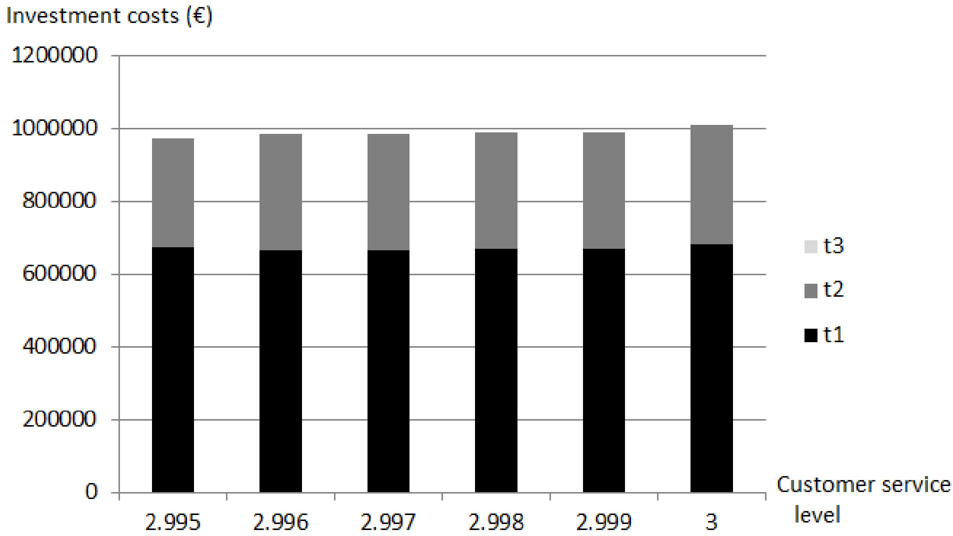


Figure 8: Investments after disruption for case C.

In this case, when the customer service level is maximized (equal to 3) the value of the investment made is the highest, since in order to satisfy all clients, it is necessary to invest in more capacity. For all points, the majority of the investment is made at the beginning of the planning horizon (t1), when the expansion of the capacity of the existing infrastructures in the supply chain occurs and new entities are installed. In time period 2, after the disruption occurs, there is also an investment in order to repair the damages. For time period t3 no investment is observed. For example, when the ENPV is maximized, which corresponds to a customer service level of 2.995, after the disruption all processes of plant F3 suffer an increase of, on average, 96% in capacity, meaning that the plant is rebuilt with almost the initial processing capacity. On the other hand, the other plants remain with the same capacities. When no more investment is allowed and in order to be able to continue supplying the costumers, there is an increase in the quantity of final products outsourced and a consequent increase in the distribution costs, as can be observed in figure 9.

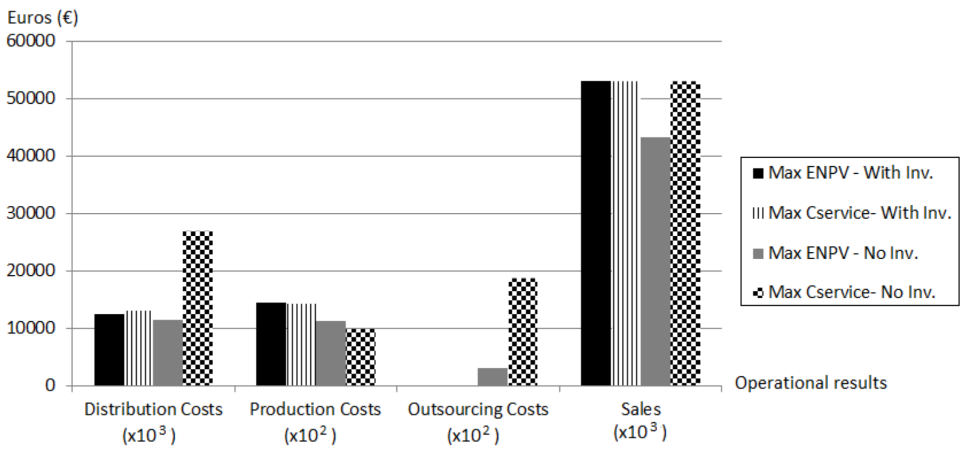


Figure 9: Operational results for case C under disruption, for the extreme cases.

The results obtained are similar for all three networks, but as case C is more flexible it presents better results. A more detailed analysis is done for this case. As shown in figure 9, when comparing the extreme cases of the Pareto-optimal curve, it is possible to conclude that when the resilience measure is maximized, the quantity of final products outsourced is higher and is even higher when there is no possibility to invest more after the disruption which results in higher distribution costs and lower production costs.

This leads us to conclude that in order to increase resilience after the occurrence of this type of disruption, supply chains, if possible, should invest in more internal processing capacity, otherwise other options such as outsourcing have to be explored.

4.3 Comparison of the results

By comparing the results obtained for the cases without disruptions (section 4.1), it is possible to see that a supply chain with more flexibility in terms of transportation links between all entities allows the achievement of better results than the others with more rigid logistic links. Although when disruptions occur the buildup of more resilient supply chains implies a reduction on the ENPV. The most flexible network presented has always better results in terms of both objectives than the others when the disruption occurred, which corroborates the literature in supply chains resilience that present flexibility of the networks as a possible resilient strategy allowing more resilience into the supply chains. When comparing the cases with and without budget to invest after a disruption occurs, we can conclude that having budget implies always better results for this type of disruption, since it allows the reconstruction of the damages caused.

In table 4 are presented the computational results obtained for each type of network. In the table,

Table 4: Computational results

	Total Variables		Binary Variables		Constraints		Gap(%)		CPU(s)	
Disruption	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Case A	2521873	1722798	2403	1602	240972	221417	0	0	2763.09	703.97
Case B	2524098	1724848	2418	1612	241726	222113	0	0	5616.75	1664.2
Case C	3082687	2166973	3645	2430	247069	227244	0.008%	0	7200	3732.6

the values presented for the gap and for the computational time are the highest obtained in the several models tested. As expected, as the complexity of the supply chain grows, the number of variables and constraints also increase, which results in an higher computational effort. When no disruption is present, cases A and B achieve optimality, but that does not happen with case C, that presents however a very low gap, since a maximum of 7200 seconds of computational time was imposed. When the disruption is implemented, as the model is only modeled for two time periods (t=2, when the disruption occurs and t=3), since the first time period results are inputs for the initial conditions of the supply chain for the model with the disruption, there are less variables and constraints, which results in the achievement of optimality for all cases with less computational time.

5 Conclusions

In this work, we have studied resilience in a closed-loop supply chain. To do that, a bi-objective model was developed for the design and planning of a generic CLSC considering uncertainty in products' demand using a scenario tree approach. The two objective functions are the maximization of the ENPV and the resilience, the latter being measured through the customer service level. Since the two objectives are contradictory, an approximation of the Pareto optimal curve was obtained through the application of the augmented epsilon-constraint method.

The model was applied to three types of networks with different levels of flexibility in terms of transportation links and their performances were compared when a disruption affected a production plant. This analysis was made considering two cases: when there is budget to invest in more capacity after the disruption occurs and when there is no budget available. We may conclude that in order to be more resilient a supply chain should invest in flexibility in terms of transportation links and redundancy, having extra capacity in its processes and infrastructures.

For future work, it is intended to incorporate uncertainty at the level of the returns and regarding resilience, other types of disruptions should be implemented as well as other metrics to measure resilience should be investigated.

Acknowledgements

The authors acknowledge the financial support from Fundação para a Ciência e Tecnologia (FCT), grant SFRH/BD/64125/2009 and research project PTDC/SEN-ENR/102869/2008.

References

[Amaro and Barbosa-Póvoa, 2009] Amaro, A. and Barbosa-Póvoa, A. (2009). The effect of uncertainty on the optimal closed-loop supply chain planning under different partnerships structure. *Computers and Chemical Engineering*, 33(12), 2144-2158.

[Amin and Zhanq, 2012] Amin, S. and Zhang, G. (2012). A multi-objective facility location model for closed-loop supply chain network under uncertain demand and return. *Applied Mathematical Modelling*, doi:http://dx.doi.org/10.1016/j.apm.2012.09.039.

[Cardoso et al., 2013] Cardoso, S., Barbosa-Póvoa, A. and Relvas, S. (2013). Design and Planning of Supply Chains with Integration of Reverse Logistics Activities under Demand Uncertainty. *European Journal of Operational Research*, 226, 436-451.

[Carvalho et al., 2012] Carvalho, H., Barroso, A., Machado, V., Azevedo, S. and Cruz-Machado, V. (2012). Supply chain redesign for resilience using simulation. *Computers and Industrial Engineering*, 62, 329-341.

[Christopher and Peck, 2004] Christopher, M. and Peck, H. (2004). Building the resilient supply chain. *International Journal of Logistics Management*, 15(2), 1-13.

[Guide and Van Wassenhowe, 2002] Guide, V. and Van Wassenhowe, L. (2002). The Reverse Supply Chain. *Harvard Business Review*, 80(2), 25-26.

[Özkir and Basligil, 2012] Özkir, V. and Basligil, H. (2012). Modelling product-recovery process in closed-loop supply chain network design. *International Journal of Production Research*, 50(8), 2218-2233.

[Pfohl et al., 2010] Pfohl, H., Kohler, H. and Thomas, D. (2010). State of the art in supply chain risk management research: empirical and conceptual findings and a roadmap for the implementation in practice. *Logistics Research*, 2,33-44.

[Salema et al., 2010] Salema, M., Barbosa-Póvoa, A. and Novais, A. (2010). Simultaneous design and planning of supply chains with reverse flows: A generic modelling formulation. *European Journal of Operational Research*, 203, 336-349.

[Sawik, 2013] Sawik, T. (2013). Selection of resilient supply portfolio under disruption risks. *Omega*, 41, 259-269.

[Schmitt and Singh, 2012] Schmitt, A. and Singh, M. (2012). A quantitative analysis of disruption risk in a multi-echelon supply chain. *International Journal of Production Economics*,139, 22-32.

[Spiegler et al., 2012] Spiegler, V., Naim, M. and Wikner, J. (2012). A control engineering approach to the assessment of supply chain resilience. *International Journal of Production Research*, DOI: 10.1080/00207543.2012.710764.

[Vugrin et al., 2011] Vugrin, E., Warren, D. and Ehlen, M. (2011). A Resilience Assessment Framework for Infrastructure and Economic Systems: Quantitative and Qualitative Resilience Analysis of Petrochemical Supply Chains to a Hurricane. *Process Safety Progress*, 30(3), 280-290.

[Zeballos et al., 2012] Zeballos, L., Gomes, M., Barbosa-Póvoa, A. and Novais, A. (2012). Addressing the uncertain quality and quantity of returns in closed-loop supply chains. *Computers and Chemical Engineering*, 47, 237-247.

Benchmarking dos Serviços dos Hospitais Portugueses: Uma Aplicação de *Data Envelopment Analysis*

Ricardo A. S. Castro*, Conceição Silva Portela†, Ana S. Camanho‡

* *Faculdade de Engenharia da Universidade do Porto*
E-mail: `ricardo.alves.castro@fe.up.pt`

† *Universidade Católica Portuguesa no Porto,*
E-mail: `csilva@porto.ucp.pt`

‡ *Faculdade de Engenharia da Universidade do Porto.*
E-mail: `acamanho@fe.up.pt`

Resumo

Neste estudo, apresentamos um modelo de avaliação de serviços hospitalares, numa perspetiva de eficiência. Foram escolhidos múltiplos *inputs* e *outputs*, após análise dos seus impactos nos gastos globais dos serviços. Um modelo de DEA foi aplicado a um serviço, Medicina Interna, impondo-se restrições aos pesos e evitando-se obter serviços com pesos pouco razoáveis. As maiores poupanças são possíveis nos medicamentos e material clínico. É feita uma comparação entre serviços eficientes e ineficientes, observando-se que também os meios complementares de diagnóstico e terapêutica (recursos) e as variáveis de acesso aos cuidados (elementos da produção) são determinantes na definição do melhor desempenho.

Palavras chave: *Data Envelopment Analysis*, Avaliação Hospitalar, Eficiência de Serviços.

1 Introdução: Avaliação de Hospitais – Motivação e Desafios

Em todos os países da Organização para a Cooperação e Desenvolvimento Económico (OCDE) tem-se assistido a um aumento constante das despesas em saúde, aumento que tem sido superior ao aumento percentual do Produto Interno Bruto (PIB) nestes países [OCDE(2011)]. Portugal e a grande maioria dos países da OCDE apresentam mecanismos de proteção dos seus cidadãos na doença e, por isso, apresentam uma preponderância dos gastos públicos em saúde, face aos gastos privados. No nosso país, a maior parte da despesa pública incorrida pelo Serviço Nacional de Saúde (SNS) deve-se aos hospitais (mais de 50%).

Os aumentos de custos com saúde têm exercido uma grande pressão sobre os défices nacionais dos países desenvolvidos, o que tem levado os diversos Estados a agir no sentido da sua contenção, tendo em vista a sustentabilidade dos sistemas de saúde no longo prazo. Muito em voga até ao fim da década de 80, a doutrina "mais mercado" na Saúde foi abandonada, mesmo pelo Banco Mundial, tendo o enfoque passado a ser a eficiência da intervenção estatal [Simões(2004)]. A necessidade de controlar mais de perto a despesa pública tem também levado à gradual aceitação da necessidade de se introduzirem mecanismos de avaliação mais profundos nos processos de tomada de decisão nas administrações públicas, nomeadamente ao nível da saúde.

As políticas nacionais para o setor da saúde, nomeadamente a que conduziu à introdução de hospitais empresa e à separação do papel do Estado como prestador e financiador (Lei 27/2002 de 8 de Novembro), supõem, pela sua natureza, a criação de mecanismos de avaliação por parte das próprias instituições hospitalares e quase forçam uma avaliação aos hospitais, para se aferir o impacto da sua implementação. Apesar de necessária, a avaliação do desempenho de um hospital não é uma atividade trivial [Costa e Lopes(2007)]. Das diferentes dificuldades, podem destacar-se a inexistência de objetivos claros e comuns para a atividade de todas as unidades, a existência de objetivos não mensuráveis [Simões(2004)] ou mesmo a existência de diferentes agentes [Costa e Lopes(2007)], no processo da prestação de cuidados de saúde. A extensão de estudos de avaliação do desempenho hospitalar, em Portugal, é, talvez por isso, ainda reduzida. Os estudos que têm sido realizados baseiam-se, maioritariamente, em programas de certificação, acreditação e *benchmarking*, promovidos por entidades anglófonas. Nestas avaliações, não é, contudo, usual utilizarem-se metodologias de fronteira para avaliar o desempenho dos hospitais, preferindo-se uma avaliação baseada numa longa lista de indicadores, sem se conseguir obter uma imagem única e direta da realidade hospitalar. Isto fica bem patente ao analisar-se os relatórios de retorno enviados pela Administração Central do Sistema de Saúde (ACSS) aos hospitais do SNS. Esta não é uma realidade exclusivamente nacional, como se pode constatar na longa lista pública de indicadores usados pelo *National Health System* britânico.

As metodologias de fronteira são, no entanto, largamente utilizadas pela comunidade académica [Hollingsworth(2008)]. Na avaliação de cuidados médicos, especialmente em hospitais ou unidades de cuidados primários, consideram-se na literatura, usualmente, duas perspetivas de análise: uma ligada à gestão, que é responsável pela afetação dos recursos necessários à atividade clínica, e outra ligada ao tratamento dos pacientes, que se designa por perspetiva clínica [Chilingerian e Sherman(2004)]. Estas perspetivas estão interligadas, já que os *outputs* da primeira coincidem com os *inputs* da segunda (elementos denominados de *outputs* intermédios). No entanto, enquanto na primeira perspetiva uma gestão eficiente terá como objetivo minimizar o *input*, para um dado nível de *output* intermédio (a capacidade instalada e disponibilizada ao médico, como dias de internamento, número de consultas, etc.), na segunda o tratamento clínico adequado dos doentes preocupa-se com a maximização dos *outcomes* positivos no estado da saúde dos indivíduos, que refletem os resultados finais sobre a sua saúde.

Tendo em conta que não existe uma referência absoluta de desempenho hospitalar, e dado o contexto aqui apresentado, considerou-se oportuno e valioso um estudo que permitisse avaliar os hospitais, comparando-os entre si. Considerou-se que um modelo de *Data Envelopment Analysis* (DEA) poderia ser um elemento fundamental na identificação das melhores práticas e na deteção das maiores fontes de ineficiência, agregando múltiplas variáveis que caracterizem a produção hospitalar. Neste estudo, apresenta-se um modelo para a avaliação da eficiência hospitalar, centrada na atividade de gestão, ao nível dos serviços. Para isso, começa-se por apresentar a técnica usada para avaliar os hospitais, a DEA. De seguida, apresenta-se forma como os hospitais têm sido avaliados, na literatura, e quais os fatores que têm sido tidos em consideração em análises de eficiência. De seguida, apresenta-se o modelo teórico que resulta da análise aos dados obtidos. Estes dados, relativos a 2008, provêm da Base de Dados dos Elementos Analíticos (contabilidade de gestão) dos hospitais do SNS, disponibilizada diretamente pela ACSS. Os restantes dados referem-se aos recursos e à produção dos hospitais, também cedidos pela ACSS. Posteriormente à idealização do modelo, ele será concretizado numa avaliação do serviço de Medicina Interna de hospitais do SNS.

2 *Data Envelopment Analysis*

A DEA é uma metodologia que permite não só tratar a complexidade inerente à comparação entre unidades que apresentem multiplicidade de *inputs* e *outputs*, como permite agregar numa única medida os resultados desta comparação complexa. O resultado da DEA é uma medida de eficiência para cada unidade de decisão (em inglês *Decision Making Unit* ou DMU), calculada com base na comparação dessa mesma unidade com uma ou várias unidades de referência, aquelas que representam o melhor desempenho observado. Como esta metodologia é baseada num cálculo de produtividade assente em múltiplos *inputs* e *outputs*, existe a necessidade de atribuir pesos distintos a cada um desses itens, de forma a evidenciar a importância relativa de cada um deles. Para eliminar arbitrariedades, o método consiste no cálculo da produtividade relativa (ou eficiência) de cada DMU, com os pesos que mais a beneficiam. Este processo de otimização permite encontrar unidades de referência que formam a fronteira da produção, lugar onde se situam as unidades eficientes. Nestas unidades, um aumento num qualquer *output* implica a redução de um outro *output*, ou o aumento de pelo menos um dos *inputs* (ou vice-versa) [Fried et al.(2008)].

Na prática, a metodologia de DEA utiliza um modelo de programação linear para calcular as eficiências de cada uma das unidades em estudo. Dependendo das assunções e das restrições impostas pelo (ou ao) objeto de estudo, o programa linear varia. Por exemplo, no caso de na situação em estudo se verificar que os aumentos de *outputs* não são proporcionais aos aumentos de *inputs* (rendimentos variáveis à escala), o modelo a usar será o BCC (do trabalho original de [Banker et al.(1984)]):

$$\max_{v,u} \theta = \sum_r u_r y_{ro} - w \tag{1}$$

$$\text{Sujeito a:} \sum_i v_i x_{io} = 1 \tag{2}$$

$$\sum_r u_r y_{rj} - \sum_i v_i x_{ij} - w \leq 0, (j = 1, \ldots, n) \tag{3}$$

$$v_i, u_r \geq 0, \forall i, r \tag{4}$$

$$w \in \Re \tag{5}$$

Neste programa, aplicado a cada uma das unidades em estudo, considera-se existir n unidades em avaliação, m *inputs* e s *outputs*, sendo o a unidade em estudo. Este programa consiste na maximização do numerador do rácio de produtividade, com as variáveis de decisão a serem os pesos dos *inputs* (v_i) e *outputs* (u_r) e uma variável livre (w). Ao denominador do rácio de produtividade impõe-se uma restrição que o faça ser igual a 1. As restantes restrições do modelo impõe que unidades de dimensão semelhante à unidade o, ao usarem os mesmos pesos, não possam ter um valor de eficiência (θ) superior a 1. É a imposição destas restrições que torna o resultado uma medida da eficiência de cada unidade, já que elas criam uma referência relativamente à qual todas as unidades se comparam. Esta medida de eficiência representa o fator pelo qual cada unidade deve multiplicar os seus *inputs*, de modo a situar-se sobre a fronteira de produção, isto é, sem alterar o seu *mix* (proporção) de *inputs* (ou de *outputs* para a orientação alternativa).

Existem outras variantes aos modelos básicos de DEA. Uma delas, consiste na introdução de restrições aos valores dos pesos. Esta variante é utilizada quando se pretende impor ao modelo características resultantes de um julgamento *a priori* do avaliador, que o torne mais realista, ou mesmo quando se pretende fazer uma análise de sensibilidade (repare-se como a livre escolha dos pesos permitida pela DEA implica a não assunção de quaisquer juízos de valor *a priori*). Uma das formas de impor este tipo de restrições consiste na utilização do método da região de confiança (traduzido do inglês *Assurance Region Method*, proposto por [Thompson et al.(1986)]), em que se enquadra o valor dos rácios entre os pesos de *inputs* (ou *outputs*) distintos [Thanassoulis et al.(2004)]:

$$\alpha \leq \frac{v_2}{v_1} \leq \beta \tag{6}$$

Estas restrições poderiam ser também aplicadas aos pesos virtuais dos *inputs* e *outputs*, isto é, o produto do valor de cada variável com os respetivos pesos. Contudo, a sua utilização implica, de forma indireta, uma limitação absoluta dos pesos do modelo e não um enquadramento relativo [Thanassoulis et al.(2004)]. A colocação da equação 6 no modelo BCC, altera a fronteira eficiente, diminui o domínio do modelo de programação linear e altera a natureza radial do modelo (para detalhes sobre o assunto, pode consultar-se [Portela e Thanassoulis(2006)]). Como consequência, as unidades veem a sua eficiência piorar ou manter-se, face ao modelo sem restrições de pesos.

3 Avaliação de Eficiência nos Hospitais com *Data Envelopment Analysis* – Revisão da Literatura

Na literatura de ciência da gestão, a DEA é a metodologia mais usada em estudos de desempenho em saúde e são os hospitais, por larga margem, os alvos preferenciais destes mesmos estudos [Hollingsworth(2008)]. Estes estudos tentam avaliar a eficiência destas organizações complexas, onde não existe um modelo de produção globalmente aceite e onde, por isso, as avaliações com DEA poderão ser muito sensíveis a erros de medição ou a uma fraca aferição da realidade, por omissão consciente ou inconsciente de variáveis [Hollingsworth et al.(1999)]. É por isto mesmo que se defende que a utilização de DEA numa avaliação de hospitais deveria cingir-se a aferir tendências gerais ou a avaliar unidades mais pequenas e concretas, como por exemplo serviços ou mesmo práticas médicas, onde os potenciais erros de medição serão menores. Os hospitais não são, por isso, considerados unidades suficientemente homogéneas para permitirem a utilização simples e direta desta técnica, apesar da literatura se centrar no seu estudo.

Ao contrário da literatura de avaliação de hospitais, que é vasta, a literatura de avaliação de unidades subhospitalares não o é. Os estudos de [Puig-Junoy(1998)] e [Kontodimopoulos e Niakas(2005)] são exemplos de análises feitas a este nível com recurso à DEA. No primeiro caso, o estudo tenta discernir o que determina o desempenho de unidades de cuidados intensivos. O segundo trabalho estuda unidades de hemodiálise gregas, e discute as dificuldades de análise de unidades subhospitalares, usando DEA. Em Portugal, encontram-se estudos semelhantes a estes, como os que avaliam Grupos de Diagnósticos Homogéneos (GDH). Como exemplo destes estudos, pode destacar-se o de [Dismuke e Sena(2001)] que, tendo por base os dois mais frequentes GDH (o das desordens cerebrovasculares, exceto ataques isquémicos transitórios, e o das falhas e ataque cardíacos) estudaram a evolução da produtividade de 3 tecnologias de diagnóstico, ao longo de 3 anos (1992 a 1994). Estes estudos avaliaram se o aumento de produtividade nestas tecnologias trouxe consigo uma diminuição da qualidade, aferida pela mortalidade. Utilizando um modelo de DEA com *outputs* indesejáveis e os índices de Malmquist-Luenberger (usados para aferir a evolução da produtividade), verificou-se que não havia evidência de tal realidade.

A aplicação da DEA ao setor hospitalar usa tradicionalmente uma orientação para *inputs*, justificando-se este facto pela predominância da perspetiva da eficiência na avaliação hospitalar. Isto resulta da preocupação que os diferentes agentes da saúde têm na contenção de recursos, uma vez que a produção hospitalar é, essencialmente, ditada pela procura da população, um fator exógeno ao hospital. Nesta categoria, as variáveis mais utilizadas correspondem ao número de camas, número de médicos, enfermeiros, outros funcionários ou mesmo os custos que estas categorias profissionais representam, assim como os custos com consumíveis (medicamentos e material clínico, principalmente) [Castro(2011)]. A maioria dos estudos analisa a eficiência técnica e, tendo em vista uma homogeneização de *outputs*, utilizam-se variáveis que fazem referência à complexidade dos doentes tratados (*case-mix*) ou ao número de serviços dos hospitais (*service-mix*). Devido à inexistência de um modelo de produção globalmente aceite, bem como à heterogeneidade da atividade hospitalar, não há consenso relativamente ao tipo de rendimentos à escala a assumir nas análises, optando-se muitas vezes pelo uso dos dois e posterior comparação.

Em Portugal, a atividade avaliativa tem vindo a ser objeto de atenção crescente e tem-se centrado, principalmente, nas recentes transformações que o setor da saúde tem sofrido. Mais recentemente, utilizou-se o DEA em grande medida, ao nível dos hospitais, para avaliar a reforma da transformação dos hospitais públicos em Sociedades Anónimas (SA) (ou Entidades Públicas Empresariais - EPE, como atualmente se denominam). Deste universo, podem destacar-se os estudos do [Tribunal de Contas(2006)], de [Moreira(2008)] e de [Gonçalves(2008)]. O primeiro tinha por objetivo medir a eficiência económica dos hospitais transformados e medir a qualidade e equidade no acesso aos cuidados de saúde. Só no estudo da eficiência económica se utilizou o DEA. Em [Moreira(2008)], utilizou-se o grupo dos hospitais do Setor Público Administrativo (SPA) como um grupo de controlo, criando-se dois tipos de fronteiras. As conclusões, baseadas em dados de 2001 a 2005, apontam para uma maior eficiência média inicial no grupo dos hospitais SPA, o que se reverteu no ano de 2005. O estudo de [Gonçalves(2008)] obteve um resultado contrário. Utilizando dados referentes ao período entre 2002 e 2004, [Gonçalves(2008)] verificou, usando os Índices de Malmquist, que a fronteira de eficiência técnica evoluiu positivamente neste período, tendo sido os hospitais SPA os que mais contribuíram para esta evolução.

Em conjunto, os estudos apresentam evidência da inadequação do método quando se avaliam hospitais, pela impossibilidade da definição de modelos que capturem mais fielmente a sua complexidade. Em Portugal, não existe, de momento, outra avaliação que, usando DEA, se foque nos serviços dos hospitais.

4 Modelo de DEA para os Serviços Médicos/Cirúrgicos dos Hospitais do SNS numa Perspetiva de Eficiência

A estrutura dos hospitais não é homogénea. Cada hospital tem a sua estrutura interna definida de acordo com a sua influência geográfica e, principalmente, de acordo com as valências/especialidades implementadas. No entanto, algo é comum a todos: os serviços são as suas unidades funcionais fundamentais. Destes, são os serviços das especialidades médicas ou cirúrgicas quem tem a responsabilidade de servir diretamente os pacientes através de internamentos, consultas externas, sessões em hospital de dia ou cirurgias em ambulatório, quatro das cinco principais linhas de produção que um hospital realiza. A estes serviços que, em 2008, foram responsáveis por 53% dos custos diretos dos hospitais analisados, chamaremos "serviços principais", em contraponto com os restantes, que denominaremos de "secções de apoio", por incluírem os restantes grupos de serviços. Verificou-se que, cada um destes grupos tem, nos custos diretos, um peso bastante inferior ao dos serviços principais. É, portanto, para os serviços principais que este estudo apresenta um modelo "ideal", com o objetivo da sua avaliação numa perspetiva de eficiência, considerando apenas a produção das suas quatro principais linhas de produção. Como é nestes serviços que se realiza a maioria da produção hospitalar, é neles que são maioritariamente imputados todos os custos das restantes secções (secções de apoio), o que reforça a sua importância num primeiro estudo.

A criação do modelo iniciou-se com uma revisão bibliográfica, que teve por finalidade identificar as variáveis mais frequentes nos estudos da área. Esta revisão pode ser consultada em [Castro(2011)]. Posteriormente, e para verificar se às variáveis mais usadas na literatura nacional e internacional, correspondiam os recursos mais relevantes à realidade hospitalar do SNS, identificaram-se as estruturas de custos médias para todos os serviços de especialidades médicas e cirúrgicas, com o objetivo de identificar as que melhor as caracterizam. Nesta análise, contabilizaram-se todos os custos incorridos pelas linhas de produção de internamento, consulta externa e hospital de dia, já que a ACSS não dispõe de informação relativa à discriminação dos custos com cirurgia em ambulatório pelos diferentes serviços. No entanto, esta linha de produção apresentou, em 2008, custos residuais face aos restantes.

Como pode ser visto na Figura 1, os tipos de custos mais representativos nos dois grupos de serviços são os mesmos (variando apenas o peso de cada um deles), à exceção do Bloco Operatório, não utilizado pelas especialidades médicas (não realizam cirurgias). As principais categorias de custos são, então, os custos com Pessoal (CP), Matérias Vendidas e Consumidas (CMVMC – representados esmagadoramente pelos medicamentos e material clínico), Meios Complementares de Diagnóstico e Terapêutica (MCDT), Serviços Administrativos (SA), Fornecimentos e Serviços Externos (FSE) e Serviços Hoteleiros (SH). Nos serviços médicos, estas categorias são responsáveis por 93% do total de custos . Aos serviços cirúrgicos, acrescenta-se o Bloco Operatório (BO) e o total destas categorias representa 94% do total.

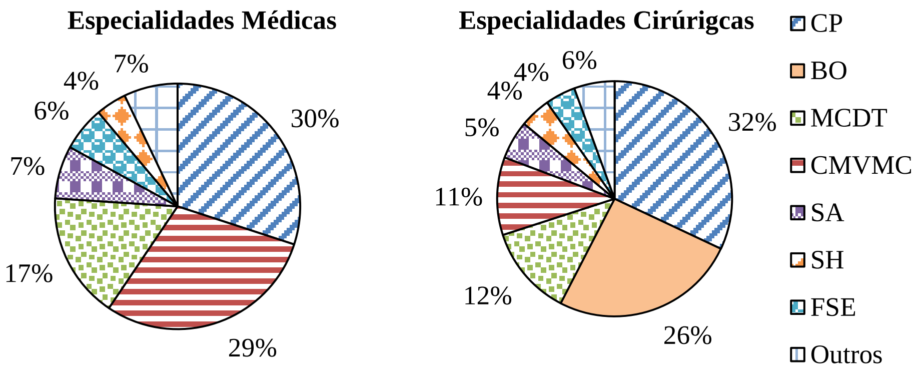


Figura 1: Estrutura dos Custos Gerais (Diretos e Imputados) dos Serviços Principais.

Sendo os custos com o pessoal uma categoria bastante heterogénea, foi necessário avaliá-la de forma mais pormenorizada. Realizando uma análise semelhante à anterior, verificou-se que 90% e 92% dos custos com pessoal nos serviços médicos e cirúrgicos, respetivamente, são atribuíveis aos médicos, enfermeiros e ao pessoal operário e auxiliar.

Após a análise à estrutura de custos dos serviços principais, foi possível definir uma lista de *inputs* e *outputs* para modelar, idealmente, o processo de transformação de recursos numa capacidade clínica de tratar doentes (*outputs* intermédios). Como se pretende calcular a eficiência técnica das unidades em avaliação, definem-se todas as variáveis, em termos de quantidades e não de valores económicos.

Para representar os recursos de pessoal, optou-se por considerar o número de horas semanais disponibilizadas pelos médicos e enfermeiros, como dois dos *inputs* do modelo. Como o pessoal operário e auxiliar tem um peso nos custos bastante inferior ao das classes clínicas (pesa, nos dois tipos de especialidades, somente 10%), optou-se por representar este recurso por uma variável mais genérica. Pretendeu-se que esta variável representasse outros recursos que se possam assumir proporcionais à dimensão de um serviço. A variável escolhida foi o número de camas. Além de representar o pessoal operário, a variável "número de camas"representa, neste modelo, os custos com os FSE (muitas vezes considerados, na literatura, proporcionais à dimensão de um serviço), com os serviços administrativos e com os serviços hoteleiros. O número de camas é, assim, uma variável que funciona como *proxy* para estas quatro categorias de recursos. Para atestar esta opção, verificou-se a existência de correlações significativas entre estes recursos e o número de camas (valores entre 0,70 e 0,94), o que confirma a proporcionalidade. Os custos incorridos pelos serviços principais, com MCDT, devem-se à prestação de diferentes serviços, por parte de unidades especializadas em meios complementares de diagnóstico e terapêutica, aos primeiros. Isto leva à necessidade de se contabilizar, de forma ponderada, todos os procedimentos, para se refletir a diferente complexidade, associada aos custos, que estes serviços acarretam. O ponderador dos procedimentos, em Portugal, é definido por portaria do Ministério da Saúde (à data do estudo, Portaria nº 839-A/2009). Os custos com MCDT seriam, então, representados idealmente por esta variável. Para representar os CMVMC, seria ideal contabilizarem-se os medicamentos e material clínico, de forma semelhante ao que é feito com os MCDT. Contudo, não existe idêntica definição de ponderadores. Assim, na prática e por agora, não resta alternativa à utilização dos custos para a sua ponderação. Na lista de recursos apresentada abaixo permanece a variável que seria ideal utilizar (e que um dia talvez se possa utilizar), enquanto na secção seguinte, em que se demonstra a aplicação prática do modelo teórico aqui proposto, se utilizam os custos incorridos com os CMVMC, como *proxy* para as quantidades de matérias ponderadas usadas pelos serviços. Para se representar os recursos gastos, pelos serviços, com o bloco operatório (BO), optou-se por escolher a variável que representa o tempo médio semanal que cada especialidade ocupa neste serviço auxiliar. Além do tempo total das intervenções, pode considerar-se o número total de intervenções como produto do BO. O número de horas que um serviço ocupou no bloco será vantajoso relativamente ao total de intervenções, por aferir, simultaneamente, a complexidade e quantidade das mesmas, assumindo-se uma qualidade uniforme dos médicos.

Por fim, é de salientar que as variáveis de recursos escolhidas, e sumariadas na Tabela 1, se referem a 89 e 92% dos custos totais das quatro linhas de produção consideradas nos serviços de especialidades médicas e cirúrgicas, respetivamente, em 2008.

Tabela 1: Variáveis de *Input* dos Modelos de DEA para os Serviços Principais

Nome	**Definição Operacional**
Nº Horas Médicos	Contabilização do nº de horas semanais de trabalho médico.
Nº Horas Enfermeiros	Contabilização do nº de horas semanais de trabalho enfermeiro.
Nº Exames Ponderados	Contabilização do nº de procedimentos de MCDT requeridos, ponderados pelos fatores definidos na portaria do Ministério da Saúde mais recente.
Nº Matérias Ponderadas	"Contabilização"do nº ponderado de medicamentos e materiais clínicos.
Nº Camas	Contabilização do nº de camas existentes.
Nº Horas de Bloco	Contabilização do nº de horas totais das intervenções cirúrgicas realizadas.

Relativamente às variáveis de *output*, a sua escolha deve refletir o que é fundamental ser medido: o que é produzido pelas diferentes linhas de produção. Nisto incluem-se o número total de doentes atendidos em cada linha de produção, que pode também ser considerado uma medida de acesso aos cuidados de saúde, e a quantidade total de tratamento produzido e disponível para todos os utentes, medida usualmente pelo número de unidades pelas quais se pode dividir cada tipo de tratamento (*proxies* desta quantidade). O número de doentes atendidos, em internamento, é usualmente contabilizado pelo número de altas num ano, em consulta externa, pelo número de primeiras consultas e, em hospital de dia, pela contabilização direta do número de doentes. O número de cirurgias em ambulatório reflete o número de doentes atendidos e as quantidades totais de tratamento produzidas, considerando-se um novo doente sempre que se realiza uma nova cirurgia. Quanto ao total de tratamento produzido nas restantes linhas, ele é usualmente medido, para o internamento, pelo número de dias total que os doentes permaneceram internados, para as consultas, pelo seu total (a soma das primeiras com as consultas de acompanhamento) e, para o hospital de dia, pelo número total de sessões a que os utentes acederam. Todos estes *outputs* encontram-se sumariados na Tabela 2.

Tabela 2: Variáveis de *Output* do Modelo de DEA para os Serviços Principais

Nome	Definição Operacional
Tempo Estadia Total	Soma do nº de dias de estadia dos pacientes que saem do internamento.
Nº Altas Internamento (corrigidas pelo *case-mix*)	Contabilização do nº total de altas de internamento multiplicadas pelos índices de *case-mix* respetivos.
Nº de Consultas Externas	Contabilização do nº de consultas externas.
Nº 1as consultas	Contabilização do nº de primeiras consultas
Nº Sessões Hospital Dia	Nº total de sessões de Hospital de Dia.
Nº Utentes Hospital Dia	Contabilização doentes com acesso a sessões de hospital de dia.
Nº Cirurgias Ambulatório (corrigidas pelo *case-mix*)	Contabilização do nº de cirurgias em ambulatório ponderadas pelo respetivo índice de *case-mix.*

Ao medir ambas as dimensões da produção, comparando duas unidades com o mesmo nível de *inputs* e de produção em quantidade (com os mesmos dias de doentes em internamento, por exemplo), a que apresente o menor acesso de pessoas aos cuidados (menor número de altas), apresenta-se como uma unidade mais ineficiente. No entanto, a um fraco acesso aos cuidados de saúde proporcionado por um hospital (baixo número de altas, por exemplo), pode estar associada uma maior complexidade das doenças que este trata, o que faz prolongar a estadia de um só doente. Para controlar esta heterogeneidade das doenças, utiliza-se aqui a variável mais comum e a única disponível em Portugal: o índice de *case-mix* [CIDES-FMUP-UP e ACSS(2011)]. Este índice apenas pode ser calculado para o internamento e para as cirurgias em ambulatório e não é usado como um *output per se*, mas como um fator de ajuste, multiplicando-se o seu valor pela variável a ajustar.

5 Análise de Eficiência ao Serviço de Medicina Interna – Caso de Estudo

5.1 Descrição do Procedimento de Implementação do Modelo

Concretizou-se, neste estudo, um modelo piloto para avaliar o desempenho de um serviço. O escolhido foi o serviço de Medicina Interna, por ser o segundo serviço principal mais dispendioso, o que lida com os casos mais comuns e o que existe num maior número de hospitais. Para este serviço médico, o modelo "ideal"tem como variáveis de *input*, o "Nº de Horas de Médicos", o "Nº de Horas de Enfermeiros", o "Nº de Exames Ponderados", o "Nº de Matérias Ponderadas"e o "Nº de Camas". Quanto aos *outputs*, sendo esta uma especialidade onde não existe hospital de dia nem cirurgias, serão relevantes as variáveis "Tempo Estadia Total", "Nº Altas Internamento", "Nº de Consultas Externas"e "Nº 1as consultas".

Ao tentar implementar o modelo exemplificativo, encontramos dificuldades na obtenção de alguns dados e na fiabilidade de outros (para uma discussão aprofundada do assunto, consultar [Castro(2011)]). Para além disto, para construir o modelo apresentado na secção anterior, não foi possível obter alguns dados, como o índice de *case-mix* do serviço de Medicina Interna, a quantidade de MCDT ponderados consumidos e as horas de trabalho de enfermeiros, embora esta informação seja recolhida pela ACSS. Devido a estas restrições, foram utilizados os custos incorridos pelos serviços com os enfermeiros, com os exames ponderados e com as matérias ponderadas, como *proxy* para as quantidades em falta. O índice de *case-mix* usado foi o índice para as especialidades médicas de cada hospital, assumindo-se este como *proxy* do de Medicina Interna, algo justificável pelo tipo de doenças tratadas por esta disciplina.

Nos dados recolhidos, contabilizaram-se 55 hospitais que apresentaram dados para o serviço em análise (hospitais que têm, de facto, esta valência), mas apenas 48 foram avaliados. Foram eliminados da amostra os hospitais que apresentavam falhas em qualquer das variáveis, quer por valores em falta, quer terem uma ordem de grandeza incompatível com a realidade medida. Apresentam-se as estatísticas descritivas dos dados na Tabela 3, que evidenciam a heterogeneidade da amostra analisada.

Tabela 3: Estatística Descritiva das Variáveis Utilizadas no Modelo de DEA

Variável	Média	Desvio Padrão	Máximo	Mínimo
Inputs				
x_1 – Horas Médicos (h)	1350	1262	6420	46
x_2 – Custos Enfermeiros (€)	1509070	1371092	6099323	27602
x_3 – Custos Exames Ponderados (€)	1476455	1532799	7980923	27687
x_4 – Custos Matérias Ponderadas (€)	1410007	1625900	8805764	297
x_5 – Nº Camas	78	58	260	0
Outputs				
y_1 – Tempo Estadia Total (dias)	28159	21068	102931	0
y_2 – Nº Altas Internamento (corrigidas pelo *case-mix*)	2680	2568	15032	0
y_3 – Nº de Consultas Externas	8350	6334	23932	529
y_4 – Nº Primeiras Consultas	1665	1436	8324	108

Utilizou-se o *software* não-comercial EMS (*Efficiency Measurement System)* para calcular os resultados de eficiência das diferentes unidades. Verificou-se que os serviços hospitalares operam com rendimentos variáveis à escala, o que significa que o estudo de *benchmarking* deve apenas comparar diretamente, unidades com dimensão semelhante. No entanto, como a amostra analisada neste estudo era pequena (48 serviços de Medicina Interna), em relação ao número de *inputs* e *outputs* definidos, a análise de DEA, assumindo rendimentos variáveis à escala, não permitiu fazer uma discriminação clara entre o desempenho das várias unidades. A maioria (28) foi mesmo considerada eficiente na análise sem restrições de pesos, por assunção de pesos nulos para várias variáveis, distintas de serviço para serviço. Conhecendo-se a realidade hospitalar, percebe-se que nenhum dos fatores deve ser negligenciado na análise, e, por esse motivo, optou-se por incluir, no modelo, restrições aos valores dos pesos, implementadas usando o método da região de confiança. Desta forma, foi possível garantir que todos os *inputs* e *outputs* têm valorizações ajustadas à realidade, bem como diferenciar os níveis de desempenho dos diversos serviços. Para se identificar as restrições a aplicar aos pesos das diferentes variáveis, começou por identificar-se a importância relativa de cada *output* e *input.* Começando pelos *outputs*, analisaram-se os custos unitários das duas linhas de produção consideradas. Os resultados apresentam-se na Tabela 4.

Tabela 4: Custos Unitários por Unidade Produzida nos Serviços de Medicina Interna

Variável de Output	Volume	Custo Unitário Médio
y_2	157.207 altas	2.091,39 €/alta
y_1	1.366.808 dias internamento	240,55 €/dia internamento
y_3	517.972 consultas	155,28 €/consulta

Estes resultados revelam que, na dimensão referente à quantidade total de tratamento produzido, o peso da variável referente ao internamento deverá ser igual a pelo menos 1,5 vezes o peso da variável referente às consultas externas (valor obtido do rácio entre os custos unitários das consultas e dos dias em internamento). Para impedir uma excessiva disparidade na razão entre os pesos destas variáveis, restringiu-se o limite superior desta relação ao valor de 5. A consideração deste limite está relacionada com a manutenção da ordem de grandeza entre os pesos. Isto permite que os pesos tomem valores de magnitude semelhante, sem, no entanto, se restringir excessivamente a capacidade de cada DMU escolher os seus pesos, o que resulta nas primeiras restrições:

$$1,5 \leq \frac{u_1}{u_3} < 5 \tag{7}$$

De forma a não tornar o modelo demasiado restritivo, apenas se restringiu esta razão entre *outputs* de linhas de produção distintas (consultas e internamento). Relativamente à dimensão referente ao acesso aos cuidados, apenas foi possível obter informação para a linha de produção de internamento (ver

Tabela 4, onde falta y_4). Nesta linha, verifica-se que esta dimensão (y_2) apresenta um custo unitário aproximadamente dez vezes superior à dimensão da quantidade de tratamento (y_1). Assim, optou-se por impor limites à relação entre os pesos que refletissem essa magnitude, usando novamente o conceito de ordem de grandeza de um número:

$$5 \leq \frac{u_2}{u_1} < 50 \tag{8}$$

Dada a inexistência de informação quanto à dimensão de acesso na linha de produção de consulta externa (y_4), considerou-se que a relação entre as suas duas dimensões seria semelhante à existente na linha de internamento.

Para identificar a importância de cada *input*, analisou-se a distribuição dos custos que cada uma das variáveis representa em todos os serviços de Medicina Interna. Os resultados apresentam-se na Figura 2, que evidencia que cada uma das cinco variáveis de *input* aqui consideradas tem uma importância muito semelhante nos custos totais do serviço em análise.

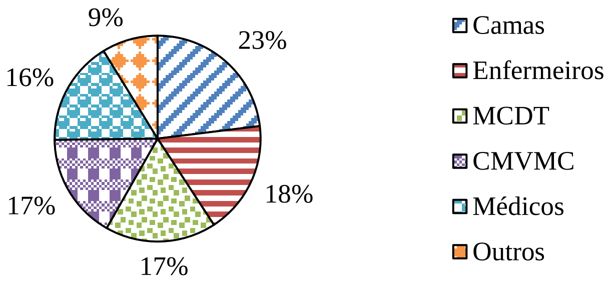


Figura 2: Distribuição dos Custos Gerais (Diretos e Imputados) nos Serviços de Medicina Interna, Representados pelas Variáveis de Interesse.

Para que se reflita nos resultados a importância relativa semelhante de cada *input*, sem se restringir demasiado a liberdade de escolha de cada DMU (como se fez no lado dos *outputs*), optou-se por forçar os hospitais a pesarem todos os *inputs*. Não seria "admissível"que um hospital se tornasse eficiente por pesar apenas alguns *inputs*, negligenciando os restantes. Nesse sentido, limitaram-se as relações entre todos os *inputs* e o *input* "número de horas médicos"(x_1), para que mantivessem a mesma magnitude. De notar que, neste caso, as restrições impostas são entre os pesos virtuais de um hospital fictício, usando *inputs* médios. A opção por restrições aos pesos virtuais deveu-se às diferentes unidades de medida das variáveis. Como se pode ver na equação 9, estas restrições virtuais convertem-se em *Assurance Regions*:

$$0,5 \leq \frac{v_i \overline{x_i}}{v_1 \overline{x_1}} < 5 \Leftrightarrow 0,5 \times \frac{\overline{x_1}}{\overline{x_i}} \leq \frac{v_i}{v_1} < 5 \times \frac{\overline{x_1}}{\overline{x_i}} \tag{9}$$

Exemplificando para o caso da variável x_2, a restrição correspondente fica:

$$0,5 \times 10^{-3} \leq \frac{v_2}{v_1} < 5 \times 10^{-3} \tag{10}$$

5.2 Apresentação e Discussão dos Resultados da Avaliação de Eficiência

Na Figura 3, apresenta-se a distribuição das eficiências obtidas para os 48 hospitais analisados. A média dos resultados obtidos foi de 74,2%, com um desvio padrão de 20,1%. O valor mínimo de eficiência obtido foi de 38,4%. Dos resultados, destaca-se a grande amplitude do intervalo de valores de eficiência, evidenciando uma grande disparidade de desempenhos. Aproximadamente um quinto dos hospitais foi considerado eficiente (10 em 48) e é semelhante o número de hospitais com um nível de eficiência baixo (7, abaixo dos 50%) e alto (9, entre os 80% e 100%). A grande maioria dos hospitais, 22, encontra-se no entanto num nível médio de eficiência (entre 50% e 80%). Esta análise permite concluir que há uma grande margem de melhoria da eficiência dos serviços de Medicina Interna, o que justifica a relevância deste assunto para a gestão hospitalar (para resultados mais detalhados, consultar [Castro(2011)])

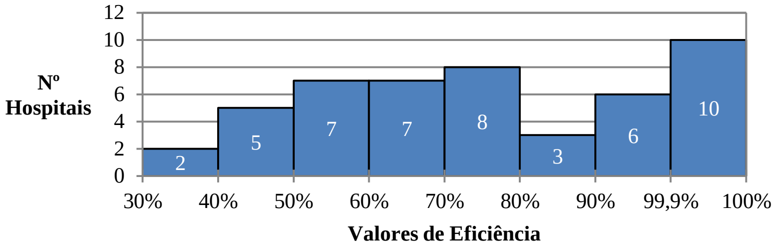


Figura 3: Distribuição dos Valores de Eficiência.

Calculando a razão entre a média dos valores das variáveis de *input* e de *output* das unidades mais ineficientes (eficiência inferior a 80%) e das eficientes, como se apresenta na Figura 4, normalizando para os valores destas últimas, percebemos que os serviços de Medicina Interna mais ineficientes apresentam tempos de internamento e número de consultas externas semelhantes aos serviços eficientes. No entanto, apresentam menos altas de internamento e menor número de primeiras consultas. Foram avaliadas como ineficientes, as unidades que, por comparação, apresentaram níveis de acesso mais baixos, o que valida a necessidade de uma utilização simultânea de duas variáveis, por linha de produção. Ainda assim, as unidades ineficientes apresentam custos bastante superiores aos das unidades eficientes, particularmente em matérias ponderadas, cujos custos são cerca de 1,5 vezes superiores. Os custos com exames ponderados e com enfermeiros também são superiores, mas não se verificam muitas diferenças no número de camas e nas horas de médicos entre os serviços eficientes e ineficientes .

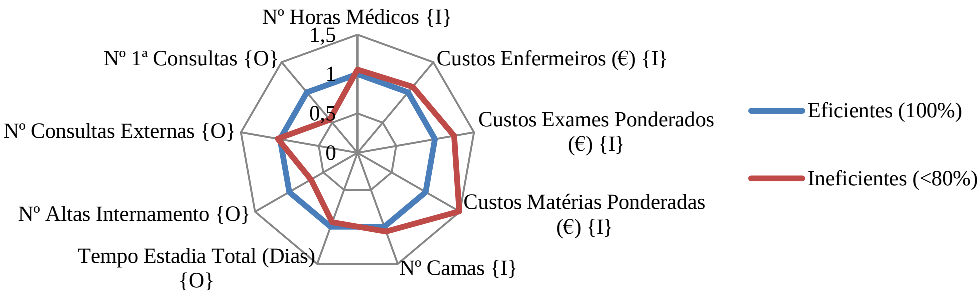


Figura 4: Comparação entre Valores Médios de Unidades Eficientes e Ineficientes.

Nos hospitais avaliados, o total das poupanças potenciais, obtidas por eliminação da totalidade das ineficiências, é apresentado na Figura 5. Este total tem uma magnitude de 100 M€. Pode concluir-se que, tal como verificado anteriormente, as matérias ponderadas, os enfermeiros e os MCDT são as áreas onde maiores poupanças serão possíveis.

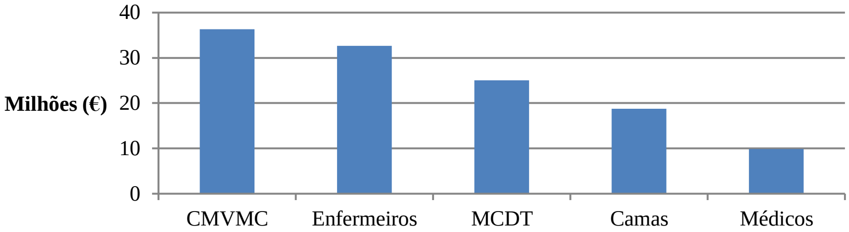


Figura 5: Potencial de Poupanças a Obter, para o SNS, em Medicina Interna.

6 Conclusões e Trabalho Futuro

Neste trabalho idealizou-se um modelo para avaliar os diferentes serviços hospitalares principais, na ótica da eficiência hospitalar. Este modelo teve por base uma revisão bibliográfica que permitiu identificar as variáveis mais comuns e passíveis de ser usadas num modelo de DEA, e um estudo exaustivo da contabilidade analítica dos hospitais do SNS, que permitiu identificar quais as que, efetivamente, são responsáveis pelos gastos mais significativos. Como exemplificado na implementação prática, quando aplicado a cada um dos serviços de um hospital, o modelo tem de ser adaptado mas não deixa de ter uma identidade genérica.

A aplicação do modelo de DEA ao serviço de Medicina Interna, permitiu comparar os níveis de eficiência deste serviço em diferentes hospitais, concluindo-se pela existência de desempenhos muito distintos e por grandes níveis de ineficiência. Verificou-se que os hospitais menos eficientes eram também aqueles que apresentavam os níveis mais baixos de acesso aos cuidados, para esta especialidade. No caso estudado, verificou-se que os maiores ganhos de eficiência podem ser realizados através de reduções nos custos dos medicamentos e material clínico.

Há ainda muito trabalho pela frente para tornar operacional, uma avaliação aos hospitais. Em primeiro lugar, faltará validar o modelo proposto junto dos principais utilizadores, os gestores hospitalares. Posteriormente será necessário idealizar uma forma de agregar os resultados de eficiência de cada serviço numa medida única a atribuir a cada hospital, de forma a permitir uma comparação global entre os hospitais (e percecionar o desempenho dos gestores hospitalares). Tendo algumas secções de apoio, um peso significativo nos custos diretos dos hospitais, será também importante fazer-se a sua avaliação. Para que uma avaliação aos hospitais fique completa, será fundamental aferir-se a eficácia no tratamento dos pacientes, através de modelos de avaliação do tratamento das doenças. Para isto será fundamental o envolvimento de clínicos, dada a sua capacidade única de aferir a capacidade das variáveis do modelo representarem adequadamente o que se pretende medir. Por outro lado, a sua colaboração é essencial para validar e apoiar a aceitação por parte da classe médica dos modelos a propor.

Apesar de todos os constrangimentos, os desenvolvimentos alcançados com este trabalho foram significativos, nomeadamente com a concretização de um modelo de avaliação ao nível dos serviços hospitalares, unidades verdadeiramente comparáveis mas até aqui não abordadas na literatura nacional. A aplicação de DEA permite comparar os hospitais recorrendo a uma medida única, mas tendo por base um conjunto alargado de indicadores, cuja importância individual é definida no interesse de cada avaliado. Espera-se, por isso, que este trabalho contribua para uma mais efetiva avaliação dos hospitais e, para já, para a identificação dos fatores que mais contribuem para a ineficiência hospitalar, em cada serviço.

Referências

[Banker et al.(1984)] Banker, R. D., Charnes, a., Cooper, W. W., 1984. Some Models for Estimating Technical and Scale Inefficiencies in Data Envelopment Analysis. **Management Science** 30 (9), 1078–1092.

[Castro(2011)] Castro, R. A. S., 2011. **Benchmarking de Hospitais Portugueses - Modelação com Data Envelopment Analysis**. Dissertação de Mestrado, Faculdade de Engenharia da Universidade do Porto.

[Chilingerian e Sherman(2004)] Chilingerian, J. A., Sherman, H. D., 2004. Health Care Applications: From Hospitals to Physicians, From Productive Efficiency to Quality Frontiers. Em: Cooper, W. W., Seiford, L. M., Zhu, J. (Eds.), **Handbook on Data Envelopment Analysis**. Kluwer Academic Publisher.

[CIDES-FMUP-UP e ACSS(2011)] CIDES-FMUP-UP, ACSS, 2011 **Índice de Case-Mix (ICM) - Portal da Codificação Clínica e dos GDH**. [último acesso a 8 de Março de 2013]. Disponível em: http://portalcodgdh.min-saude.pt/index.php/Índice_de_Case-Mix_(ICM)

[Costa e Lopes(2007)] Costa, C., Lopes, S., 2007. **Avaliação do Desempenho dos Hospitais Públicos em Portugal Continental - 2005**. Escola Nacional de Saúde Pública - Universidade Nova de Lisboa.

[Dismuke e Sena(2001)] Dismuke, C., Sena, V., 2001. Is there a trade-off between quality and productivity? The case of diagnostic technologies in Portugal. **Annals of Operations Research**, 101–116.

[Donabedian(1980)] Donabedian, A., 1980. The definition of quality and approaches to its assessment. Vol. 1. **Health Administration Press**, Ann Arbor, Michigan.

[Fried et al.(2008)] Fried, H. O., Lovell, C. A. K., Schmidt, S. S., 2008. Efficiency and Productivity. Em: Fried, H. O., Lovell, C. A. K., Schmidt, S. S. (Eds.), **The Measurement of Productive Efficiency and Productivity Growth**. Oxford University Press, New York, Ch. 1.

[Gonçalves(2008)] Gonçalves, L. F. M., 2008. **Análise da eficiência dos hospitais SA e SPA segundo uma abordagem de fronteira de eficiência**. Tese de Doutoramento, ISCTE, Lisboa.

[Hollingsworth(2008)] Hollingsworth, B., 2008. The measurement of efficiency and productivity of health care delivery. **Health Economics** 17 (10), 1107–1128.

[Hollingsworth et al.(1999)] Hollingsworth, B., Dawson, P. J., Maniadakis, N., Jul. 1999. Efficiency measurement of health care: a review of non-parametric methods and applications. **Health care management science** 2 (3), 161–72.

[Kontodimopoulos e Niakas(2005)] Kontodimopoulos, N., Niakas, D., Feb. 2005. Efficiency measurement of hemodialysis units in Greece with data envelopment analysis. **Health Policy** 71 (2), 195–204.

[Moreira(2008)] Moreira, S., 2008. Análise da Eficiência dos Hospitais –Empresa: Uma Aplicação da *Data Envelopment Analysis*. **Banco de Portugal –Boletim Económico Primavera**.

[OCDE(2011)] OECD, 2011. **Health at a Glance 2011: OECD Indicators**. OECD Publishing.

[Portela e Thanassoulis(2006)] Portela, M. C. A. S., Thanassoulis, E., 2006. Zero weights and non-zero slacks: Different solutions to the same problem. **Annals of Operations Research** 145 (1), 129–147.

[Puig-Junoy(1998)] Puig-Junoy, J., 1998. Technical efficiency in the clinical management of critically ill patients. **Health Economics** 277, 263–277.

[Simões(2004)] Simões, J., 2004. **Retrato Político da Saúde - Dependência do Percurso e Inovação em Saúde: da Ideologia ao Desempenho**. Livraria Almedina.

[Thanassoulis et al.(2004)] Thanassoulis, E., Portela, M. C. a., Allen, R., 2004. Incorporating Value Judgements in DEA. Em: Cooper, W. W., Seiford, L. M., Zhu, J. (Eds.), **Handbook on Data Envelopment Analysis**. Kluwer Academic Publishers.

[Thompson et al.(1986)] Thompson, R. G., Singleton Jr., F., Thrall, R. M., Smith, B. A., 1986. Comparative site evaluations for locating a high-energy physics lab in Texas. **Interfaces** (December), 35–49.

[Tribunal de Contas(2006)] Tribunal de Contas, 2006. **Relatório Global de Avaliação do Modelo de Gestão dos Hospitais do SEE –Período de 2001-2004**. Relatório Técnico n.º 20/06 –Audit. I-II.

Routing and assignment of clients of garden maintenance services

J. Orestes Cerdeira*, Manuel Cruz†, Ana Moura‡

* *Departamento de Matemática, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa & Centro de Estudos Florestais, Instituto Superior de Agronomia, Universidade Técnica de Lisboa, Portugal*
E-mail: `orestes@isa.utl.pt`

† *Instituto Superior de Engenharia & Laboratório de Engenharia Matemática, Politécnico do Porto, Portugal*
E-mail: `mbc@isep.ipp.pt`

‡ *Instituto Superior de Engenharia/LEMA, Politécnico do Porto & Centro de Matemática da Universidade do Porto, Portugal*
E-mail: `aim@isep.ipp.pt`

Abstract

We address a routing/assignment problem posed by Neoturf, which is a Portuguese company working in the area of project, building and garden's maintenance. The aim is to define a procedure for scheduling and routing efficiently its clients of garden maintenance services. The company has two teams available throughout the year to handle all the maintenance jobs. Each team consists of two or three employees with a fully-equipped vehicle capable of carrying out every kind of maintenance service. At the beginning of each year, the number and frequency of maintenance interventions to conduct during the year, for each client, are agreed. Time windows are established so that visits to the client should occur only within these periods. There are clients that are supposed to be always served by the same team, but other clients can be served indifferently by any of the two teams. Since clients are geographically spread over a wide region, the total distance travelled while visiting clients is a factor that weighs heavily on the company costs. Neoturf is concerned with reducing these costs, while satisfying agreements with its clients. We give a mixed integer linear programming formulation for the problem, discuss limitations on the size of instances that can be solved to guarantee optimality, present a modification of the Clark and Wright heuristic for the vehicle routing with time windows, and report preliminary computational results obtained with Neoturf data.

Keywords: Routing; Travelling Salesman Problem; Multiple Time Windows.

1 Introduction

In this paper we address a routing/assignment problem posed by Neoturf, which is a Portuguese company working in the area of project, building and garden's maintenance. One of the services provided by Neoturf is the maintenance of private gardens of residential customers (about 60), whose demands are mainly periodic short time interventions (usually 1 to 3 hours). In the beginning of each year, the number and frequency of maintenance interventions to conduct during the year are accorded with each client. Thus, a minimum and maximum periods of time separating two consecutive interventions on the same client are settled. The amount of work highly depends on seasonality. The company allocates to this service two teams (each consists of two or three employees) during the whole year, which may be reinforced with an additional third team during summer. Each team has a van fully equipped with the tools needed to perform the maintenance jobs. There are customers who should be always served by the same team, while others can be served by any team.

Time windows were established so that visits to the client should occur only within these periods. The clients are geographically spread along an area around Oporto of approximately 10 000 km^2. In 2011, these teams traveled more than 60 000 km, with a significant impact on the costs.

Neoturf aims at finding a procedure to scheduling and routing clients efficiently so to reduce costs, while satisfying the agreements with the clients. The scheduling of clients for each day should be planed on a basis of short periods of time (say ten consecutive working days), since unforeseeable events (e.g., weather conditions, client not available at the time previously arranged) may force to postpone planned interventions and to re-settle the designed scheduling.

The routing of customers in each period can be viewed as a variation of the multiple travelling salesman problem with multiple time windows, in which (i) some customers, but not all, are to be visited by a certain vehicle (team); and (ii) no more than one route is assigned on each day to each vehicle (team). We give a mixed integer linear programming formulation model for the problem, discuss limitations on the size of instances that can be solved to guarantee optimality, present a modification of the Clark and Write heuristic for the vehicle routing with time windows [Clarke and Wright, 1964], and report computational results obtained with Neoturf data.

2 Formulation

We consider the year partitioned into consecutive short periods of time (say 10 consecutive working days) and, for each period P of m consecutive working days, we classify clients as

- mandatory, those for which an intervention has to take place during period P, i.e., the number of days since the last visit till the end of period P exceeds the maximum number of consecutive days which can elapse without any intervention taking place, according to what has been agreed with the client;
- discarded, those for which no intervention is expected to take place during period P, i.e., the number of days since the last visit till the end of period P is lower than the number of consecutive days that were agreed to elapse before a new intervention takes place;
- admissible, those for which an intervention may or may not take place during period P.

Let C be the set of clients to be served in period P. We start with C consisting of all mandatory and admissible clients. If no feasible scheduling is found, the decision maker may consider, among other options, to redefine C removing some or all admissible clients from the period P. The routing of clients in C, respecting time windows and minimising travelling and waiting times, is closely related to the *Travelling Salesman Problem with Time Windows* (TSPTW) .

As the travelling salesman problem (TSP) is NP-hard, so is the TSPTW. Savelsbergh [Savelsbergh, 1985] showed that even the problem of deciding whether a feasible solution for the TSPTW exist is NP-complete.

Our problem can be viewed as a TSPTW, where each client can have multiple time windows, and in which certain clients in C have to be visited by (vehicle) team E_0, other clients have to be visited by team E_1, and the remaining clients can be served indifferently either by team E_0 or by team E_1. We denote the sets of those clients by C_0, C_1 and $C_{0,1}$, respectively.

We based our formulation on the so-called *big M formulation* of the TSPTW (model 1 in [Ascheuer et al, 2001]).

We construct a directed weighted graph $G = (V, A, \rho)$ as follows (see Figure 1). The set of vertices V is equal to $C \cup B$, where each vertex b_i^k of B, with $i = 0, \cdots, m$ and $k = 0, 1$, is the i-th "day (fictitious) copy" of the depot for team E_k. There is an arc (u, v) linking client u to client v if there is any possibility to serve v immediately after visiting u, by a same team. Arcs with both directions link each vertex b_i^k, $i = 1, \cdots, m-1$, with every client of $C_k \cup C_{0,1}$, for $k = 0, 1$. There is an arc from b_0^k to every vertex in $C_k \cup C_{0,1}$, for $k = 0, 1$, but there is no arc with head b_0^k. There is an arc from every vertex in $C_k \cup C_{0,1}$ to b_m^k, $k = 0, 1$, but no arc with tail b_m^k. The other arcs in set A are $(b_0^k, b_1^k), (b_1^k, b_2^k), \cdots, (b_{m-1}^k, b_m^k)$, with $k = 0, 1$, and no more arcs exist linking pairs of vertices in B. For $v \in V$, we use V_v^+ and V_v^- to denote the out-neighborhood and in-neighborhood of v, respectively, i.e., $V_v^+ = \{u \in V : (v, u) \in A\}$ and $V_v^- = \{u \in V : (u, v) \in A\}$.

A scheduling of clients assigned to team E_k will be read on graph G as a directed path Q_k from b_0^k to b_m^k. The clients that are to be visited on day i are the vertices of C on the subpath of Q_k linking b_{i-1}^k to b_i^k. The order of vertices on that path specifies the order by which the corresponding clients should be visited. If arc (b_{i-1}^k, b_i^k) is included in path Q_k it means that no interventions on clients of set C will occur on day i for team E_k.

We define the weight ρ_{uv} of every arc $(u, v) \in A$ as the time to travel on arc (u, v), except when $u, v \in B$, where $\rho_{uv} = 0$.

For each vertex $v \in C$, let $T_v^j = [e_v^j, l_v^j]$ be the j-th time-window of client v, $j = 1, \cdots, nT_v$, where nT_v is number of time windows of vertex v, $e_v^j < l_v^j < e_v^{(j+1)}$, and e_v^j and l_v^j are the release time and the deadline time of the j-th time-window of client v, respectively. The release time and deadline time specify minimum and maximum instants for the start of the intervention at the client. For vertices of B, define $T_{b_0^k}^1 = [ST, ST]$ and $T_{b_i^k}^1 = [EN + 24(i-1), EN + 24(i-1)]$, for $i = 1, \cdots, m$ and $k = 0, 1$, where ST and EN are, respectively, the daily service start hour and the daily service end hour.

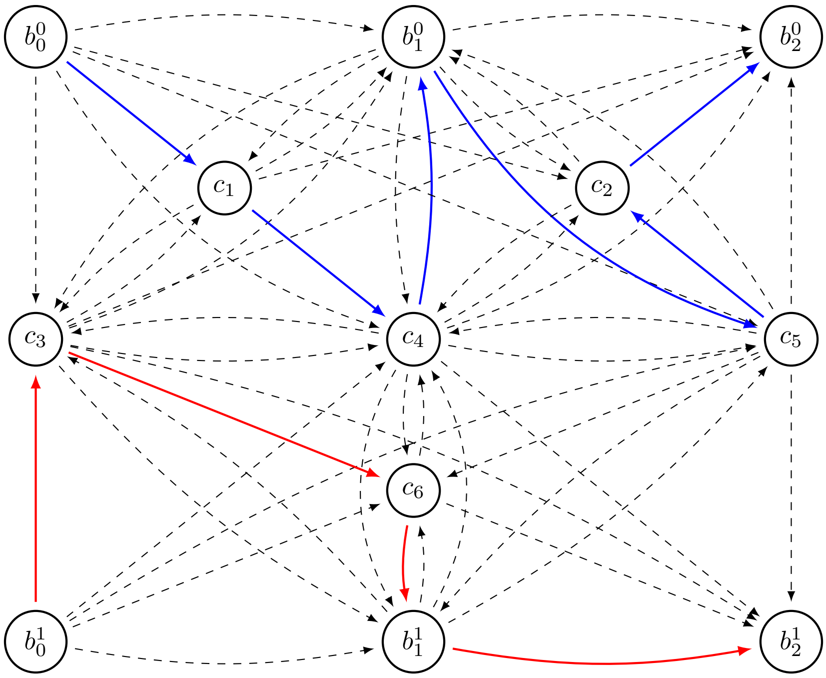


Figure 1: An example of a directed graph G, and a feasible solution for a two days period. Vertices b_0^0, b_1^0, b_2^0 and b_0^1, b_1^1, b_2^1 are the "fictitious copies" of the depot for team E_0 and team E_1, respectively. The subsets of the set of clients $C = \{c_1, \ldots, c_6\}$ are $C_0 = \{c_1, c_2\}$, $C_1 = \{c_6\}$ and $C_{0,1} = \{c_3, c_4, c_5\}$. The scheduling of clients assigned to team E_0 is represented by the directed path consisting of continuous (blue) arcs $Q_0 = (b_0^0, c_1, c_4, b_1^0, c_5, c_2, b_2^0)$. Clients c_1, c_4 and c_5, c_2 are visited by that order on days one and two, respectively. The scheduling of clients assigned to team E_1 is represented by the directed path consisting of continuous (red) arcs $Q_1 = (b_0^1, c_3, c_6, b_1^1, b_2^1)$. Clients c_3, c_6 are visited by that order on day one and no client is visited on day two.

For $v \in C$, let t_v be the processing time on client v, and set $t_{b_0^k} = 0$ and $t_{b_i^k} = ST + 24 - EN$, for $i = 1, \ldots, m$.

To formulate the problem we use the following variables.

- $x_{uv} = \begin{cases} 1, & \text{if arc } (u,v) \in A \text{ is selected,} \\ 0, & \text{otherwise;} \end{cases}$
- $y_v^j = \begin{cases} 1, & \text{if vertex } v \in V \text{ is served in the time-window } T_v^j, \text{ with } j \leq nT_v, \\ 0, & \text{otherwise;} \end{cases}$
- $a_v = \begin{cases} 0, & \text{if vertex } v \in V \text{ is assigned to team } E_0, \\ 1, & \text{if vertex } v \in V \text{ is assigned to team } E_1; \end{cases}$
- $s_v \geq 0$ is the start time, i.e., time-instant in which the service starts at vertex $v \in V$;
- $w_v \geq 0$ is the waiting-time to start the service at client $v \in C$, if the vehicle arrives early than the release time.

We deem minimise the sum of travel-time, waiting-time on clients, and number of working days. We thus have the following objective function.

$$\text{Min} \sum_{(u,v) \in A} (\rho_{uv} + \Delta_{uv}) x_{uv} + \sum_{v \in C} w_v \tag{1}$$

where $\Delta_{uv} = EN - ST$ if $v = b \in B \backslash \{b_0^0, b_0^1\}$ and $\Delta_{uv} = 0$ for the remaining arcs (u,v), to ensure that optimal solutions will have the minimum number of working days (i.e., the maximum number of arcs (b_{i-1}^k, b_i^k)).

The following equations

$$\sum_{u \in V} x_{vu} = 1, \quad \forall v \in V \backslash \{b_m^0, b_m^1\}, \tag{2}$$

$$\sum_{u \in V} x_{uv} = 1, \quad \forall v \in V \backslash \{b_0^0, b_0^1\}, \tag{3}$$

ensure there will be exactly one arc leaving every vertex $v \neq b_m^k$, and exactly one arc entering every vertex $v \neq b_0^k$.

To force that each client is visited exactly in one of its time-windows, we add equations

$$\sum_{j \leq nT_v} y_v^j = 1, \quad \forall v \in V. \tag{4}$$

To guarantee that the start time occurs within the selected time-window and that vehicle has enough time to travel from u to v, we use the following constraints

$$\sum_{j \leq nT_v} e_v^j y_v^j \leq s_v \leq \sum_{j \leq nT_v} l_v^j y_v^j, \quad \forall v \in V, \tag{5}$$

$$s_u + t_u + \rho_{uv} - (1 - x_{uv})M \leq s_v, \quad \forall (u,v) \in A, \tag{6}$$

where $M > 0$ is large enough (say $M = 24m$) to guarantee that the left hand side is non positive whenever $x_{uv} = 0$, and thus making constraint (6) not active when $x_{uv} = 0$.

Note that constraints (2),(3) together with (6), ensure that the set of selected arcs defines a directed path linking b_0^k to b_m^k, for $k = 0, 1$, where every vertex of V is included exactly once in exactly one of the two paths.

The following inequalities define upper bounds on the waiting times on clients.

$$w_v \geq s_v - (s_u + t_u + \rho_{uv}) - (1 - x_{uv})M', \quad \forall (u,v) \in A, v \in C, \tag{7}$$

where $M' > 0$ is large enough (say $M' = 24m$) to guarantee that the right hand side is non positive whenever $x_{uv} = 0$, thus turning the constraint (7) redundant when $x_{uv} = 0$.

The following conditions guarantee that the team assigned to every client v in $C_{0,1}$ is the same team that has visited vertex u, whenever arc (u,v) is in the solution.

$$a_v \leq 1 - x_{uv} + a_u, \quad \forall (u,v) \in A \tag{8}$$

$$a_v \geq x_{uv} - 1 + a_u, \quad \forall (u,v) \in A \tag{9}$$

$$a_v = k, \quad \forall v \in C_k \cup \{b_0^k, b_1^k, \ldots, b_m^k\}, \; k = 0, 1 \tag{10}$$

Indeed, if $x_{uv} = 1$, $a_v = a_u$, and if $x_{uv} = 0$, the inequalities (8) and (9) are redundant.

The range of the variables is established as follows.

$$a_v \in \{0,1\}, \quad \forall v \in C_{0,1} \tag{11}$$

$$x_{uv} \in \{0,1\}, \quad \forall (u,v) \in A \tag{12}$$

$$y_v^j \in \{0,1\}, \quad \forall v \in V, \text{ and } j \leq nT_v \tag{13}$$

$$s_v \geq 0, \quad \forall v \in V \tag{14}$$

$$w_v \geq 0, \quad \forall v \in C \tag{15}$$

The above model (1)-(15) gives a mixed integer linear programming formulation for the problem of routing clients of C on a given period of m days, by two teams. The objective function (1) was defined to minimise travel-time and waiting-time on clients in the minimal number of days. Other alternative goals could be considered. For instance, minimising the total completion-time, i.e., the time of the last service on period P. This could be achieved introducing variable F, imposing the constraints $F \geq s_v + t_v, \forall v \in C$, and defining as objective function: $\min F$. This would give solutions with a minimum number of consecutive working days, and leaving the non working days, if any, at the end of period P. Solutions that define a sequence of consecutive non working-days finishing at the end of period P permit to anticipate the next period. However, the objective function (1) expresses the goals specified by Neoturf. The existence of intermittent non working days is not a issue for Neoturf, as it permits to assign the members of the team to other activities.

3 Methods of resolution

We used the NEOS Server [NEOS Server, 2013] platform to test the model (1)-(15). The implementation was made in AMPL [AMPL, 2013] modelling language and ran using the commercial solver Gurobi.

On the tests that we carried out, only for periods not exceeding five days, Gurobi produced the optimal solutions. For different instances consisting of periods of more than five consecutive working days NEOS Server returned either "timeout" or "out of memory". No improvements were achieved when considering different parameterizations on *threads*, *mipgap* or *timelimit*.

Given these limitations, and despite the relatively small number of days of the periods that are to be considered for planning, we decided to waive from optimality guaranteed, and use an implementation of Clarke and Wright (C&W) [Clarke and Wright, 1964] heuristic for the vehicle routing problem with multiple time windows (VRPMTW) available in MATLAB [MATLAB, 2013] .

There are two main issues in applying C&W heuristic to our problem. First, C&W algorithm does not distinguish between clients from C_0, C_1 and $C_{0,1}$. Thus, solutions may include in the same routes clients from C_0 together with clients from C_1.

The second issue follows from the assumption behind C&W algorithm that there are enough vehicles available for the routes determined by the algorithm. Thus, the same team may be assigned, on the same day, to more than one route with incompatible time windows (i.e., services to clients in different routes overlap in time).

To handle the first issue we proceeded as follows.

- We duplicated the number m of days of period P.
- For all clients in C_1, we added $24 \times m$ hours to the release and deadline times of every time window.
- For all clients in $C_{0,1}$ we duplicated the number of time windows and, beside the original ones, we also added $24 \times m$ hours to the release and deadline times of every original time window.

Since each client is visited exactly once, in the whole period (now with $2m$ days), within one of its time windows, setting the time windows of clients C_0 on the first m days and the time windows of clients C_1 on days $m + 1$ to $2m$, ensures that clients from C_0 will not be put together in the same routes with clients from C_1.

Duplicating as described above the number of time windows of clients $C_{0,1}$, and given that each will be served exactly once, defines a partition of these clients into those that will be served in the first m days (together with clients of C_0), and those that will be served in days $m + 1$ to $2m$ (together with clients of $C_{0,1}$).

To address the second issue we use matchings in bipartite graphs.

Suppose the number of routes assigned to a team is less than or equal to m, and there is more than one route on the same day. We consider a bipartite graph where vertices of bi-class R represent routes and vertices of the other bi-class D represent the m days. There is an edge $[r, d]$, with $r \in R$ and $d \in D$, if and only if route r can be done (w.r.t time windows) in day d. We then find the maximum matching [Hopcroft and Karp, 1973] of this graph. If it has $|R|$ edges, then it indicates how routes should be distributed by the m days of the period, with no more than one route per day. If the maximum matching has less than $|R|$ edges, or $|R| > |D|$, we propose that the decision maker considers: assigning an extra-team for this period; increasing the number of days in the forecast period, or reducing the number of admissible clients for the period, and repeat the whole process.

4 Computational results and conclusion

Here we report some computational experiments carried out with Neoturf data. We call total time to the sum of travel-time and waiting-times, i.e., the values of the objective function not accounting for parameters Δ.

As previously referred, we only succeeded to obtain solutions using Gurobi for very small periods of days. For these small instances the gap of total routing times of the solutions obtained with C&W heuristic w.r.t. the optimal values (i.e., $(T(C\&W) - OPT)/OPT$, where $T(C\&W)$ and OPT, are the total time of the solution obtained with C&W heuristic and the optimal total time, respectively) did not exceed 5%.

We then compared the planning that Neoturf had established for a 14 days period (18-Feb-2013 till 3-Mar-2013) with the one produced with C&W heuristic. The solution produced with C&W has an total time of 8h54m (waiting-time = 0, and 105h24m if working-time is also considered) to serve the 27 clients in 7 and 9 working days for teams E_0 and E_1, respectively. The planning of Neoturf consisted of 14h02m total-time (waiting-time=1h00, and 110h32m considering working-time), 8 days for team E_0 and 11 days for team E_1.

This gives a reduction on total-time of about 37%(= 100 × (14h02m - 8h54m)/14h02m), that significantly decreases costs resulting from distances travelled, specially because the two teams travel around 60 000 km/year.

Neoturf was very pleased with the solutions obtained. Yet we believe that results may be improved using heuristics for routing more sophisticated than C&W, and exploring models alternative to (1)-(15). We intend to pursuit on this direction.

Acknowledgments

The authors are grateful to two referees for several suggestions on a previous version of the manuscript.

The problem addressed in this paper was presented by Neoturf during the 86th European Study Group with Industry, held at ISEP/IPP, School of Engineering, Polytechnic of Porto, 7 - 11 May 2012. The authors are also grateful to Neoturf for all its availability, by answering all questions and providing all the necessary data to perform this work.

The three authors were supported by the Portuguese Foundation for Science and Technology (FCT). J. O. Cerdeira was funded through the project PEst-OE/AGR/UI0239/2011 and the *Centro de Estudos Florestais* under FEDER/POCI. M. Cruz was supported by *Laboratório de Engenharia Matemática.* A. Moura was funded through the project PEst-C/MAT/UI0144/2011 and the *Centro de Matemática da Universidade do Porto*, under POCTI/POSI, with Portuguese and European Community structural funds.

References

[AMPL, 2013] AMPL modelling language, March 2013 Available online at: http://www.ampl.com/.

[Ascheuer et al, 2001] Ascheuer, N., Fischetti, M., and Grötschel, M. (2001) Solving the asymmetric travelling salesman problem with time windows by branch-and-cut. *Mathematical Programming*, Ser. A 90: 475-506.

[Baker, 1983] Baker, E. K. (1983) An exact algorithm for the time-constrained travelling salesman problem. *Operations Research*, 31(5): 938-945.

[Clarke and Wright, 1964] Clarke, G. and Wright, J. W. (1964) Scheduling of Vehicles from a Central Depot to a Number of Delivery Points. *Operations Research*, 12(4): 568-581.

[Hopcroft and Karp, 1973] Hopcroft, J. E., Karp, R. M. (1973) An $n^{5/2}$ algorithm for maximum matchings in bipartite graphs. *SIAM Journal on Computing*, 2(4): 225-231.

[MATLAB, 2013] MATLAB, The MathWorks, Inc., Natick, Massachusetts, United States. http://www.mathworks.com/products/matlab/

[NEOS Server, 2013] NEOS Server, March 2013 Available online at: http://www.neos-server.org/.

[Savelsbergh, 1985] Savelsbergh, M. W. P. (1985) Local search in routing problems with time windows. *Annals of Operations Research* 4(1): 285-305.

Discrete lot sizing and scheduling on parallel machines: description of a column generation approach

António J.S.T. Duarte[†], J.M.V. Valério de Carvalho[‡]

[†] *Escola Superior de Tecnologia e Gestão, Instituto Politécnico de Bragança, Portugal*
E-mail: `aduarte@ipb.pt`

[‡] *Departamento de Produção e Sistemas, Universidade do Minho, Portugal*
E-mail: `vc@dps.uminho.pt`

Abstract

In this work, we study the discrete lot sizing and scheduling problem (DSLP) in identical parallel resources with (sequence-independent) setup costs and inventory holding costs. We propose a Dantzig-Wolfe decomposition of a known formulation and describe a branch-and-price and column generation procedure to solve the problem to optimality. Preliminary results show that the lower bounds provided by the reformulated model are stronger than the lower bounds provided by the linear programming (LP) relaxation of the original model.

Keywords: DLSP, lot sizing, scheduling, setup costs, column generation, branch-and-price.

1 Introduction

Since the introductory work of Wagner and Whitin [Wagner and Whitin, 1958] a great amount of research has been done on the discrete lot sizing and scheduling problem (DLSP). The original model has been extended from single-item to multiple-item and from single resource to multiple-resource configurations. Also, additional constraints and different cost structures have been studied. Other studies aim at proposing and/or strengthening compact mixed integer linear (MILP) formulations in order to solve larger and more complex instances. Examples of relevant research works on this problem are [van Hoesel et al., 1991], [van Hoesel and Kolen, 1994], [van Eijl and van Hoesel, 1997], [Gicquel et al., 2011] and [Gicquel et al., 2012].

Most of the published research for problems with parallel resources is devoted to heuristics.

In this work we propose a Dantzig-Wolfe decomposition to a common integer linear (ILP) formulation and a branch-and-price algorithm to solve the problem to optimality. For the single resource problem a similar column generation approach is presented in [Cattrysse et al., 1993].

For the parallel resource configurations the authors are not aware of similar approaches, although the used decomposition is very close the one used in [Manne, 1958] and in [Lasdon and Terjung, 1971]. However, on those works, the problem of finding the optimal integer solution was not addressed. Also, the problem does have some similarities with the capacitated lot sizing and scheduling problem for which there is also some published research involving column generation, such as [Degraeve and Jans, 2007] and [Caserta and Voß¸, 2013].

A relatively recent review of methods for this problem can be found in [Karimi et al., 2003].

In section 2 we provide a formal description of the problem. In section 3 we present a compact original ILP formulation. In section 4 we present a minimum cost flow model that can be used to readily compute upper bounds. In section 5 a Dantzig-Wolfe decomposition for the ILP formulation is proposed along with the resulting master problem and subproblem. In section 6 a dynamic programming approach to the resulting subproblem is presented. Three different branching schemes to solve the problem to optimality are presented in section 7. Finally we present some results showing that the lower bounds provided by the reformulated model are stronger than the lower bounds provided by the linear programming relaxation of the original model.

2 Problem description

There are R identical parallel resources, indexed with $r = 1, \dots, R$, I items to be processed, indexed with $i = 1, \dots, I$, and T discrete and equal periods of time, indexed with $t = 1, \dots, T$. In each time period, any given machine will be producing one *demand unit* of a given item or will be idle.

Without loss of generality, we define the *demand unit* for a given item as the quantity of that item that is possible to process in one machine during one time period. In practice, this can be seen as a minimum lot size for each item. From this point on, demands will be expressed in integer demand units.

Each item has the following associated coefficients: a vector of demands along the planning horizon, $d_i = \{d_{i1}, \dots, d_{iT}\}$; a startup cost, s_i, which is the cost of starting the production of a different item in a given resource, which is resource and time independent; an inventory holding cost, h_i, defined as the cost of holding one demand unit of item i over one time period (time independent).

The objective is to decide a production schedule (assigning machines to items over the different time periods) that minimizes the sum of startup and holding costs while meeting the required demands (back-orders are not allowed).

3 ILP formulation

Because the resources are identical, in our formulation, we use the aggregate variables, as defined in [Gicquel et al., 2012]. The complete set of variables is:

x_{it} : number of resources producing item i on period t. Variables x_{i0} are defined in order to account for the number of startups in period 1 and should be made equal to a value that reflects the state of the various resources at the start of period 1;

y_{it} : number of resources where production of item i is started on period t and a startup cost is incurred;

z_{it} : number of demand units of item i carried as inventory from period t to period $t+1$. Variables z_{i0} are defined and should be fixed to reflect the inventory level at the start of period 1.

The complete ILP formulation is the following:

$$\min \quad \sum_{i=1}^{I}\sum_{t=1}^{T}(s_i y_{it} + h_i z_{it}) \tag{1}$$

$$\text{s.t.} \quad z_{i(t-1)} + x_{it} = d_{it} + z_{it} \qquad i = \{1, \dots, I\},\ t = \{1, \dots, T\} \tag{2}$$

$$y_{it} \geq x_{it} - x_{i(t-1)} \qquad i = \{1, \dots, I\},\ t = \{1, \dots, T\} \tag{3}$$

$$\sum_{i=1}^{I} x_{it} \leq R \qquad t = \{1, \dots, T\} \tag{4}$$

$$x_{it} \geq 0 \text{ and integer} \qquad i = \{1, \dots, I\},\ t = \{1, \dots, T\} \tag{5}$$

$$y_{it} \geq 0 \text{ and integer} \qquad i = \{1, \dots, I\},\ t = \{1, \dots, T\} \tag{6}$$

$$z_{it} \geq 0 \text{ and integer} \qquad i = \{1, \dots, I\},\ t = \{1, \dots, T\} \tag{7}$$

Note that x_{i0} and z_{i0} are actually constants that reflect the initial state of the resources and the initial inventory levels. From this point on, for simplicity and without loss of generality we will assume these constants to be 0.

The objective function (1) sums the startup costs and the holding inventory costs. Constraints (2) express the inventory balance at each period. Constraints (3) ensure that a startup cost is incurred whenever the number of resources used for a given item increases. Finally, constraints (4) limit the number of resources used in each time period, and constraints (5), (6) and (7) specify the type and limits of the variables.

Using a similar formulation and a standard optimization package on a personal computer, [Gicquel et al., 2012] reported that they could not solve instances with $I = 10$, $R = 2$ and $T = 50$ within 30 minutes of computation. It is clear that solving this formulation directly is not practical, even for small instances.

4 Minimum cost flow formulation

When performing branch-and-bound it is important to be able to compute upper bounds. In this section we propose a minimum cost flow formulation for the DLSP. The formulation is incomplete in the sense that inventory costs are accounted but not the startup costs, which means that the optimal solutions of the network flow problem, when they exist, are feasible to the DLSP, but not guaranteed to be optimal. A similar network for single item problems appears on [Zangwill, 1968].

Consider the following acyclic directed network. There is one supply node, S, whose supply is equal to RT. Consider also a set of T transshipment nodes, one for each time period, named $T_1, \dots, T_T$. There are arcs from S to T_t with cost 0 and capacity equal to R.

Each of the T_t nodes will be connected to I demand nodes named $D_{1t}, \dots, D_{It}$. The demand on the D_{it} nodes will be equal to d_{it} and the arcs from T_t to D_{it} have a cost of 0 and unlimited capacity (in practice, the limit will be R). The flow on these arcs has the same meaning as variables x_{it} of the ILP formulation.

Another set of directed arcs will depart from each D_{it} node to the node $D_{i(t+1)}$. These arcs have a cost equal to h_i and unlimited capacity. The flow on these arcs has the same meaning as variables z_{it} of the ILP formulation.

Finally, in order to balance the supply and the demand, consider an additional demand node, D_{idle}, whose demand, d_{idle}, is computed as[1]

$$d_{idle} = RT - \sum_{i=1}^{I} \sum_{t=1}^{T} d_{it}$$

Finally, an arc with cost equal to zero and unlimited capacity, should connect S and D_{idle}. The flow on this arc represents the global capacity excess on the resources.

The complete network is represented on Figure 1. Note that z_{i0} and z_{iT} can be used to account for, respectively, initial and final inventory levels, if there is need for them to be non-zero.

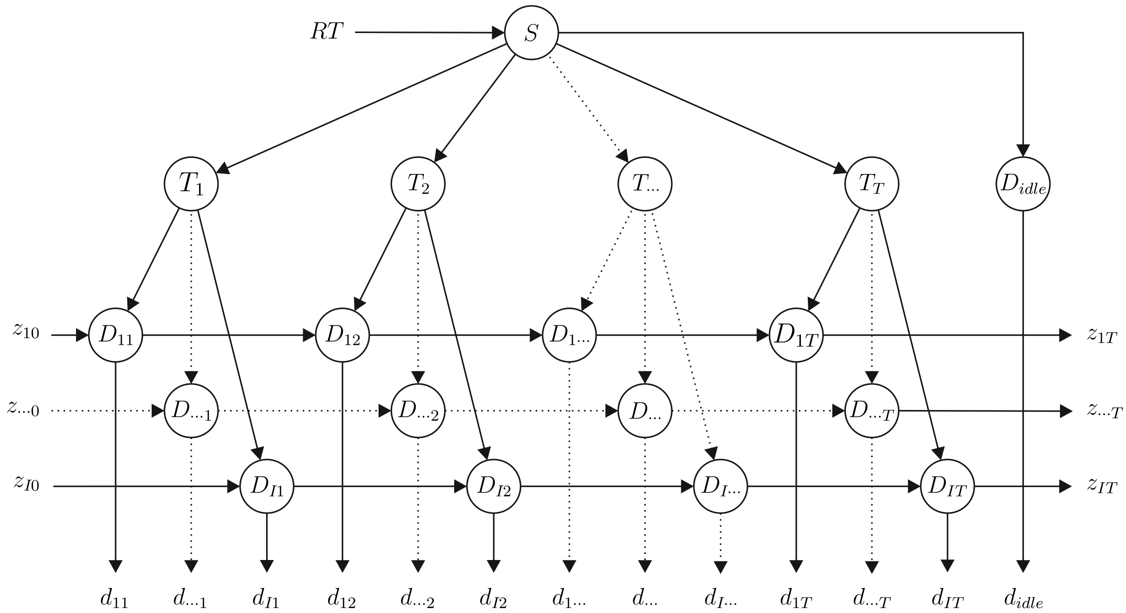


Figure 1: Minimum cost flow network representation.

Because the flow in arcs (T_t, D_{it}) has the same meaning as variables x_{it} of the ILP formulation, this network can be used to compute feasible solution to the DLSP that can be used as upper bounds, taking advantage of fast and widely available state-of-the-art minimum cost flow algorithms.

5 Dantzig-Wolfe decomposition

In this section we apply and present a standard Dantzig-Wolfe decomposition to the ILP formulation presented in section 3.

[1] Note that, if d_{idle} is negative, the problem is infeasible due to a global lack of resource capacity. If d_{idle} is non-negative, the problem can still be infeasible due to demand imbalances over time. A trivial way to check feasibility is to use the same principle to compute the idle capacity at every time period t', *i.e.*, $d_{idle}^{t'} = Rt' - \sum_{i=1}^{I} \sum_{t=1}^{t'} d_{it}$.

The ILP formulation has a block angular structure. With the exception of (4), which are coupling constraints, all other constraints can be grouped into I blocks, one for each product item. In our decomposition we will leave constraints (4) in the master problem and group all the constraints that refer to item i to a polyhedron named P_i.

Because any polyhedron P_i is a convex region, any point belonging to P_i can be represented as a convex combination of extreme points. Let p_{ik} be such points. For any P_i polyhedron there will be K_i extreme points, so that $k = 1, \ldots, K_i$. Let $\lambda_{ik} \geq 0$ be the weight of each extreme point in a given combination such that, for any given i, $\sum_{k=1}^{K_i} \lambda_{ik} = 1$. After variable substitution, the master problem will be:

$$\min \quad \sum_{i=1}^{I} \sum_{k=1}^{K_i} c_{ik} \lambda_{ik} \tag{8}$$

$$\text{s.t.} \quad \sum_{i=1}^{I} \sum_{k=1}^{K_i} a_{ikt} \lambda_{ik} \leq R \qquad t = \{1, \ldots, T\} \tag{9}$$

$$\sum_{k=1}^{K_i} \lambda_{ik} = 1 \qquad i = \{1, \ldots, I\} \tag{10}$$

$$\lambda_{ik} \geq 0 \text{ and integer} \qquad i = \{1, \ldots, I\},\ k = \{1, \ldots, K_i\} \tag{11}$$

In this reformulated model, columns can be interpreted as potential schedules for a single item, i, where c_{ik} is the cost of the schedule (including startup and inventory holding costs) and a_{ikt} is number of resources used by the schedule in period t.

Because it is not practical to enumerate all the potential single item schedules, they have to be dynamically generated. Based on the dual solution of the master problem, the subproblems will generate valid and cost attractive schedules to be included in the solution of the master problem.

Each P_i polyhedron will give origin to a different subproblem. Let π_t and ν_i be the dual variables associated with constraints (9) and (10), respectively. Subproblem i will have the following formulation:

$$\min \quad \sum_{t=1}^{T} (s_i y_{it} + h_i z_{it} - \pi_t x_{it}) - \nu_i \tag{12}$$

$$\text{s.t.} \quad z_{i(t-1)} + x_{it} = d_{it} + z_{it} \qquad t = \{1, \ldots, T\} \tag{13}$$

$$y_{it} \geq x_{it} - x_{i(t-1)} \qquad t = \{1, \ldots, T\} \tag{14}$$

$$0 \leq x_{it} \leq R \text{ and integer} \qquad t = \{1, \ldots, T\} \tag{15}$$

$$y_{it} \geq 0 \text{ and integer} \qquad t = \{1, \ldots, T\} \tag{16}$$

$$z_{it} \geq 0 \text{ and integer} \qquad t = \{1, \ldots, T\} \tag{17}$$

The subproblem is a single item DLSP on parallel resources. Note that the bounds on x_{it} in constraints (15) are included to avoid the generation of invalid schedules that will never be part of an optimal integer solution to the master problem.

After optimization, for a new column, $c_{ik} = \sum_{t=1}^{T} (s_i y_{it} + h_i z_{it})$ and, hence, the subproblem optimal objective function value is the reduced cost of that column. A generated column is added to the master problem, only if its reduced cost is negative. Also, coefficients a_{ikt} of the new column are equal to x_{it}.

Clearly, if the solution of the reformulated model has only integer variables, then an integer solution to DLSP can be computed. Nevertheless, one relevant characteristic of this problem is that an integer solution to DLSP can also be computed from non-integer variables of the reformulated model, whenever the solution of the reformulated model corresponds to an integer solution in the space of the original variables. This is fully exploited in the branch-and-price algorithm, because the solution in the space of the original variables has to be computed to derive the branching constraints; the branching scheme is presented in section 7.

The following proposition defines the set of conditions that a solution to the master problem must possess in order to be an integer solution to the DLSP:

Proposition 1. *For a solution to the DLSP problem to be integer, it is sufficient that all λ_{ik} variables are integer or that all x_{it} variables are integer, with*

$$x_{it} = \sum_{k=1}^{K_i} a_{ikt} \lambda_{ik} \tag{18}$$

Proof. This proposition can be easily demonstrated because λ_{ik} are binary variables that represent a single item schedule among all the resources, and, if they are all integer, they represent a valid solution. Variables x_{it} are the original formulation variables that represent the number of resources used by item i in time period t. Thus, if all x_{it} are integer, they represent a valid solution. □

Consider a new free decision variable, y'_{it} defined as $y'_{it} = x_{it} - x_{i(t-1)}$. This decision variable represents the change in the number of resources producing item i from period $t-1$ to period t. If there is an increase in the number of resources used, y'_{it} will be positive (equal to the formerly defined y_{it}) and, if there is a decrease, it will be negative. Given this definition, the following proposition is also true:

Proposition 2. *Given the sets of variables x_{it}, y'_{it} and z_{it}, if one of those sets is integer, then, the others must also be integer.*

Proof. Variables y'_{it} represent the variation in the number of used resources for a given item and can be computed from x_{it} as stated above. Hence if one of the sets is integer the other is also integer. Variables z_{it} are inventory levels and so $z_{it} = z_{i(t-1)} + x_{it} - d_{it}$. Because d_{it} are integer values, the previous reasoning still applies. □

6 Subproblem optimization

In this section we present a dynamic programming algorithm to solve the subproblem, a single item DLSP. The algorithm evaluates function $F_t(z, r)$ that represents the minimum cost to get z inventory level at the end of period t with r resources setup for the production of the considered item. If we assume that all resources are idle at instant 0, and the initial inventory is 0, then, $F_0(0, 0) = 0$. At each stage transition, we must decide how many resources will be allocated to the production of the considered item, i. Let $x_{it} = \{0, \ldots, R\}$ be that value. Then, from state (z, r) at stage $t-1$ we can reach, at stage t, states $(z' = z - d_{it} + x_{it}, r' = x_{it})$ as long as $z' \geq 0$, because inventory can not be negative. The objective function will be computed in the following way:

$$F_t(z', r') = \begin{cases} F_{t-1}(z, r) - \pi_t r' + h_i z' + s_i(r' - r) & \text{if } r' > r \\ F_{t-1}(z, r) - \pi_t r' + h_i z' & \text{if } r' \leq r \end{cases} \tag{19}$$

At each stage, the maximum theoretical number of states will be equal to $(R+1)(z_t^+ - z_t^- + 1)$, where z_t^- and z_t^+ are bounds on the inventory level at the end of period t and can be computed as follows:

$$z_t^- = \max\left(0, d_{i(t+1)} - R + z_{t+1}^-\right) \tag{20}$$

$$z_t^+ = \min\left(\sum_{l=1}^{t}(R - d_{il}), \sum_{l=t+1}^{T} d_{il}\right) \tag{21}$$

In equation (20) computation is recursive and should be initialized with $z_T^- = 0$, stating that the minimum inventory at the end of period T should be 0 (see discussion on section 4). The computations reflect the fact that, when the demand exceeds R, there will be need for inventory at the end of the previous period or periods.

Concerning the equation (21), the maximum inventory is the minimum value between the achievable inventory at the end of period t using maximum capacity and the maximum inventory needs to satisfy demand from inventory for the rest of the planning horizon (once again, assuming that the final inventory should be 0).

Note that these bounds can be used to improve (7) in the ILP formulation and (17) in the subproblem formulation and can be easily modified in the presence of initial and final inventories.

The above mentioned number of states is the theoretical maximum because if, for some state, $F_t(z, r)$ equals or exceeds ν_i, further transitions from that state can be ignored, because the reduced cost of the new column would not be negative and, hence, the column would not be attractive.

7 Branching

Solving the relaxed master problem to optimality does not guarantee an integer solution. For that reason, in order to find an integer optimal solution it is necessary to identify and eliminate fractional solutions. Branching is a standard procedure to achieve that goal.

As it is widely known, when performing column generation, branching on the master problem variables (λ_{ik}) is not a good idea, because it leads to column regeneration whenever a branching decision of the type $\lambda_{ik} \leq 0$ is made.

Given proposition 2, presented in section 5, the sets x_{it}, y'_{it} and z_{it} are natural candidates for branching. The choice should be made based on the results of computational performance tests.

Note that the original variables y_{it} cannot be used for branching because, although integrality on x_{it} implies integrality on y_{it}, the converse is not true. For example, consider the number of resources used (a_{ikt} vectors) in two four-period schedules for a given item: (0,4,4,4) and (4,4,3,1). Suppose that, in the optimal solution of a given node, both λ_{ik} are at a level of 0.5. As it can be easily seen, $x_{ik} = (2, 4, 3.5, 2.5)$ while $y_{ik} = (2, 2, 0, 0)$. This solution would be fractional, while the y_{it} vector would be integer. In this case, the vector y'_{it} would be (2,2,-0.5,-1) and, hence, not integer.

The following subsections present the 3 possible branching schemes along with the adjustments to the subproblem structure.

7.1 Branching on x_{it}

When branching upon the x_{it} variables, in node j, two branches of the problem are created. On one branch (the left branch) the constraint

$$x_{it} \leq \lfloor x_j^* \rfloor \tag{22}$$

is added, where x_j^* represents some non-integer value. On the other branch (the right branch) the following constraint is added instead:

$$x_{it} \geq \lceil x_j^* \rceil \tag{23}$$

With respect to finding the optimal solution of the model at a given node j, it is necessary to call the subproblems for attractive columns not yet included in the master problem. In node j, besides the initial constraints, the master problem has other sets of constraints, denoted as P_{it}^j, with $i = 1, \ldots, I$ and $t = 1, \ldots, T$, resulting from all the branching decisions imposed on each different variable x_{it}.

Let $\rho_{it,j}^p$ be the dual variable associated with constraint p, with $p \in P_{it}^j$. Thus, in order for the subproblem to correctly identify the attractive columns, in the objective function (12) and in the recursive equation (19), π_t must be replaced with $(\pi_t + \rho_{it}^j)$, where ρ_{it}^j is the sum of all dual variables, $\rho_{it,j}^p$, associated with constraints $p \in P_{it}^j$, which are imposed on the variable x_{it} at node j, *i.e.*, $\rho_{it}^j = \sum_{p \in P_{it}^j} \rho_{it,j}^p$.

7.2 Branching on z_{it}

Branching on the z_{it} variables requires some additional manipulations. Developing $z_{it} = z_{i(t-1)} + x_{it} - d_{it}$ recursively yields the following (assuming the starting inventory is 0):

$$z_{it} = \sum_{l=1}^{t} (x_{il} - d_{il}) \tag{24}$$

To translate z_{it} to the master space, once again, equation (18) should be used. Using the same approach as before, on node j we want to branch on variable z_{it}, whose fractional value is z_j^*. The left and right branching constraints will be, respectively:

$$z_{it} \leq \lfloor z_j^* \rfloor \tag{25}$$

$$z_{it} \geq \lceil z_j^* \rceil \tag{26}$$

Using the same notation as in section 7.1, if ρ_{it}^j is the sum of the dual variables that refer to constraints imposed on the variable z_{it}, the modification to objective function (12) and to the recursive equation (19) is the replacement of h_i by $(h_i - \rho_{it}^j)$.

7.3 Branching on y'_{it}

Let ${y'_j}^*$ be the fractional value of y'_{it} that we wish to branch upon on node j. The constraints to impose on the left and right branches are, respectively,

$$y'_{it} \leq \lfloor {y'_j}^* \rfloor \tag{27}$$

$$y'_{it} \geq \lceil {y'_j}^* \rceil \tag{28}$$

On these equations, y'_{it} can be replaced with $x_{it} - x_{i(t-1)}$ and projected to the master problem space using equation (18). Once again, as in the previous sections, let ρ^j_{it} be the sum of the dual variables whose associated constraints refer to variable y'_{it}.

In this case, the modifications to the subproblem structure are more complex than in the previous branching schemes presented on sections 7.1 and 7.2.

In the case of the ILP formulation there is the need of creating a set of variables to account for decreases in the number of used resources. Let's name those variables y^-_{it}. In the objective function (12) a new term associated with this new variables must be included rendering the following objective function:

$$\sum_{t=1}^{T} \left((s_i - \rho^j_{it}) y_{it} + h_i z_{it} - \pi_t x_{it} + \rho^j_{it} y^-_{it} \right) - \nu_i \tag{29}$$

Also, an additional set of constraints must be included (similar to constraints (14)):

$$y^-_{it} \geq x_{i(t-1)} - x_{it} \qquad t = \{1, \ldots, T\} \tag{30}$$

Also, in the subproblem formulation that resulted from the decomposition, the y_{it} variables have no upper bound because it is implicitly assumed that their coefficients on the objective function are always positive. Because this last assumption is no longer true, an upper bound on y_{it} equal to max $\left(0,\ x_{it} - x_{i(t-1)}\right)$ must be enforced in the ILP subproblem formulation. The same logic applies to the y^-_{it} variables: an upper bound equal to max $\left(0,\ x_{i(t-1)} - x_{it}\right)$ must be enforced. For simplicity, the necessary additional constraints are omitted here.

The recursive equation (19) needs also to be modified and, after the necessary modifications, it will be:

$$F_t(z', r') = \begin{cases} F_{t-1}(z, r) - \pi_t r' + h_i z' + (s_i - \rho^j_{it})(r' - r) & \text{if } r' > r \\ F_{t-1}(z, r) - \pi_t r' + h_i z' + \rho^j_{it}(r - r') & \text{if } r' \leq r \end{cases} \tag{31}$$

With this changes, the subproblem will correctly process the additional dual information.

8 Computational results

To assess the quality of the bounds, a total of 10 instances was solved. The first 3 instances are very small and were used for debugging purposes. The other 7 instances are also small, with only 4 items, 4 resources and 5 time periods (10 periods for the last two), and were generated using the procedure described in [Gicquel et al., 2012].

The results are shown in table 1 and the columns have the following meaning: columns R, I and T are, respectively, the number of resources, the number of items and the number of periods; column *UC* is the used capacity, in percentage; column *ILPRel* is the optimal value of the LP relaxation of the ILP formulation presented in section 3; column *Network* is the optimal value for the network flow model presented in section 4; column *MRoot* is the optimal value for the master problem of the reformulated model at the root node, with the asterisk denoting integer solutions where no branching was necessary; the last column is the value of the integer optimal solution, after branching as needed.

Table 1: Preliminary computational results.

Instance	R	I	T	UC	ILPRel	Network	MRoot	Optimal
1	3	3	4	100	21.25	28	25*	25
2	4	3	4	100	80.00	92	84	84
3	4	3	4	81.25	55.25	74	57	57
4	4	4	5	95	1426.33	2454	1534*	1534
5	4	4	5	90	1118.25	3219	1350	1350
6	4	4	5	85	667.33	2005	753*	753
7	4	4	5	80	990.67	1688	1105*	1105
8	4	4	5	75	1221.33	2088	1412*	1412
9	5	5	10	100	2813.17	5411	3282.5	3368
10	5	5	10	88	1574.32	4688	1673*	1673

Although the tested instances were very small it can be seen that the reformulated model is stronger than the original formulation, as the lower bound given by its LP relaxation is better than the one given by the LP relaxation of the original model. With the exception of instance 9, this bound is equal to the optimal solution value.

Concerning the network flow model, it can be seen that the provided upper bound can be arbitrarily bad. Because the model only considers inventory holding costs, when this costs are small compared to the startup costs (as it happens in instances 4 to 10)[2] the bound is very high.

As these preliminary results were encouraging, an implementation was developed in C# (Microsoft .NET framework 4.5) using ILOG CPLEX 12.5.0.1 for optimization, with the default parameters, and run in a laptop with a Intel Core i7 3610QM @ 2.30GHz CPU. The branching scheme is based on the x_{it} variables, as described in section 7.1.

All the test instances were generated using the procedure described in [Gicquel et al., 2012]. Furthermore, the instances have similar characteristics, namely, there are 4 sets of instances:

- set A: small instances ($R = 2$, $I = 10$ and $T = 50$);
- set B: instances with a large number of periods ($R = 2$, $I = 10$ and $T = 150$);
- set C: instances with a large number of items ($R = 2$, $I = 25$ and $T = 50$);
- set D: instances with a large number of resources ($R = 10$, $I = 10$ and $T = 50$).

These sets were combined with 5 levels of used capacity (75%, 80%, 85%, 90% and 95%). For each combination, 3 instances were generated, resulting in a total of 60 instances.

The computational results are shown in table 2, where each line contains aggregate results for the 3 instances in each combination described above, and the columns have the following meaning: column *UC* refers to the used capacity; columns *Nodes* and *Cols* are the average number of nodes in the branch-and-price tree and the average number of columns generated, respectively; columns *TMIP* and *TBP* are average times (in seconds) to solve to optimality the ILP formulation presented in section 3 (*TMIP*) using the CPLEX MIP Solver and the proposed branch-and-price framework (*TBP*), respectively; columns *SMIP* and *SBP* show the number of instances solved to optimality using each procedure within a time limit of 30 minutes; column *LBInc* shows the average increase, in percentage of the ILP formulation LP relaxation bound, to the LP relaxation of the reformulated model[3]; finally, column *Gap* shows the average gap, in percentage, between the LP relaxation of the root node and the optimal (or best) solution found[4].

Table 2: Extended computational results.

Instance set	UC	Nodes	Cols	TMIP	SMIP	TBP	SBP	LBInc	Gap
	75	1.7	602.3	2.36	3	0.39	3	89.4	0.01
A: $R = 2$,	80	3.0	512.0	0.90	3	0.55	3	70.0	0.06
$I = 10$ and	85	1.0	774.0	4.09	3	0.45	3	85.6	0.00
$T = 50$	90	5.7	1031.0	1.20	3	0.41	3	67.0	0.16
	95	34.0	1686.0	4.85	3	0.66	3	69.0	0.32
	75	251.3	4795.7	219.47	2	14.42	3	87.3	0.30
B: $R = 2$,	80	3844.3	22495.0	-	0	255.79	3	96.4	0.26
$I = 10$ and	85	2051.3	35514.0	-	0	19.07	2	75.9	0.43
$T = 150$	90	617.0	18415.7	-	0	91.90	3	78.8	0.32
	95	2624.0	78555.7	-	0	1046.95	2	75.6	0.48
	75	3.7	583.0	1.20	3	0.52	3	88.0	0.03
C: $R = 2$,	80	1.0	716.0	3.20	3	0.57	3	90.9	0.00
$I = 25$ and	85	35.7	1040.7	5.44	3	0.53	3	105.3	0.05
$T = 50$	90	55.7	1010.7	5.09	3	0.50	3	85.9	0.13
	95	700.3	1710.7	4.73	3	1.10	3	71.5	0.18
	75	30.3	573.0	0.99	3	5.60	3	8.9	0.29
D: $R = 10$,	80	5.7	858.3	1.44	3	5.90	3	10.2	0.03
$I = 10$ and	85	1228.3	3080.0	1.97	3	50.21	3	6.3	0.37
$T = 50$	90	465.3	5410.0	1.94	3	144.79	3	7.8	0.28
	95	3185.0	12928.3	6.85	3	214.27	2	9.3	0.67

These results seem to confirm the quality of the lower bound of the reformulated model, as shown by the increase in the bound when comparing with the ILP formulation and the small gap to the optimal solution.

[2] holding costs are drawn from a discrete uniform distribution between 5 and 10 and startup costs from a discrete uniform distribution between 100 and 200.

[3] Using the notation presented in section 8, $LBInc = 100 \times (MRoot - ILPRel) / ILPRel$.

[4] If *Best* represents the optimal or best solution found, $Gap = 100 \times (Best - MRoot) / MRoot$.

Except for the instances in set D, where the generic MIP solver outperformed the proposed framework, in all the other instances the results are favorable to the proposed framework. For the instances in set B, which are difficult to solve with the ILP formulation, our framework solved 13 out of 15 instances, while the MIP solver solved only 2.

Future research efforts should try to fully understand the results for set D and improve the performance for those set of instances, probably with the help of additional cuts, different branching schemes and/or with an heuristic approach.

9 Acknowledgments

The authors want to thank the anonymous reviewers of the IO2013 conference for the insightful comments to the first version of this paper.

References

[Caserta and Voß¸, 2013] Caserta, M. and Voß¸, S. (2013). A math-heuristic Dantzig-Wolfe algorithm for capacitated lot sizing. *Annals of Mathematics and Artificial Intelligence*, pages 1–18.

[Cattrysse et al., 1993] Cattrysse, D., Salomon, M., Kuik, R., and van Wassenhove, L. N. (1993). A dual ascent and column generation heuristic for the discrete lotsizing and scheduling problem with setup times. *Management Science*, 39(4):477–486.

[Degraeve and Jans, 2007] Degraeve, Z. and Jans, R. (2007). A new Dantzig-Wolfe reformulation and branch-and-price algorithm for the capacitated lot-sizing problem with setup times. *Operations Research*, 55(5):909–920.

[Gicquel et al., 2011] Gicquel, C., Minoux, M., and Dallery, Y. (2011). Exact solution approaches for the discrete lot-sizing and scheduling problem with parallel resources. *International Journal of Production Research*, 49(9):2587–2603.

[Gicquel et al., 2012] Gicquel, C., Wolsey, L. A., and Minoux, M. (2012). On discrete lot-sizing and scheduling on identical parallel machines. *Optimization Letters*, 6(3):545–557.

[Karimi et al., 2003] Karimi, B., Fatemi Ghomi, S. M. T., and Wilson, J. M. (2003). The capacitated lot sizing problem: a review of models and algorithms. *Omega*, 31(5):365–378.

[Lasdon and Terjung, 1971] Lasdon, L. S. and Terjung, R. C. (1971). An efficient algorithm for multi-item scheduling. *Operations Research*, 19(4):946–969.

[Manne, 1958] Manne, A. S. (1958). Programming of economic lot sizes. *Management Science*, 4(2):115–135.

[van Eijl and van Hoesel, 1997] van Eijl, C. A. and van Hoesel, C. P. M. (1997). On the discrete lot-sizing and scheduling problem with Wagner-Whitin costs. *Operations Research Letters*, 20(1):7–13.

[van Hoesel and Kolen, 1994] van Hoesel, S. and Kolen, A. (1994). A linear description of the discrete lot-sizing and scheduling problem. *European Journal of Operational Research*, 75(2):342–353.

[van Hoesel et al., 1991] van Hoesel, S., Wagelmans, A., and Kolen, A. (1991). A dual algorithm for the economic lot-sizing problem. *European Journal of Operational Research*, 52(3):315–325.

[Wagner and Whitin, 1958] Wagner, H. M. and Whitin, T. M. (1958). Dynamic version of the economic lot size model. *Management Science*, 5(1):89–96.

[Zangwill, 1968] Zangwill, W. I. (1968). Minimum concave cost flows in certain networks. *Management Science*, 14(7):429–450.

A criação de horários no Ensino Superior Português: uma solução real para o problema real

Pedro Fernandes†, Carla Sofia Pereira*, Armando Barbosa†

† *Bullet Solutions, Porto, Portugal*
E-mail: pedro.fernandes@bulletsolutions.com

* *CIICESI-ESTGF, Instituto Politécnico do Porto, Felgueiras, Portugal*
E-mail: carla.pereira@eu.ipp.pt

† *Bullet Solutions, Porto, Portugal*
E-mail: armando.barbosa@bulletsolutions.com

Resumo

Ao longo de um ano letivo, em cada Instituição de Ensino, muitos dias e muitos recursos são gastos a tentar gerir manualmente algo que pode ser automatizado e otimizado – a criação dos horários. Numa fase em que a necessidade de minimizar custos passou a fazer parte da realidade diária de todas as Instituições de Ensino, é impensável continuar a trabalhar deste modo. Apesar da criação automática de horários ser um dos problemas mais estudados pela comunidade científica, quase todos os trabalhos têm sido baseados em modelos muito simplificados da realidade, não tendo, desse modo, uma aplicação prática. Neste artigo é apresentado um resumo do trabalho realizado pela empresa Bullet Solutions ao longo dos últimos anos, desde a construção de um modelo realista do problema, passando pela idealização e desenvolvimento de algoritmos capazes de apresentar soluções válidas e de qualidade em ambientes altamente restritos, até à análise final dos resultados obtidos com a utilização em ambiente real do *software* desenvolvido.

Palavras chave: Geração automática, Horários, Ensino, Heurísticas, Otimização combinatória

1 Introdução

O problema da criação de horários para o Ensino é tipicamente definido como o escalonamento de um conjunto de aulas entre docentes e alunos, num grupo de salas, em períodos de tempo, respeitando um determinado número de restrições [Schaerf, 1999] [Bonutti et al., 2012] [Lewis, 2007].

Ao longo de um ano letivo, em cada Instituição de Ensino, muitos dias e muitos recursos são gastos a tentar encontrar manualmente uma solução que respeite todas as regras existentes (uma solução viável) e que ao mesmo tempo vá de encontro às expectativas de todos os intervenientes (uma solução de qualidade).

As diferenças entre os sistemas académicos de cada país e até entre Instituições do mesmo país, fazem com que exista uma quantidade significativa de modelos e formulações do problema [Schaerf, 1999] [Bonutti et al., 2012] [Alvarez-Valdes, Crespo and Tamarit, 2002], variando a estrutura e dependência da informação base, o conjunto de restrições utilizadas ou o conceito de qualidade da solução.

Devido ao seu caráter combinatório e complexidade associada, este é um dos problemas mais estudados pela comunidade científica e pela área da Investigação Operacional em particular [Schaerf, 1999] [Murray, Müller and Rudová, 2007] [Lewis, 2007]. Contudo, quase todos os trabalhos têm sido baseados em modelos muito simplificados da realidade, não tendo, desse modo, uma aplicação prática.

Numa fase em que a necessidade de minimizar custos passou a fazer parte da realidade diária de todas as Instituições de Ensino, é impensável continuar a gerir de um modo manual algo que pode ser automatizado e otimizado, produzindo soluções muito mais rápidas e de melhor qualidade, facilitando toda a posterior gestão de alterações ao longo do ano e libertando os recursos afetos a esta tarefa para outros trabalhos que realmente acrescentem valor.

Neste artigo são apresentados alguns dos principais resultados do trabalho efetuado pela empresa Bullet Solutions, que se iniciou em 2005 com um levantamento de requisitos detalhado, com o objetivo de construir um modelo real e completo do problema. Este modelo esteve na base da construção, no final de 2006, da primeira versão do produto Bullet TimeTabler Education (BTTE) – gerador automático e otimizado de horários – e tem vindo a ser atualizado e melhorado ao longo dos anos.

Ao contrário do que é defendido em alguns estudos, foi possível construir um modelo e posteriormente um produto completo e adaptado às diversas especificidades existentes, estando atualmente o *software* BTTE a ser utilizado com sucesso em mais de metade das Instituições de Ensino Superior de Portugal, incluindo as 10 maiores referências do país.

Este artigo está estruturado em 6 secções. De modo a fazer um enquadramento com outros trabalhos realizados sobre o problema de criação de horários no Ensino, é apresentada na secção 2 uma análise do estado da arte. Na secção 3 é descrito o cenário real da criação de horários no Ensino Superior Português. A secção 4 apresenta os algoritmos desenvolvidos para a resolução do problema. Na secção 5 são apresentados os casos de estudo e os principais resultados. Por fim, na secção 6, podem ser encontradas as principais conclusões do trabalho realizado.

2 Revisão da literatura

A criação automática de horários para o Ensino é um problema estudado pela comunidade científica desde a década de 60 do século passado. Ao longo destes quase 50 anos, largas centenas de trabalhos foram publicados, com as mais diversas formulações do problema e técnicas de resolução, podendo ser encontrado em [Schaerf, 1999] um resumo dos estudos mais relevantes até ao início deste século.

Apesar da enorme atenção dada a este problema, dos esforços colocados na tentativa de o resolver e dos consideráveis avanços conseguidos, continua a existir uma grande distância entre os modelos utilizados na investigação efetuada e a realidade encontrada nas Instituições de Ensino. Em [McCollum, 2006] é abordado sem qualquer tipo de complexo o longo caminho que ainda é necessário percorrer para diminuir essa distância, sendo efetuada uma excelente análise crítica das opções tomadas nos estudos realizados e da impossibilidade de as aplicar na prática. Por um lado, o autor compreende a necessidade de restringir o problema de modo a poder existir um modelo de comparação de estudos e métodos de resolução, mas ao mesmo tempo mostra que apenas trabalhando sobre um modelo real, ou perto de real, poderão ser aplicados com sucesso, na prática, os resultados de toda essa investigação.

Em 2012 é publicado um novo estudo [Bonutti et al., 2012] que pretende suprir a enorme lacuna identificada em [McCollum, 2006]. Nesse trabalho, os autores procuram criar um modelo mais próximo da realidade encontrada nas Instituições de Ensino, de modo a poder servir de base para novos estudos, procurando assim a aproximação que permita transformar a investigação efetuada na solução para os problemas reais existentes. Apesar de ter sido dado um passo importante, a modelação apresentada está ainda longe de estar completa.

Ainda antes dessa tentativa formal de criação de um modelo mais abrangente, podem ser encontradas algumas abordagens com formulações do problema interessantes [Daskalaki and Birbas, 2005] [Daskalaki, Birbas and Housos, 2004], já com a preocupação de incluir na análise um grande conjunto de restrições e objetivos esquecidos na esmagadora maioria dos estudos existentes e cruciais para um modelo real. Duas abordagens muito práticas, com visões muito direcionadas para a utilização final, podem ser encontradas em [Alvarez-Valdes, Crespo and Tamarit, 2002] e [Murray, Müller and Rudová, 2007]. Nesses estudos existe uma preocupação em modelar casos reais de Instituições de Ensino, procurando posteriormente os métodos de resolução mais adequados.

Adicionalmente, poderá ser consultado [Lewis, 2007], onde são comparados estudos que abordam o problema com recurso à utilização de meta-heurísticas, pela importância e contributo destas técnicas para a área em questão e [Blum and Roli, 2003], onde é apresentada a importância das meta-heurísticas em problemas combinatórios complexos e onde é efetuada a análise e comparação dos diversos métodos existentes. Igualmente relevantes são dois estudos onde é discutida a importância e o impacto de uma correta definição das estruturas de vizinhança para a resolução de problemas combinatórios de grande dimensão e que podem ser encontrados em [Abdullah, Burke and McCollum, 2005] e [Ahuja et al., 2002].

Para além do problema já abordado associado à falta de um modelo representativo de base, o distanciamento entre a investigação e as necessidades reais referido em [McCollum, 2006] origina um segundo problema: o foco de quase todos os estudos existentes está na otimização das soluções encontradas, quando em muitos casos o maior desafio real que as Instituições enfrentam passa por encontrar uma solução viável para o problema. O facto de quase todos os estudos contemplarem uma ínfima parte das restrições reais existentes faz com que não se sinta propriamente esta dificuldade, estando o fator diferenciador da investigação associado à qualidade da solução final encontrada. É sem dúvida fundamental obter uma solução final com a melhor qualidade possível, mas não pode ser esquecido o primeiro passo, ou seja, encontrar as técnicas adequadas para conseguir produzir uma solução inicial viável, em ambientes muitíssimo restritos como são os da esmagadora maioria das Instituições de Ensino.

3 A criação de horários no Ensino Superior Português

3.1 Conceitos fundamentais

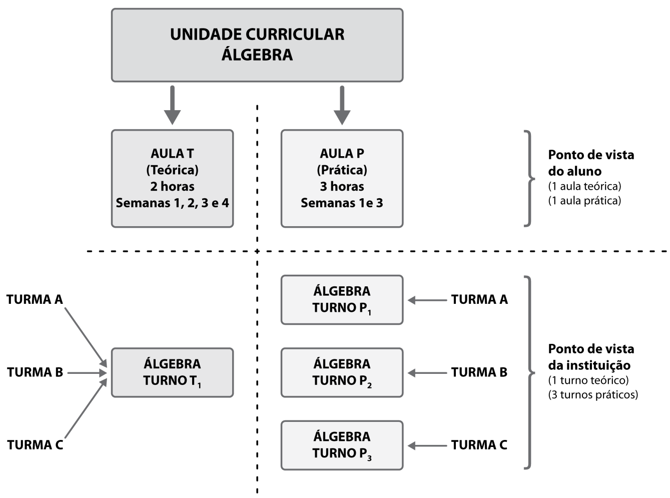
O sistema de ensino português apresenta uma grande diversidade de cenários e níveis de complexidade. No entanto, estes fatores estão muito mais relacionados com especificidades próprias (regras que têm de ser cumpridas ou objetivos que devem ser atingidos) do que propriamente com a estrutura do problema.

Numa Instituição de Ensino podem ser oferecidos diversos **cursos**. Cada curso pode ter vários **planos curriculares** e cada plano curricular é composto por um conjunto de **unidades curriculares**. Em cada plano curricular existe um conjunto de **turmas**, ou seja, grupos de alunos que seguirão esse plano curricular e partilharão o mesmo horário.

Cada Instituição tem um conjunto de **salas**, que podem estar situadas em zonas geográficas diferentes. Cada sala tem a sua capacidade e um conjunto de características que a diferencia. A Instituição tem igualmente um conjunto de **docentes**, que lecionam as unidades curriculares existentes.

As **aulas** representam a forma como é lecionada uma determinada unidade curricular, do ponto de vista do aluno. Cada aula é caracterizada por uma carga letiva de uma ou mais **tipologias** (teórica, prática, por exemplo) e pelo **espaço temporal** em que ocorre (semanas). Cada aula pode ter um ou mais **turnos**. Os turnos representam a forma como é lecionada uma unidade curricular do ponto de vista da Instituição, ou seja, quantas vezes é repetida a mesma aula para diferentes conjuntos de alunos. Os turnos são constituídos por uma ou mais turmas. Na figura 1 é apresentado um exemplo para o caso da unidade curricular Álgebra, com duas aulas (T e P), uma delas de tipologia teórica com duração 2 horas e outra de tipologia prática com duração 3 horas, que ocorrem em semanas diferentes. A aula T é composta por um turno (T_1) que agrega as três turmas existentes (Turma A, Turma B, Turma C). A aula P é composta por três turnos (P_1, P_2, P_3), estando as três turmas separadas, uma em cada turno.

Figura 1: Exemplo do modo de funcionamento de uma unidade curricular.



Assim, para criar os horários de uma Instituição, deverá ser colocado na grelha horária um conjunto de **eventos**. Cada evento é composto por um turno (conjunto de alunos de uma ou mais turmas), por uma ou mais salas e por um ou mais docentes. Um evento pode ser lecionado numa sala pré-determinada, numa sala de um grupo de salas possíveis ou em mais do que uma sala em simultâneo. Um evento pode ser lecionado por um docente pré-determinado, por um docente de um grupo de docentes possíveis ou por mais do que um docente em simultâneo. O evento será colocado numa determinada posição da grelha horária, tendo uma duração e ocorrendo num conjunto de semanas, respeitando todas as restrições existentes. A grelha horária deverá ser configurável, permitindo a indicação do tamanho do *slot* mínimo, ou seja, dos intervalos possíveis de alocação de eventos (por exemplo, se for definido um *slot* de 60 minutos, a grelha horária será dividida em intervalos de 60 minutos, podendo ser iniciados eventos em cada um desses intervalos). A duração do evento deverá ser um múltiplo do *slot* definido (no exemplo anterior, um evento de 120 minutos deveria ser definido como tendo uma duração equivalente a dois *slots*).

A introdução de um conceito de múltiplas semanas na criação dos horários acrescenta uma complexidade enorme ao problema, uma vez que a solução do mesmo não passa pela "simples" criação de um horário base para uma semana que se repete nas semanas seguintes.

3.2 Restrições

Uma solução é viável se respeitar todas as restrições impostas previamente. Existem regras próprias de cada Instituição, restrições legais ou impossibilidades físicas que têm de ser cumpridas.

Todas as restrições que são apresentadas de seguida podem ser encontradas no produto BTTE, com a possibilidade de serem parametrizadas individualmente:

- Cada recurso (seja docente, turma ou sala) só pode ter um evento atribuído em cada momento;
- Cada sala só pode receber um número de alunos que respeite a sua capacidade;
- Eventos que têm de ser contíguos;
- Eventos que não podem ser lecionados no mesmo dia;
- Eventos que têm de ser sobrepostos;
- Eventos que têm de ser separados (podendo, no entanto, ocorrer no mesmo dia);
- Indisponibilidades de todos os intervenientes (instituição, salas, turmas, unidades curriculares, docentes e turnos), ou seja, espaços temporais onde não podem ser atribuídos eventos, para cada recurso de cada tipo de entidade envolvida. As pausas para intervalos de refeição, que podem ser flexíveis ou rígidas, são uma extensão das indisponibilidades atrás referidas, nos casos dos docentes e alunos;
- Limites de horas diárias de eventos que podem ser atribuídos a docentes e turmas;
- Limites de horas consecutivas de eventos que podem ser atribuídos a docentes e turmas;
- Tipologias de eventos, dentro de uma unidade curricular, que têm de respeitar uma ordem de ocorrência na semana (os eventos práticos só podem ser lecionados depois dos teóricos, por exemplo);
- Intervalo mínimo necessário entre eventos com início em dias diferentes, para docentes e turmas (de modo a não permitir que um aluno ou docente que tenha eventos num dia até muito tarde, tenha um novo evento atribuído no dia seguinte demasiado cedo);
- Tempo mínimo de deslocação entre eventos, para docentes e alunos (para garantir que há um tempo suficiente para se poderem deslocar entre zonas geográficas diferentes em eventos consecutivos);
- Garantir um número mínimo de dias livres para determinados docentes (compressão do horário para libertar tempo aos docentes para outras atividades).

É importante salientar que, numa situação real, a "linha" que separa algumas situações que têm obrigatoriamente de ser respeitadas (restrições) de outras que preferencialmente deveriam ser respeitadas (objetivos) é ténue e varia de Instituição para Instituição. Assim, esses casos poderão ser encontrados tanto no subcapítulo das restrições como no dos objetivos, uma vez que poderão ser utilizados de um modo ou de outro, de acordo com a parametrização que o utilizador efetuar na aplicação.

3.3 Objetivos

A qualidade de uma solução está associada ao cumprimento de um conjunto de objetivos. As relações entre esses objetivos e o modo como é penalizado o seu incumprimento formam a função objetivo que avalia a qualidade de uma solução viável.

Na realidade portuguesa, podem ser encontrados diversos conceitos de qualidade e os mais variados objetivos que devem ser atingidos para podermos dizer que estamos perante uma boa solução. De seguida são apresentados os objetivos existentes no BTTE, que podem ser parametrizados individualmente:

- Eventos em dias diferentes que devem começar à mesma hora;
- Eventos que devem ser lecionados na mesma sala;
- Eventos que devem ser contíguos;
- Eventos que devem ser sobrepostos;
- Em cada período do dia (manhã, tarde ou pós-laboral), procurar colocar eventos de tipologias diferentes, nos horários das turmas (para uma turma, num determinado período do dia, não devem existir apenas eventos da mesma tipologia);
- Conseguir períodos livres (sem eventos atribuídos) nos horários das turmas e docentes;
- Procurar garantir um mínimo de horas por período para turmas e docentes (de modo a evitar que estes tenham de se deslocar à Instituição apenas por um período de tempo reduzido);
- Respeitar preferências dos eventos por determinados períodos (eventos preferencialmente atribuídos a um período do dia);
- Evitar furos nos horários de turmas, docentes e salas;
- Evitar trocas de salas nos horários de turmas e docentes;
- Evitar as alterações logísticas das salas (tentar que os eventos que forem atribuídos a uma sala tenham necessidades de equipamento o mais semelhantes possível);
- Utilização das salas mais indicadas para cada evento (tanto do ponto de vista do aproveitamento da sua capacidade, como de possíveis indicações de preferência que existam);

- Respeitar preferências de todos os intervenientes (instituição, salas, turmas, unidades curriculares, docentes e turnos), ou seja, espaços temporais onde preferencialmente não devem ser atribuídos eventos, para cada recurso de cada tipo de entidade envolvida, podendo ser definidos dois níveis de preferência;
- Tipologias de eventos, dentro de uma unidade curricular, que devem respeitar uma ordem de ocorrência na semana (os eventos práticos só devem ser lecionados depois dos teóricos, por exemplo);
- Conseguir um número mínimo de dias livres para determinados docentes (tentativa de compressão do horário para libertar tempo aos docentes para outras atividades);
- Docentes que devem lecionar eventos nos mesmos dias (docentes com horários semelhantes).

O conceito de qualidade varia de Instituição para Instituição e até dentro da própria Instituição. Assim, o BTTE permite que a construção da função objetivo seja definida pelo utilizador (tanto os objetivos que a constituem, como a ponderação individual de cada um desses objetivos), podendo, desse modo, adaptar-se às diversas necessidades existentes.

4 Algoritmos desenvolvidos

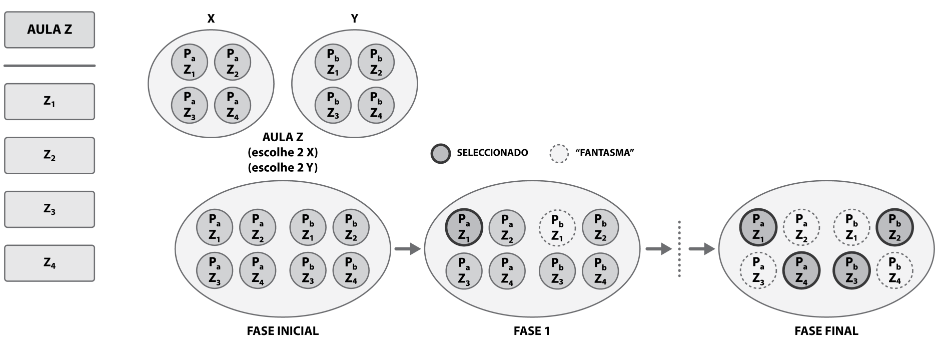
4.1 Heurística construtiva: construção da solução inicial

A criação de horários numa Instituição de Ensino só se encontra concluída quando todos os eventos estiverem escalonados, respeitando todas as restrições existentes. Uma criação parcial, ou seja, uma solução que não inclua todos os eventos que terão de ser lecionados, não pode ser aplicada na prática. Adicionalmente, uma solução viável pode ser melhor ou pior do que outra, respeitando mais ou menos os objetivos estabelecidos que avaliam a qualidade final da mesma.

A solução inicial é construída a partir de uma grelha horária vazia. Em cada momento de construção existe uma lista com todos os eventos que ainda não foram colocados na grelha horária. No entanto, nem todos os elementos desta lista serão colocados na grelha final. Alguns deles são eventos "fantasma" criados para permitir ao algoritmo uma maior flexibilidade no seu espaço de pesquisa. Para uma melhor compreensão deste conceito, observe-se o exemplo que é apresentado de seguida.

Se o Docente P_a lecionar dois turnos da Aula Z (composta pelos turnos Z_1, Z_2, Z_3, Z_4) e o Docente P_b lecionar outros dois turnos da mesma Aula, não havendo qualquer indicação contrária, tanto o Docente P_a como o Docente P_b podem lecionar qualquer um dos quatro turnos existentes, sendo necessário apenas garantir que cada um lecionará sempre dois e só dois turnos. Assim, na lista de eventos para colocar existirão as combinações P_aZ_1, P_aZ_2, P_aZ_3, P_aZ_4, P_bZ_1, P_bZ_2, P_bZ_3, P_bZ_4, sendo que do grupo de eventos do P_a (conjunto X) serão colocados dois eventos (sendo os outros dois eventos "fantasma"), o mesmo acontecendo com o grupo de eventos do P_b (conjunto Y). À medida que são selecionados eventos para a sua colocação na grelha final, é atualizada a informação, é controlado o número de eventos atribuídos a cada conjunto e são "adormecidos" os eventos necessários (por exemplo, se for colocado o evento P_aZ_1, o evento P_bZ_1 deixa de ser válido, uma vez que o turno Z_1 já foi atribuído ao Docente P_a). Na figura 2 é apresentada a situação anterior, desde o estado inicial (ponto de partida do algoritmo), até ao estado final (quando todos os eventos estiverem colocados na grelha horária).

Figura 2: Exemplo de utilização de eventos "fantasma".



Tendo a lista de eventos por colocar definida, é necessário escolher qual o evento que vai ser colocado na grelha horária. Para essa seleção, foi definido um critério de urgência, que é calculado com base em quatro fatores: o grau de dependência existente entre eventos associados por restrições, uma vez que se pode tornar extremamente complexo colocar grupos de eventos que estejam associados através de regras de contiguidade ou sobreposição (quanto mais eventos envolvidos, mais urgente é a colocação desses eventos); a sobreposição da disponibilidade de todos os recursos envolvidos no evento, que basicamente define em quantos locais diferentes o evento pode ser colocado em cada momento (quanto menos locais disponíveis, mais urgente é o evento); o refinamento do fator anterior, eliminando adicionalmente locais potencialmente viáveis mas que deixaram de o ser devido a outras restrições como, por exemplo, o limite de horas diárias de eventos (quanto menos locais disponíveis, mais urgente é o evento); o maior ou menor leque de salas possíveis para receber o evento (quanto menor for o número de salas disponíveis, maior é a probabilidade desse local se tornar rapidamente indisponível e, portanto, mais urgente é o evento).

Após a seleção do evento mais urgente, procede-se à avaliação de qual o melhor local para o colocar. A estrutura de dados montada permite avaliar apenas os locais viáveis, não sendo perdido tempo a analisar locais inválidos. A escolha do melhor local de colocação é definida com base em dois critérios fundamentais: o local de melhor qualidade, procurando construir uma solução inicial que já tenha preocupação em atingir os objetivos definidos pelo utilizador na criação da função objetivo; o local que causar um menor impacto nos restantes eventos por colocar, procurando evitar escolhas que levem a uma considerável limitação das possibilidades de colocação dos eventos que ainda terão de ser colocados.

Após a colocação de cada evento é efetuada a seleção de um novo evento mais urgente, recalculando o critério de urgência. Com a colocação do evento, todos os mapas de dados são atualizados. A estrutura de informação montada no produto BTTE permite atualizar e manipular todos os mapas de dados necessários para estes cálculos em tempo real. Assim, em cada momento, é possível ter uma noção exata das possibilidades de colocação de cada evento e da sua urgência. É igualmente possível saber qual é o impacto que a colocação de qualquer evento terá em todos os eventos que ainda não foram colocados.

Mesmo tentando ter todos os cuidados na sequência pela qual são tratados os eventos, em problemas muito restritos, como são quase todos os casos reais, é natural que, em determinados momentos, o evento que foi escolhido como o mais urgente não tenha nenhum local possível de colocação. Quando estas situações ocorrem, não adianta continuar a colocar novos eventos sem encontrar primeiro uma solução para o evento atual. Essas novas colocações apenas continuariam a restringir o problema.

Após terem sido testados, sem sucesso, métodos tradicionais de *backtracking* para resolver este problema, foram desenvolvidos métodos de tratamento próprios e adaptados ao modelo existente. Assumindo que, se existir uma solução possível para o problema, existirá pelo menos um encaixe de todos os eventos na grelha, isso significa que não é necessário retirar eventos já colocados, mas sim movê-los para outras posições, de modo a tentar abrir um novo espaço para o evento com dificuldades. A estrutura de informação existente e os mapas de dados permitem, de novo, que estes movimentos não sejam aleatórios, aumentando imenso a probabilidade de efetuar os movimentos corretos. Este é um dos pontos cruciais para conseguir encontrar uma solução viável num ambiente real restrito. Atualmente, o produto BTTE tem incorporado métodos de fuga resultantes de vários anos de aperfeiçoamento, permitindo ultrapassar este tipo de problemas com bastante destreza.

Assim que os métodos de fuga conseguem encontrar um local de colocação para o evento em dificuldades, este é colocado e prossegue normalmente a construção da solução inicial.

A solução inicial é encontrada quando todos os eventos tiverem sido colocados com sucesso.

4.2 Heurística melhorativa: obtenção de soluções de melhor qualidade

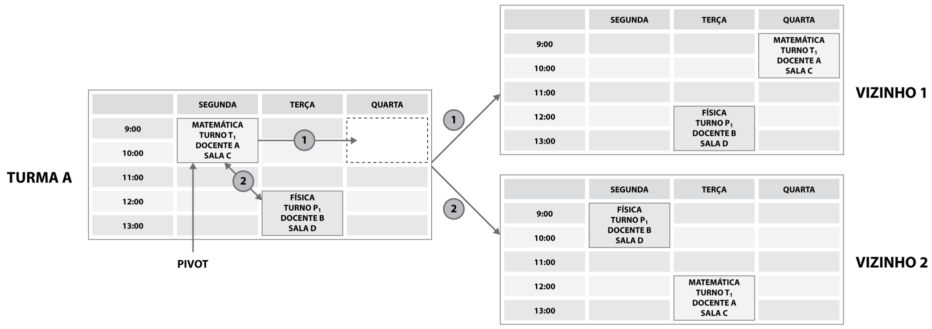
Encontrada a solução inicial para o problema (ponto de partida), é iniciada a fase de otimização, onde a partir de métodos apropriados, são procuradas progressivamente soluções de melhor qualidade.

No BTTE a pesquisa de novas soluções é baseada em estruturas de vizinhanças, ou seja, numa construção iterativa de conjuntos de vizinhos (soluções com pequenas alterações em relação à atual) que são avaliados e aceites (ou não) de acordo com determinados parâmetros, permitindo desse modo uma exploração do espaço de soluções. Para além dos diferentes modos de construção de vizinhanças, os algoritmos implementados passam por três grandes fases: normal, intensificação e diversificação. Cada uma destas fases foi especificada para atingir um determinado fim, sendo o seu funcionamento em conjunto a chave para um resultado final otimizado. Apesar de terem sido desenvolvidas de raiz heurísticas totalmente adaptadas ao problema existente e modelo criado, existiram conceitos e princípios de meta-heurísticas como, por exemplo, a Pesquisa Tabu [Blum and Roli, 2003], que serviram de base para a investigação efetuada.

A opção por utilizar mais do que uma estrutura de vizinhança permitiu diversificar a pesquisa, tornando-a mais robusta e completa. Assim, foram implementados dois modos diferentes de criar vizinhos, ou seja, novas soluções: trocas diretas de eventos; trocas de eventos em dois passos.

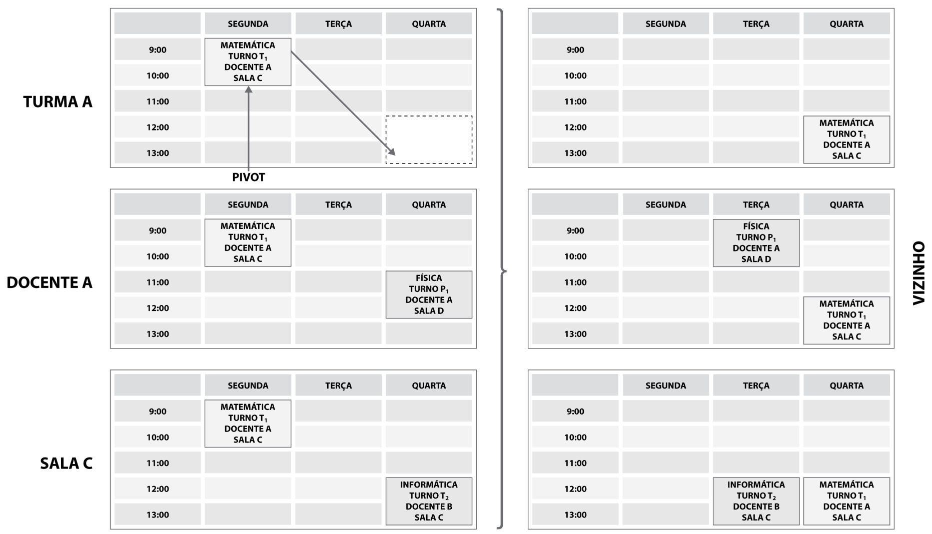
No primeiro caso, é escolhido um evento *pivot* e, com o apoio dos mapas de dados existentes, são selecionados potenciais locais de destino. Os locais de destino podem estar livres ou ocupados. Se estiverem livres, há imediatamente um vizinho válido. Se estiverem ocupados, o evento ou eventos que estão a ocupar o local de destino serão movimentados para o local original do evento *pivot*, criando um vizinho válido se for possível efetuar a troca. A figura 3 ilustra a primeira estrutura de vizinhança, utilizando dois eventos de uma turma.

Figura 3: Primeira estrutura de vizinhança.



No segundo caso, a grande diferença está no modo de tratamento dos eventos que estão a ocupar o local de destino, sendo feita uma avaliação mais completa das sobreposições existentes e procurando locais de colocação alternativos para além do local original do evento *pivot*. A figura 4 ilustra a segunda estrutura de vizinhança, utilizando três eventos e as entidades (turma, docente e sala) associadas ao *pivot*.

Figura 4: Segunda estrutura de vizinhança.



Com a primeira estrutura de vizinhança conseguem-se resultados mais rápidos e com menos alterações em cada iteração (os vizinhos são mais "parecidos", uma vez que são construídos com base em trocas diretas de eventos ou em movimentações para locais livres). Com a segunda estrutura de vizinhança conseguem-se encontrar novos vizinhos interessantes apenas com uma iteração (os vizinhos podem ter diferenças maiores, uma vez que o número de eventos incluídos na troca pode ser superior, bem como os potenciais locais de destino de cada um deles).

Em cada iteração é sempre aceite o melhor vizinho existente (vizinho com melhor valor da função objetivo), mesmo que este vizinho tenha uma qualidade inferior à da solução atual. Após diversos testes optou-se por esta abordagem, uma vez que na esmagadora maioria das situações esta técnica de aceitação permitiu, por um lado, fugir a mínimos locais e, por outro, não colocar em risco a recuperação rápida para soluções de qualidade.

Dependendo da fase de pesquisa (normal, intensificação ou diversificação), a utilização das vizinhanças descritas tem características próprias. Na fase normal pretende-se uma pesquisa mais rápida e uma exploração mais livre do espaço de soluções. Assim, o número de vizinhos criado em cada iteração é limitado, quer pela diminuição do número de eventos *pivot* utilizados, quer pela diminuição do número de potenciais locais de destino avaliados. Na fase de intensificação pretende-se explorar mais detalhadamente as zonas de pesquisa mais promissoras. Nesta fase, o algoritmo regressa à solução de melhor qualidade que encontrou na fase normal, fazendo uma pesquisa mais profunda em seu redor (aumentando o número de eventos *pivot* utilizados e avaliando a totalidade dos potenciais locais de destino de cada um deles). O critério de paragem em cada uma destas fases passa por um número de iterações consecutivas sem conseguir melhorar a solução atual, assumindo, nesse momento, que a pesquisa deve ser mais profunda junto das melhores soluções (passando da fase normal para a fase de intensificação), ou direcionada para outra zona do espaço de soluções (passando da fase de intensificação para a fase de diversificação). Na fase de diversificação, através da adulteração temporária da função objetivo (durante um determinado número de iterações), consegue-se simular um novo ponto de partida para uma nova pesquisa (uma nova solução inicial), sem perder o tempo necessário para criar uma solução de raiz e sem perder demasiada qualidade, uma vez que a função objetivo é apenas ligeiramente alterada e não ignorada.

O processo de otimização termina por indicação do utilizador, sendo nessa altura devolvida a melhor solução encontrada.

5 Casos de Estudo

5.1 Enquadramento e principais características

Foram analisados 20 casos reais de Instituições portuguesas de Ensino Superior, clientes do produto BTTE. Procurou-se que os casos selecionados fossem o mais heterogéneos possível, com diversos graus de complexidade e dimensão, de modo a representar a realidade portuguesa.

Cada um dos casos de estudo, apresentados na tabela 1, pode ser categorizado através de 11 fatores críticos (FC) de introdução de complexidade no problema: 1) número de semanas diferentes (quanto maior, mais complexo); 2) existência de eventos contíguos e sobrepostos (quanto maior, mais complexo); 3) número de eventos a colocar (quanto maior, mais complexo); 4) número de turmas (quanto maior, mais complexo); 5) número de docentes (quanto maior, mais complexo); 6) existência de restrições de disponibilidade dos recursos (quanto maior, mais complexo); 7) sobreposição de regimes (se existir, mais complexo); 8) existência de elevada concorrência pelos recursos físicos (quanto maior, mais complexo); 9) existência de elevada diversidade de agregação de turmas na constituição dos turnos (quanto maior, mais complexo); 10) dimensão do *slot* da grelha horária, em minutos (quanto menor, mais complexo); 11) número de objetivos a analisar (quanto maior, mais complexo).

Tabela 1: Casos de estudo e fatores críticos de introdução de complexidade no problema.

	FC1	FC2	FC3	FC4	FC5	FC6	FC7	FC8	FC9	FC10	FC11
Caso 1	2	nenhum	144	28	84	médio	não	baixo	baixo	30	21
Caso 2	1	nenhum	341	50	100	baixo	não	médio	baixo	60	16
Caso 3	1	nenhum	90	55	92	médio	sim	médio	médio	30	15
Caso 4	1	nenhum	169	38	75	baixo	não	baixo	baixo	60	15
Caso 5	1	nenhum	130	17	34	baixo	não	baixo	baixo	30	17
Caso 6	2	nenhum	200	39	83	médio	sim	baixo	baixo	15	16
Caso 7	4	poucos	485	52	117	alto	sim	alto	baixo	30	16
Caso 8	1	nenhum	350	350	150	alto	não	alto	alto	5	15
Caso 9	2	nenhum	478	64	161	baixo	não	baixo	baixo	90	16
Caso 10	1	nenhum	453	83	139	alto	sim	médio	baixo	60	15
Caso 11	1	poucos	650	70	100	baixo	não	médio	médio	90	18
Caso 12	1	nenhum	200	47	101	médio	sim	médio	baixo	30	16
Caso 13	1	poucos	532	94	362	alto	não	alto	alto	30	16
Caso 14	2	poucos	605	205	667	alto	sim	médio	médio	30	19
Caso 15	2	nenhum	1200	127	621	médio	sim	médio	baixo	30	17
Caso 16	1	nenhum	1350	185	532	baixo	não	médio	baixo	30	18
Caso 17	4	poucos	1800	239	512	médio	não	médio	baixo	30	19
Caso 18	13	muitos	1855	825	600	médio	não	alto	médio	30	16
Caso 19	12	nenhum	1985	300	471	baixo	não	alto	baixo	30	15
Caso 20	15	nenhum	3500	1346	1250	médio	sim	médio	médio	30	17

5.2 Resultados e discussão

Uma vez que o modelo utilizado pela Bullet Solutions não é encontrado em nenhuma publicação da área, não é possível efetuar uma comparação com outros estudos. Assim, a opção passou por apresentar as diferenças significativas encontradas entre o processo anterior manual de criação de horários em cada uma das Instituições e os resultados que atualmente conseguem obter com o BTTE.

O tempo necessário para o BTTE apresentar resultados finais otimizados pode ser dividido em três parcelas: 1) tempo médio necessário para carregar e/ou validar a informação no sistema (TVBTTE); 2) tempo médio necessário para o cálculo da solução inicial (TSIBTTE); 3) tempo médio necessário para o cálculo de uma solução otimizada (TOBTTE). A soma destes tempos representa o tempo total de utilização do produto (TTBTTE), que pode ser comparado com o tempo total que era gasto no processo anterior manual (TTPM). O tempo TTPM foi estimado pelos responsáveis pelo processo de criação dos horários em cada uma das Instituições. Estes resultados são apresentados na tabela 2.

A máquina utilizada para as simulações foi um Intel Core 2 Duo a 2,53 GHz, com 4 GB de RAM. Foram efetuadas 20 simulações para cada um dos 20 casos de estudo, com dados reais das Instituições, sendo apresentados de seguida os valores médios obtidos.

Tabela 2: Casos de estudo e tempos médios associados ao processo de criação de horários.

	TVBTTE (HH:MM:SS)	TSIBTTE (HH:MM:SS)	TOBTTE (HH:MM:SS)	TTBTTE (HH:MM:SS)	TTPM (HH:MM:SS)	Poupança (Horas/%)
Caso 1	4:30:00	0:00:45	0:30:00	5:00:45	48:00:00	43h/89,56%
Caso 2	11:00:00	0:01:00	0:40:00	11:41:00	130:00:00	118h/91,01%
Caso 3	3:00:00	0:00:30	0:30:00	3:30:30	32:00:00	28h/89,04%
Caso 4	5:30:00	0:00:15	0:15:00	5:45:15	72:00:00	66h/92,01%
Caso 5	5:00:00	0:00:05	0:45:00	5:45:05	48:00:00	42h/88,02%
Caso 6	6:30:00	0:01:00	1:00:00	7:31:00	72:00:00	64h/89,56%
Caso 7	16:00:00	0:05:00	2:00:00	18:05:00	120:00:00	102h/84,93%
Caso 8	12:00:00	1:00:00	12:00:00	25:00:00	144:00:00	119h/82,64%
Caso 9	10:00:00	0:01:00	0:40:00	10:41:00	192:00:00	181h/94,44%
Caso 10	24:00:00	0:02:00	0:30:00	24:32:00	168:00:00	143h/85,40%
Caso 11	22:00:00	0:04:00	2:00:00	24:04:00	216:00:00	192h/88,86%
Caso 12	8:00:00	0:01:00	1:00:00	9:01:00	72:00:00	63h/87,48%
Caso 13	18:00:00	0:01:00	2:00:00	20:01:00	192:00:00	172h/89,57%
Caso 14	22:00:00	0:10:00	2:00:00	24:10:00	288:00:00	264h/91,61%
Caso 15	38:00:00	0:15:00	5:00:00	43:15:00	408:00:00	365h/89,40%
Caso 16	44:00:00	0:45:00	8:00:00	52:45:00	504:00:00	451h/89,53%
Caso 17	50:00:00	1:00:00	10:00:00	61:00:00	432:00:00	371h/85,88%
Caso 18	60:00:00	8:00:00	18:00:00	86:00:00	528:00:00	442h/83,71%
Caso 19	80:00:00	2:00:00	12:00:00	94:00:00	408:00:00	314h/76,96%
Caso 20	120:00:00	4:00:00	120:00:00	244:00:00	936:00:00	692h/73,93%

Nos 20 casos estudados, a opinião dos utilizadores apontou para a obtenção de resultados finais de melhor qualidade com a utilização do BTTE, quando comparados com os horários que eram criados manualmente. De igual modo, a enorme poupança de tempo (média acima dos 85%) e de recursos envolvidos no processo, permite às Instituições encarar o problema de criação dos horários de um modo muito mais seguro e considerar a aquisição do produto BTTE como uma aposta ganha.

Da análise das informações e resultados apresentados nas tabelas anteriores, existem alguns casos que importa salientar. O Caso 8 parece simples, mas apresenta tempos de geração bastante superiores a outros casos aparentemente semelhantes. A verdade é que ao utilizar um *slot* de 5 minutos na grelha horária (FC10) e tendo uma elevada diversidade de agregação de turmas na constituição dos turnos (FC9), a complexidade de cálculo aumenta imediatamente. O aumento significativo do tempo necessário de otimização a partir do Caso 15 é explicado, num primeiro momento, pelo número de eventos que existem no problema (FC3), claramente superior aos 14 exemplos anteriores. Os Casos 18, 19 e 20 envolvem um número de semanas diferentes (FC1) muito superior aos restantes, tendo o Caso 18 uma dificuldade adicional muito específica relacionada com o elevado número de restrições de eventos contíguos e sobrepostos (FC2). O Caso 20 é o mais completo e complexo de todos, com uma enorme diversidade de fatores a terem impacto no cálculo da solução.

6 Principais conclusões

As principais conclusões que se podem retirar do trabalho desenvolvido estão diretamente relacionadas com o sucesso comercial da aplicação BTTE. Apenas com uma correta modelação do problema (considerando que "correta" significa que um utilizador a vê como um espelho da realidade) é possível criar a base de informação necessária para o processo de criação dos horários, bem como todas as relações existentes. Apenas com algoritmos realmente robustos é possível criar soluções válidas para problemas tão restritos como os encontrados em quase todas as Instituições de Ensino. Apenas com métodos adicionais de otimização é possível encontrar resultados que vão de encontro às expectativas dos utilizadores, quer em relação à qualidade apresentada, quer em relação à velocidade com que é conseguida.

O facto de terem sido obtidos resultados de qualidade, com poupanças na ordem dos 85% do tempo investido no processo, permite considerar que os algoritmos implementados na aplicação têm um considerável grau de robustez e uma grande facilidade de adaptação aos diversos cenários encontrados.

A implementação e utilização da aplicação BTTE permitiu melhorias significativas em processos relacionados com a elaboração de horários, resultando em ganhos de produtividade adicionais. Assistiu-se à organização e centralização da informação em grande parte das Instituições, à eliminação de informação redundante e à eliminação de erros. Existe um melhor controlo da gestão sobre as reais necessidades da distribuição do serviço docente, verificando-se poupanças consideráveis após a utilização do BTTE.

Referências

[Abdullah, Burke and McCollum, 2005] Abdullah, S. and Burke, E. K. and McCollum, B. (2005). An investigation of variable neighbourhood search for university course timetabling. *The 2nd Multidisciplinary International Conference on Scheduling: Theory and Applications (MISTA)*: 413–427.

[Ahuja et al., 2002] Ahuja, R. K. and Ergun, Ö. and Orlin, J. B. and Punnen, A. P. (2002). A survey of very large-scale neighborhood search techniques. *Discrete Applied Mathematics*, 123(1-3): 75–102.

[Alvarez-Valdes, Crespo and Tamarit, 2002] Alvarez-Valdes, R. and Crespo, E. and Tamarit, J. M. (2002). Design and implementation of a course scheduling system using Tabu Search. *European Journal of Operational Research*, 137(3): 512–523.

[Blum and Roli, 2003] Blum, C. and Roli, A. (2003). Metaheuristics in Combinatorial Optimization: Overview and Conceptual Comparison. *ACM Computing Surveys (CSUR)*, 35(3): 268–308.

[Bonutti et al., 2012] Bonutti, A. and De Cesco, F. and Di Gaspero, L. and Schaerf, A. (2012). Benchmarking curriculum-based course timetabling: formulations, data formats, instances, validation, visualization, and results. *Annals of Operations Research*, 194(1): 59–70.

[Daskalaki and Birbas, 2005] Daskalaki, S. and Birbas, T. (2005). Efficient solutions for a university timetabling problem through integer programming. *European Journal of Operational Research*, 160(1): 106–120.

[Daskalaki, Birbas and Housos, 2004] Daskalaki, S. and Birbas, T. and Housos, E. (2004). An integer programming formulation for a case study in university timetabling. *European Journal of Operational Research*, 153(1): 117–135.

[Lewis, 2007] Lewis, R. (2007). A survey of metaheuristic-based techniques for University Timetabling problems. *OR Spectrum*, 30(1): 167–190.

[McCollum, 2006] McCollum, B. (2006). University timetabling: Bridging the gap between research and practice. *Proceedings of the 6th International Conference on the Practice and Theory of Automated Timetabling*: 15–35.

[Murray, Müller and Rudová, 2007] Murray, K. and Müller, T. and Rudová, H. (2007). Modeling and Solution of a Complex University Course Timetabling Problem. *Practice and Theory of Automated Timetabling VI: 6th International Conference, PATAT 2006, Brno, Czech Republic, August 30-September 1, 2006, Revised Selected Papers*, 3867: 189–209.

[Schaerf, 1999] Schaerf, A. (1999). A Survey of Automated Timetabling. *Artificial Intelligence Review*, 13(2): 87–127.

Análise da eficiência das microempresas do setor do retalho no interior de Portugal: uma aplicação Data Envelopment Analysis

Fernandes, António B.*, Vaz, Maurício A.†

* *Escola Superior de Tecnologia e Gestão, Instituto Politécnico de Bragança, Portugal*
*E-mail:*antoniof@ipb.pt

† *Escola Superior de Tecnologia e Gestão, Instituto Politécnico de Bragança, Portugal,*
*E-mail:*mav@ipb.pt

Resumo

A avaliação das organizações e a determinação do desempenho, obtido pelo exercício da gestão, tem sido uma preocupação constante de gestores e detentores de capital. Nos dias de hoje, a questão coloca-se com maior acuidade quer pela competitividade acrescida quer pela dimensão e complexidade das empresas. Pretende-se, com este trabalho, fazer a aplicação da metodologia DEA - Data Envelopment Analysis a um conjunto de microempresas do setor do comércio a retalho situadas no interior de Portugal. Os resultados obtidos constituem uma medida simples de comparar o desempenho relativo das diversas empresas com base nos valores de inputs e outputs observados.

Palavras chave: Micro Empresas, DEA, Desempenho, Eficiência.

1 Introdução

Em Portugal, à semelhança de muitos outros países na Europa, nomeadamente Alemanha, Espanha, Grécia, Itália, Bélgica e outros, as Pequenas e Médias Empresas (PME) constituem a maioria das empresas do tecido empresarial Português, representando uma fatia importante da oferta de emprego. As PME são a base da economia de mercado, sendo essencial, a sua existência, para o desenvolvimento de um país ou até mesmo de uma região [Hu & Schive, 1998]. Dada a importância que este tipo de empresas tem, na economia global, procura-se neste trabalho, avaliar a sua eficiência. Para tal, comparam-se os desempenhos, de um conjunto de microempresas do interior de Portugal do setor de atividade da divisão 47 - Comércio a retalho, exceto de veículos automóveis e motociclos. São utilizados dados de 35 microempresas para o ano 2009, recorrendo ao DEA (Data Envelopment Analysis). O DEA é uma técnica que recorre a modelos de programação matemática para analisar as combinações ótimas entre inputs e outputs, com base no desempenho observado de unidades produtivas. Essas combinações constituem uma fronteira e permitem determinar os níveis de ineficiência relativa [Moreira, 2008]. Com este trabalho pretende-se, de alguma forma, contrariar a subjetividade da análise do desempenho das microempresas. A evidência empírica revela que, fundamentalmente as instituições bancárias, fazem as suas análises assentes numa bateria de indicadores ponderados subjetivamente. O modelo aplicado pretende definir uma medida única de eficiência relativa, considerando os valores observados na generalidade das empresas homogéneas, tendo em conta que os pesos são determinados endogenamente pelo próprio modelo e não pelo analista. O artigo encontra-se estruturado da seguinte forma. O ponto 2 contextualiza o enquadramento das microempresas no tecido empresarial português. No ponto 3 apresenta-se de forma breve a metodologia, caracteriza-se a amostra e as variáveis e descreve-se o procedimento de estimação. No ponto 4 comentam-se os resultados, apresentando-se as conclusões no ponto 5.

2 Enquadramento das microempresas em Portugal

Hu e Schive [1998] abordam uma lição das jovens árvores da floresta como elas lutam para cima através da sombra entorpecedora dos seus rivais mais velhos. Muitas sucumbem no caminho, outras apenas sobrevivem. No entanto, as que se tornam mais fortes, a cada ano que passa, recebem uma parcela maior de luz e ar com cada aumento da sua altura, ganhando progressivamente posições aos seus maiores rivais. As PME têm armas de sobrevivência no mercado diferentes das grandes empresas. Atendendo à sua dimensão, as PME não são capazes de desfrutar de economias de escala na produção em massa, ou nas economias obtidas na diversificação de produtos, no entanto, podem reagir rapidamente aos sinais do mercado [Jenkins, 2006]. Hu e Schive [1998] consideram que, enquanto a inovação não é um pré-requisito para a sobrevivência das PME, a capacidade de recuperar rapidamente, certamente já se torna uma exigência. Segundo dados do Instituto Nacional de Estatística de 2010, como se verifica pela análise da tabela 1, das 1.144.150 empresas existentes em Portugal, 95,81% têm menos de 10 trabalhadores ao seu serviço o que é revelador da importância das microempresas no tecido empresarial português. Desta forma, a importância deste conjunto de empresas, para a economia nacional, manifesta-se naturalmente, em termos de emprego (com 44,51% do total das pessoas ao serviço) e em termos de volume de negócios nacional representam 25,41%, sendo por isso fundamental, na nossa perspetiva, analisar o seu desempenho.

Tabela 1: Estrutura do tecido empresarial Português

	Total	$<$ de 10 PT		10 - 49 PT		50 - 249 PT		$\geq$ 250 PT	
n.º de Empresas	1.144.150	1.096.155	95,8%	41.308	3,6%	5.792	0,5%	895	0,1%
Volume de Neg. (10^3 €)	356.390.110	90.546.236	25,4%	84.321.035	23,7%	77.952.716	21,9%	103.570.123	29,1%
Pessoas ao Serviço	3.843.268	1.710.671	44,5%	778.710	20,3%	556.619	14,5%	797.268	20,7%

As PME, comparativamente às grandes empresas, apresentam caraterísticas de gestão muito próprias, nomeadamente: à medida que aumenta a dimensão das empresas, aumenta a sua capacidade de endividamento global, designadamente de médio e longo prazo; à medida que aumenta a capacidade de autofinanciamento da empresa, menos as empresas recorrem ao endividamento; o financiamento externo destas empresas é essencialmente bancário, dado, entre outros fatores, a inacessibilidade ao mercado de capitais, e este é fundamentalmente de curto prazo, uma vez que as empresas apresentam dificuldade em ceder garantias adicionais, particularmente relevantes, em períodos de dificuldade comercial [Vieira & Novo, 2010]. Desta forma, o sucesso, das microempresas, depende muito da capacidade do seu gestor, da sua estrutura bem como da sua capacidade de posicionamento no mercado. Correia, Teixeira e Rebelo [2011] verificaram que, neste tipo de empresas, a intensidade competitiva e a turbulência de mercado têm influência negativa em indicadores de performance como a rendibilidade do ativo, retorno sobre o investimento e rendibilidade operacional das vendas. Os autores consideram que, "a estes resultados não será decerto alheia a maior exigência estrutural a que as empresas estão sujeitas neste tipo de contexto, pois a intensidade competitiva e a turbulência de mercado condicionam a capacidade da empresa em otimizar e rentabilizar a estrutura existente e pode forçá-la a sacrificar as suas margens de lucro para se manter competitiva"[Correia et al., 2011:10]. Segundo os dados do INE[1], a dispersão das empresas em Portugal é significativa. Ao analisar-se o rácio número de empresas/população ativa verifica-se uma distribuição homogénea pelo território nacional, como se verifica na tabela 2. Algumas regiões do interior apresentam o referido rácio com valores superiores à média nacional, como é o caso de: Alto de Trás-os-Montes com um rácio de 0,2533 para as empresas em geral e 0,2477 para as microempresas; Beira Interior Norte com 0,2327 para as empresas em geral e 0,2289 para as microempresas; e Beira Interior Sul com 0,2327 para as empresas em geral e 0,2265 para as microempresas. Podemos assim concluir que o número de empresas por habitante não apresenta diferenças significativas no território nacional, com exceção das grandes empresas.

Tabela 2: Distribuição das empresas nacionais por população ativa

Região	**Total**		**$<$ 10 PT**		**10 - 49 PT**		**50 - 249 PT**		**$\geq$ 250 PT**	
	Distr. Empr.	**Empr./ Pop.**	**Distr. Empr.**	**Empr./ Pop.**	**Distr. Empr.**	**Empr./ Pop.**	**Distr. Empr.**	**Empr./ Pop.**	**Distr. Empr.**	**Empr./ Pop.**
Portugal	100,00%	0,2278	95,81%	0,2182	3,61%	0,0082	0,51%	0,0012	0,08%	0,0002
Continente	95,86%	0,2294	91,85%	0,2198	3,45%	0,0083	0,49%	0,0012	0,08%	0,0002
Norte	31,99%	0,2084	30,39%	0,1980	1,39%	0,0091	0,19%	0,0012	0,02%	0,0001
Centro	21,68%	0,2349	20,83%	0,2256	0,75%	0,0081	0,10%	0,0011	0,01%	0,0001
Lisboa	29,69%	0,2417	28,53%	0,2323	0,96%	0,0078	0,16%	0,0013	0,04%	0,0003
Alentejo	7,12%	0,2377	6,90%	0,2304	0,19%	0,0065	0,02%	0,0007	0,00%	0,0001
Algarve	5,39%	0,2789	5,20%	0,2694	0,17%	0,0086	0,02%	0,0008	0,00%	0,0001
Açores	2,25%	0,2238	2,16%	0,2155	0,07%	0,0071	0,01%	0,0010	0,00%	0,0002
Madeira	1,89%	0,1694	1,79%	0,1606	0,09%	0,0077	0,01%	0,0010	0,00%	0,0001

[1]INE http://www.ine.pt/ dados obtidos em 23-04-2013.

3 Aplicação da metodologia DEA

O DEA é uma ferramenta estatística não-paramétrica e recorre a técnicas de programação matemática, possibilitando medir a eficiência na presença de múltiplos inputs e outputs, sem ser necessária a especificação de uma forma funcional, e em que os respetivos pesos são determinados endogenamente [Moreira, 2008]. A metodologia DEA, ao permitir a determinação pelo próprio modelo dos "pesos"a atribuir aos inputs e aos outputs, transforma-se num poderoso instrumento de análise pois que retira o carater subjetivo inerente a outras perspetivas vulgarmente utilizadas, nomeadamente o recurso ao método dos rácios e "score z"ou à análise de regressão [Vaz, 2000]. A medida simples do lucro não permite determinar objetivos realistas ou responder à questão que meios devem ser afetos à atividade, podendo a resposta ser obtida pelo DEA. Na avaliação do desempenho tem-se por base desempenhos observados, e não desempenhos médios ou padrão. As unidades mais eficientes definem uma fronteira em relação à qual se medem as eficiências relativas das outras unidades. Em termos matemáticos a determinação da medida do desempenho será definida através do recurso à programação linear dado que estamos em presença de problemas de otimização. O ponto de partida, do modelo proposto por Charnes, Cooper, e Rhodes, [1978] CCR, consiste em definir os outputs (Y) e inputs (X) que devem ser selecionados, sem especificar, à priori, com que pesos devem ser ponderados, e que entram na definição de medida de eficiência, sendo esta determinada para cada unidade de índice zero por:

$$
\max\left\{ h_{j_o} = \frac{\sum_{r=1}^{s} \mu_r \, y_{rj_o}}{\sum_{i=1}^{m} \nu_i \, x_{ij_o}} \;\middle|\;
\begin{array}{ll}
\dfrac{\sum_{r=1}^{s} \mu_r \, y_{rj}}{\sum_{i=1}^{m} \nu_i \, x_{ij}} \leq 1, & j = 1, \ldots, n, \\
\nu_i > 0, & i = 1, \ldots, m, \\
\mu_r > 0, & r = 1, \ldots, s
\end{array}
\right\} \tag{1}
$$

A primeira questão que se coloca é a da definição ou determinação dos pesos Ur e Vi. A grande inovação do modelo é que este considera os referidos pesos como incógnitas a determinar pelo próprio modelo. A solução ótima, para cada unidade, formará a combinação de pesos capaz de otimizar a sua medida de eficiência. O conjunto das unidades com valores de eficiência iguais a um poderá ser considerado o conjunto de unidades de referência para as restantes unidades de eficiência inferior a um [Vaz, 1995]. A questão fundamental que nos preocupa é a de, em presença de uma grande variedade de fatores que possam influenciar a eficiência de uma organização, como determinarmos uma medida simples que traduza essa mesma eficiência. Os diversos aspetos do desempenho podem ser abordados por recurso a diferentes modelos designadamente: modelo de eficiência custo, modelo de eficiência de mercado e modelo de determinação do potencial, retirando conclusões do conjunto de informações por eles produzidas [Vaz, 2000]. Estes modelos, que têm aplicações diversas, respondem também a necessidades diversas em termos de objetivos e resultados da análise pretendida [Norman & Stoker, 1991]. O modelo de eficiência custo é relevante para apreciar o uso económico de recursos utilizados na obtenção de determinado output. O modelo de eficiência de mercado tem como objetivo específico avaliar o desempenho do gestor e da unidade no passado recente. O modelo de determinação do potencial é especialmente relevante para estabelecer o potencial de negócio da unidade, particularmente se se questiona, por deficiências, a gestão anterior ou o encerramento do estabelecimento. Banker, Charnes e Cooper [1984] definem ainda um modelo de eficiência técnica. Este modelo considera que a maior eficiência é sempre possível pela forma de uma diminuição dos inputs mantendo o nível de outputs, aumento dos outputs mantendo o nível de inputs, ou as duas situações em simultâneo. Enquanto que, no modelo apresentado em (1), rendimentos à escala constantes, o rácio obtido, para minimização de inputs, é igual ao rácio obtido para maximização de outputs, nos rendimentos variáveis à escala esses dois rácios são diferentes, somando-se uma constante ao output.

$$\max\left\{ h_{j_o} = \frac{\sum_{r=1}^{s} \mu_r \; y_{rj_o} + C}{\sum_{i=1}^{m} \nu_i \; x_{ij_o}} \;\right|$$

$$\left.\begin{array}{ll} \dfrac{\sum_{r=1}^{s} \mu_r \; y_{rj}}{\sum_{i=1}^{m} \nu_i \; x_{ij}} \leq 1, & j = 1, \ldots, n, \\ \nu_i > 0, & i = 1, \ldots, m, \\ \mu_r > 0, & r = 1, \ldots, s \end{array}\right\} \tag{2}$$

A constante (C) define rendimentos crescentes, decrescentes ou constantes à escala. Existem muitos outros modelos de análise DEA, os quais podem ser aprofundados no trabalho de Banker, Cooper, Seiford, Thrall, e Zhu [2004].

3.1 Definição da amostra

Para classificar as empresas, quanto à sua dimensão, foram consideradas três variáveis: postos de trabalho, volume de negócios ou o total do ativo líquido, utilizadas pela recomendação da Comissão Europeia de 3 de Abril de 1996 n° 96/280/CE. Ou seja: postos de trabalho inferiores a 10, volume de negócios ou o ativo total líquido inferior a 2.000.000€. As empresas foram selecionadas de entre um conjunto de 170 empresas, sediadas no interior de Portugal, pertencentes a uma base de dados de um estudo de doutoramento. A escolha teve por base ainda as características semelhantes que as empresas têm entre si, quer em termos de atividade económica, quer em termos geográficos. Deste grupo de empresas foram selecionadas 35, tendo como critério serem consideradas microempresas tendo por base os três critérios anteriormente referidos, terem a estrutura jurídica de sociedades e pertencerem à divisão do CAE 47 . A exigência do CAE é fundamental para se garantir a homogeneidade entre as empresas dado que, como refere Moreira [2008], é uma característica muito importante na aplicação da metodologia DEA, uma vez que se trata de uma análise de eficiência relativa e não-paramétrica. Os dados são contabilísticos e foram recolhidos pela declaração IES - Informação Empresarial Simplificada do ano de 2009.

3.2 Escolha dos inputs e outputs

As variáveis utilizadas são vendas e prestação de serviços como outputs e ativo não corrente, custo das mercadorias vendidas e matérias consumidas (CMVMC) e postos de trabalho efetivos (PT) como inputs. A escolha destes outputs e inputs teve como base o recurso a informação contabilística, produzida pelas empresas, que os autores reputam de fiável e que satisfazem as exigências do modelo. Para evitar redundância da informação produzida pelas variáveis, elaborou-se o tabela 3 com as correlações dos inputs e outputs. Como se verifica, todas as correlações são estatisticamente significativas, com um grau de significância de 0,01 ou 0,05 não se verificando uma elevada correlação entre os inputs e os outputs

De seguida indica-se alguma informação relativa às empresas, que constituem a amostra e que definem os outputs e inputs, designadamente: Vendas (Y1); Prestação de Serviços (Y2); Ativo não Corrente (X1); Custo das Mercadorias Vendidas (X2) e número de Postos de Trabalho (X3).

3.3 Modelo de eficiência custo

Assim, de acordo com a metodologia anterior, considera-se para o modelo de eficiência custo, com vista à determinação da eficiente utilização de recursos, os seguintes fatores[2]:

[2]Para análise dos resultados utilizou-se o software livre EMS: Efficiency Measurement System, Version 1.3.

Tabela 3: Correlações das variáveis

		Vendas	Prestação Serviços	Ativo não Corrente	CMVMC	PT
Vendas (Y1)	Correlação de Pearson	1	0,250	,382*	,980**	,437**
	Sig. (2 extremidades)		0,148	0,024	0,000	0,009
	N	35	35	35	35	35
Prestação Serviços (Y2)	Correlação de Pearson	0,250	1	0,195	0,313	0,160
	Sig. (2 extremidades)	0,148		0,261	0,067	0,359
	N	35	35	35	35	35
Ativo não Corrente (X1)	Correlação de Pearson	,382*	0,195	1	,374*	,348*
	Sig. (2 extremidades)	0,024	0,261		0,027	0,041
	N	35	35	35	35	35
CMVMC (X2)	Correlação de Pearson	,980**	0,313	,374*	1	,379*
	Sig. (2 extremidades)	0,000	0,067	0,027		0,025
	N	35	35	35	35	35
PT (X3)	Correlação de Pearson	,437**	0,160	,348*	,379*	1
	Sig. (2 extremidades)	0,009	0,359	0,041	0,025	
	N	35	35	35	35	35

*. A correlação é significativa no nível 0,05 (2 extremidades).
**. A correlação é significativa no nível 0,01 (2 extremidades).

Tabela 4: Estatística Descritiva

	N	Mínimo	Máximo	Média	Desvio padrão
Vendas (€) (Y1)	35	3.667,00	990.381,00	299.276,91	271.831,17
Prestação Serviços (€) (Y2)	35	0,00	55.666,26	4.147,20	11.688,80
Ativo não Corrente (€) (X1)	35	379,00	372.699,00	59.343,77	77.337,75
CMVMC (€) (X2)	35	2.001,50	959.053,68	239.282,88	246.820,20
PT (unid.) (X3)	35	1	6	2,97	1,40
N válido	35				

Outputs:

Y1 – Vendas, em euros. Representa a venda de mercadorias e de produtos [C. B. Vaz, Camanho, & Guimarães, 2009].

Y2 – Prestação de Serviços, em euros.

Inputs:

X1 – Ativo não corrente, em euros. Representa o ativo não corrente da empresa líquido de depreciações e amortizações [Kao & Hwang, 2010].

X2 – Custo das mercadorias vendidas e das matérias consumidas em euros.

X3 – Postos de Trabalho em quantidade. Representa os postos de trabalhos remunerados e não remunerados afetos à empresa [Moreira, 2008].

O software utilizado no tratamento da informação permite-nos, para a minimização de inputs ou maximização dos outputs, com rendimentos constantes à escala e rendimentos variáveis à escala, a obtenção dos resultados para os dois modelos. O modelo seguido, tem a estrutura convexa, distância radial com orientação no input.

4 Resultados

4.1 Homogeneidade das Empresas

As empresas constantes da amostra podem considerar-se homogéneas, dado apresentarem atividade idêntica, CAE 47, e usarem meios similares no seu processo produtivo. Para além disso, procura-se comprovar esta homogeneidade analisando o rácio CMVMC/Vendas e Ativo não Corrente/Vendas que refletem, no primeiro caso, o contributo para os resultados designado de complementar da margem bruta e, no segundo caso, a estrutura do ativo não corrente por unidade vendida. Apenas são trabalhados estes dois rácios para permitir a representação gráfica. Pela análise do gráfico 1 verifica-se que, a maior parte das empresas, se encontram concentradas, com exceção das DMUs 1, 10 e 20 e a DMU 26 que não consta no gráfico. No entanto, estas empresas foram mantidas para efeitos de análise, uma vez que, embora apresentem resultados diferentes nada justifica que não sejam incluídas no modelo. Desta forma, a estrutura produtiva das empresas é similar, porquanto para produção de uma unidade de output 1 são consumidas idênticas combinações de input 1 e input 2.

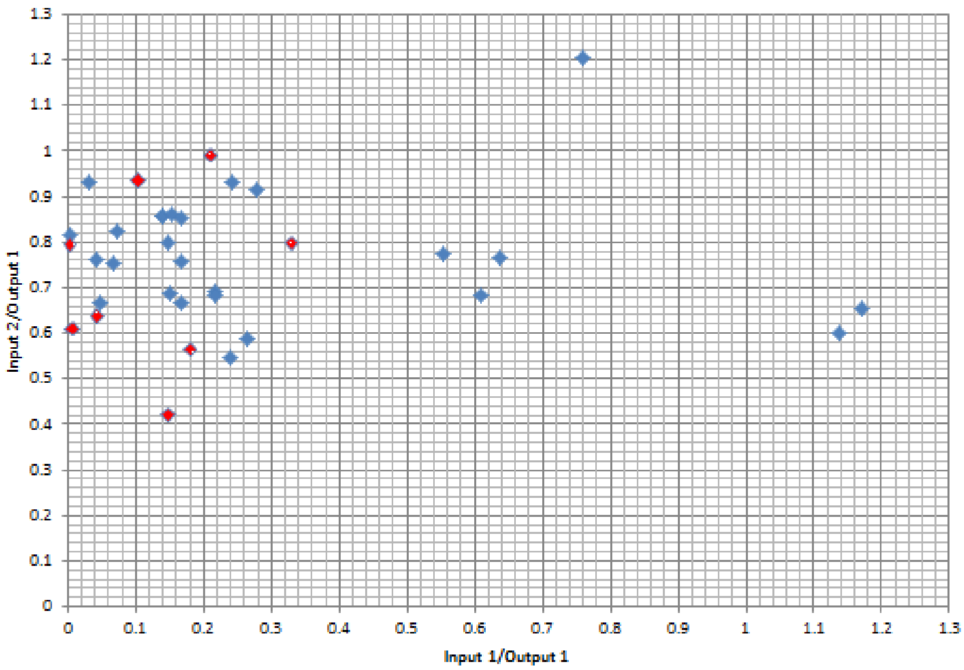


Figura 1: Homogeneidade das Empresas

4.2 Rendimentos constantes à escala

Para efeitos de análise, importa detetar as áreas de eficiência. Para tal procede-se à análise dos outputs e inputs virtuais calculados na tabela 5. Trata-se de determinar o produto dos outputs e inputs pelos respetivos pesos. Na tabela 5 a coluna "Benchmarks"indica, para cada unidade ineficiente, as unidades que lhe servem de referência ou padrão, isto é, relativamente às quais ela apresenta um grau mais baixo de realização, para condições idênticas. O sentido que deve ser atribuído às unidades de referência de uma unidade ineficiente é o de que com o mesmo conjunto de pesos que maximizam a função objetivo dessa unidade e que a apontam como ineficiente, as unidades de referência se apresentam eficientes. Os pesos ν_i e μ_r são os valores que otimizam a função definida para cada unidade. São os parâmetros que para a unidade em avaliação lhe são mais favoráveis. Os pesos, determinados endogenamente, são obtidos para que a unidade respetiva se apresente na sua melhor perspetiva.

Na tabela 5, um valor particular de μ_r e ν_i representa em que medida o output referido contribui para o seu valor de eficiência, quer a unidade seja ou não eficiente, identificando portanto uma área particular de eficiência. Um valor particular de $\nu_i x_i$ indica a melhor área de eficiência relativa da unidade na utilização dos recursos. O sentido que deve ser atribuído às unidades de referência de uma unidade ineficiente é o de que, com o mesmo conjunto de pesos que maximizam a função objetivo dessa unidade e que a apontam como ineficiente, as unidades indicadas como de referência apresentam-se como eficientes [Vaz, 2000]. Alguns valores dos pesos apresentam valor zero. A interpretação deste facto é diferente consoante esse peso diz respeito a uma unidade eficiente ou a uma unidade ineficiente. Assim, numa unidade eficiente o valor zero do peso de uma variável significa que a unidade em causa não pode ser padrão de eficiência no valor apresentado por essa variável. Uma unidade eficiente é elemento de referência apenas nos valores atingidos simultaneamente nas variáveis em que os pesos são diferentes de zero. Para as unidades ineficientes o peso zero significa que o desempenho dessa unidade apresenta um desvio no valor atingido por essa variável. Isto porque, uma unidade só é eficiente, relativamente às do grupo mais ou menos homogéneo a que pertence, se, atingindo o valor um na maximização da função objetivo os desvios nos inputs e nos outputs forem simultaneamente todos nulos [Vaz, 2000]. Como exemplo de DMU claramente eficientes veja-se na tabela 6 as DMUs 22 e 35.

4.3 Rendimentos variáveis à escala

Verifica-se, pela análise do tabela 7 que a medida de eficiência obtida, considerando rendimentos variáveis à escala, é para as diversas unidades sempre superior, em sentido lato, à obtida pelo recurso à mesma metodologia considerando rendimentos constantes à escala.

Tabela 5: Inputs e Outputs virtuais com Rendimentos Constantes à Escala

DMU	Score	X1 IV	X2 IV	X3 IV	Y1 OV	Y2 OV	Benchmarks
1	80,40%	0,0000	0,7824	0,2176	1,0000	0,0000	14 (0,0604) 22 (0,3114)
2	64,50%	0,0000	0,8253	0,1747	1,0000	0,0000	14 (0,3090) 22 (0,8265)
3	100,00%	0,0000	0,4900	0,5100	1,0000	0,0000	1
4	100,00%	0,0000	0,7823	0,2177	0,8915	0,1085	0
5	100,00%	0,2006	0,0000	0,7994	1,0000	0,0000	1
6	64,50%	0,0000	0,8253	0,1747	1,0000	0,0000	14 (0,3090) 22 (0,8265)
7	100,00%	0,0000	0,0000	1,0000	0,3634	0,6366	1
8	100,00%	1,0000	0,0000	0,0000	0,0268	0,9732	0
9	69,75%	0,0000	0,8966	0,1034	1,0000	0,0000	14 (0,3848) 22 (0,1203)
10	83,34%	0,0000	1,0000	0,0000	0,0000	1,0000	35 (0,8525)
11	57,45%	0,2083	0,7917	0,0000	1,0000	0,0000	14 (0,0836) 22 (0,9047)
12	85,70%	0,0532	0,0000	0,9468	1,0000	0,0000	14 (1,1701) 23 (0,2583)
13	77,75%	0,0000	0,7944	0,2056	1,0000	0,0000	14 (0,1999) 22 (0,8664)
14	100,00%	0,0186	0,9157	0,0657	1,0000	0,0000	22
15	84,95%	0,0853	0,9147	0,0000	1,0000	0,0000	14 (0,0400) 22 (0,0744)
16	61,85%	0,0000	1,0000	0,0000	1,0000	0,0000	22 (0,4256)
17	73,67%	0,0000	0,7337	0,2663	1,0000	0,0000	14 (0,0308) 22 (0,3222)
18	70,39%	0,0000	0,7232	0,2768	1,0000	0,0000	14 (0,1019) 22 (1,2548)
19	76,82%	0,0000	0,7813	0,2187	1,0000	0,0000	14 (0,1140) 22 (0,5973)
20	76,58%	0,0000	0,6827	0,3173	1,0000	0,0000	14 (0,0687) 22 (1,8113)
21	71,75%	0,0000	1,0000	0,0000	1,0000	0,0000	22 (2,0445)
22	100,00%	0,0875	0,9125	0,0000	1,0000	0,0000	22
23	100,00%	0,1645	0,0000	0,8355	1,0000	0,0000	1
24	86,12%	0,0201	0,2252	0,7547	1,0000	0,0000	3 (0,0481) 7 (0,0898) 14 (1,0425)
25	70,97%	0,0000	0,8109	0,1891	1,0000	0,0000	14 (0,0728) 22 (0,2457)
26	77,38%	0,0000	1,0000	0,0000	1,0000	0,0000	22 (0,0341)
27	75,63%	0,0000	0,9024	0,0976	1,0000	0,0000	14 (0,9034) 22 (0,1575)
28	73,39%	0,1062	0,8938	0,0000	1,0000	0,0000	14 (0,1475) 22 (0,3917)
29	99,17%	0,1242	0,8758	0,0000	1,0000	0,0000	5 (0,6107) 14 (0,0280)
30	68,12%	0,0000	0,8359	0,1641	1,0000	0,0000	14 (0,2746) 22 (0,6099)
31	63,83%	0,0000	0,8842	0,1158	1,0000	0,0000	14 (0,4496) 22 (0,2831)
32	76,79%	0,0000	0,8862	0,1138	1,0000	0,0000	14 (0,9254) 22 (0,5316)
33	73,11%	0,0000	0,8229	0,1771	1,0000	0,0000	14 (0,1706) 22 (0,4752)
34	78,36%	0,0000	0,6388	0,3612	1,0000	0,0000	14 (0,0065) 22 (1,1656)
35	100,00%	0,4209	0,5791	0,0000	0,0000	1,0000	1

Tabela 6: Pesos dos Inputs e Outputs com Rendimentos Constantes à Escala

DMU	Score	X1 IW	X2 IW	X3 IW	Y1 OW	Y2 OW	Benchmarks
1	80,40%	0,0000	0,0000	0,2176	0,0000	0,0000	14 (0,0604) 22 (0,3114)
2	64,50%	0,0000	0,0000	0,0437	0,0000	0,0000	14 (0,3090) 22 (0,8265)
3	100,00%	0,0000	0,0000	0,5100	0,0000	0,0000	1
4	100,00%	0,0000	0,0000	0,0544	0,0000	0,0000	0
5	100,00%	0,0005	0,0000	0,7994	0,0000	0,0000	1
6	64,50%	0,0000	0,0000	0,0437	0,0000	0,0000	14 (0,3090) 22 (0,8265)
7	100,00%	0,0000	0,0000	0,3333	0,0000	0,0000	1
8	100,00%	0,0001	0,0000	0,0000	0,0000	0,0000	0
9	69,75%	0,0000	0,0000	0,0517	0,0000	0,0000	14 (0,3848) 22 (0,1203)
10	83,34%	0,0000	0,0000	0,0000	0,0000	0,0000	35 (0,8525)
11	57,45%	0,0000	0,0000	0,0000	0,0000	0,0000	14 (0,0836) 22 (0,9047)
12	85,70%	0,0000	0,0000	0,1894	0,0000	0,0000	14 (1,1701) 23 (0,2583)
13	77,75%	0,0000	0,0000	0,0685	0,0000	0,0000	14 (0,1999) 22 (0,8664)
14	100,00%	0,0000	0,0000	0,0219	0,0000	0,0000	22
15	84,95%	0,0001	0,0000	0,0000	0,0000	0,0000	14 (0,0400) 22 (0,0744)
16	61,85%	0,0000	0,0000	0,0000	0,0000	0,0000	22 (0,4256)
17	73,67%	0,0000	0,0000	0,2663	0,0000	0,0000	14 (0,0308) 22 (0,3222)
18	70,39%	0,0000	0,0000	0,0692	0,0000	0,0000	14 (0,1019) 22 (1,2548)
19	76,82%	0,0000	0,0000	0,1093	0,0000	0,0000	14 (0,1140) 22 (0,5973)
20	76,58%	0,0000	0,0000	0,0635	0,0000	0,0000	14 (0,0687) 22 (1,8113)
21	71,75%	0,0000	0,0000	0,0000	0,0000	0,0000	22 (2,0445)
22	100,00%	0,0000	0,0000	0,0000	0,0000	0,0000	22
23	100,00%	0,0000	0,0000	0,2785	0,0000	0,0000	1
24	86,12%	0,0000	0,0000	0,1887	0,0000	0,0000	3 (0,0481) 7 (0,0898) 14 (1,0425)
25	70,97%	0,0000	0,0000	0,1891	0,0000	0,0000	14 (0,0728) 22 (0,2457)
26	77,38%	0,0000	0,0005	0,0000	0,0003	0,0000	22 (0,0341)
27	75,63%	0,0000	0,0000	0,0244	0,0000	0,0000	14 (0,9034) 22 (0,1575)
28	73,39%	0,0000	0,0000	0,0000	0,0000	0,0000	14 (0,1475) 22 (0,3917)
29	99,17%	0,0003	0,0000	0,0000	0,0000	0,0000	5 (0,6107) 14 (0,0280)
30	68,12%	0,0000	0,0000	0,0547	0,0000	0,0000	14 (0,2746) 22 (0,6099)
31	63,83%	0,0000	0,0000	0,0386	0,0000	0,0000	14 (0,4496) 22 (0,2831)
32	76,79%	0,0000	0,0000	0,0228	0,0000	0,0000	14 (0,9254) 22 (0,5316)
33	73,11%	0,0000	0,0000	0,0885	0,0000	0,0000	14 (0,1706) 22 (0,4752)
34	78,36%	0,0000	0,0000	0,1204	0,0000	0,0000	14 (0,0065) 22 (1,1656)
35	100,00%	0,0000	0,0000	0,0000	0,0000	0,0000	1

Tabela 7: Inputs e Outputs virtuais com Rendimentos Variáveis à Escala

DMU	Score	X1 IV	X2 IV	X3 IV	Y1 OV	Y2 OV	Benchmarks
1	100,00%	0,0000	0,4173	0,5827	1,0000	0,0000	6
2	65,49%	0,0000	0,9593	0,0407	1,0000	0,0000	4 (0,1800) 14 (0,2598) 22 (0,5602)
3	100,00%	0,0000	0,0000	1,0000	1,0000	0,0000	2
4	100,00%	0,0000	1,0000	0,0000	0,9538	0,0462	8
5	100,00%	1,0000	0,0000	0,0000	1,0000	0,0000	1
6	65,49%	0,0000	0,9593	0,0407	1,0000	0,0000	4 (0,1800) 14 (0,2598) 22 (0,5602)
7	100,00%	0,0472	0,0685	0,8844	0,3823	0,6177	0
8	100,00%	1,0000	0,0000	0,0000	0,3045	0,6955	0
9	76,52%	0,0062	0,6078	0,3860	1,0000	0,0000	1 (0,0671) 3 (0,1889) 14 (0,2652) 17 (0,4789)
10	83,65%	0,0000	1,0000	0,0000	0,0000	1,0000	26 (0,1475) 35 (0,8525)
11	57,48%	0,2369	0,7631	0,0000	1,0000	0,0000	14 (0,0830) 15 (0,0130) 22 (0,9040)
12	100,00%	0,9991	0,0005	0,0004	1,0000	0,0000	0
13	78,54%	0,0000	0,9506	0,0494	1,0000	0,0000	4 (0,0907) 14 (0,1747) 22 (0,7346)
14	100,00%	1,0000	0,0000	0,0000	1,0000	0,0000	13
15	100,00%	0,3699	0,6301	0,0000	1,0000	0,0000	3
16	75,06%	0,1996	0,5069	0,2935	1,0000	0,0000	15 (0,2380) 17 (0,1439) 22 (0,2632) 26 (0,3549)
17	100,00%	0,1644	0,2527	0,5830	1,0000	0,0000	6
18	75,38%	0,0260	0,9740	0,0000	1,0000	0,0000	4 (0,3850) 14 (0,0095) 22 (0,6055)
19	81,67%	0,0043	0,6296	0,3661	1,0000	0,0000	1 (0,1691) 14 (0,0962) 17 (0,2938) 22 (0,4410)
20	89,68%	0,0000	1,0000	0,0000	1,0000	0,0000	4 (0,5696) 22 (0,4304)
21	89,68%	0,0000	1,0000	0,0000	1,0000	0,0000	4 (0,4812) 22 (0,5188)
22	100,00%	0,0723	0,8173	0,1104	1,0000	0,0000	15
23	100,00%	0,2336	0,0000	0,7664	1,0000	0,0000	0
24	100,00%	0,0000	1,0000	0,0000	1,0000	0,0000	1
25	100,00%	0,0000	0,0000	1,0000	1,0000	0,0000	1 (0,2347) 3 (0,0705) 5 (0,0224) 17 (0,2939) 26 (0,2807) 29 (0,0979)
26	100,00%	0,0000	1,0000	0,0000	1,0000	0,0000	3
27	75,86%	0,0000	0,9787	0,0213	1,0000	0,0000	4 (0,0754) 14 (0,8835) 22 (0,0411)
28	75,05%	0,1230	0,8770	0,0000	1,0000	0,0000	14 (0,1265) 15 (0,5184) 22 (0,3551)
29	100,00%	0,2817	0,7183	0,0000	1,0000	0,0000	1
30	69,17%	0,0013	0,7094	0,2893	1,0000	0,0000	1 (0,0107) 14 (0,2703) 17 (0,1845) 22 (0,5345)
31	65,44%	0,0000	0,7957	0,2043	1,0000	0,0000	1 (0,4630) 14 (0,4262) 22 (0,1109)
32	80,19%	0,0000	1,0000	0,0000	1,0000	0,0000	14 (0,9421) 24 (0,0579)
33	78,16%	0,0032	0,6888	0,3079	1,0000	0,0000	1 (0,1457) 14 (0,1529) 17 (0,4441) 22 (0,2574)
34	83,66%	0,0000	1,0000	0,0000	1,0000	0,0000	4 (0,0949) 22 (0,9051)
35	100,00%	0,4414	0,5586	0,0000	0,0000	1,0000	1

Também o número de unidades eficientes aumentou como seria de esperar. Esta razão prende-se com o facto de o modelo com rendimentos variáveis entrar em linha de conta com a escala. A medida de eficiência obtida pelo modelo em rendimentos constantes é em termos médios de 81,21 % enquanto que, quando se utiliza rendimentos variáveis esse valor é 87,60%. Esta situação vem em resultado do facto de mais empresas atingirem 100% de eficiência quando se usa o modelo variáveis à escala.

5 Conclusão

O objetivo inicial deste trabalho é determinar o desempenho das microempresas, do setor do comércio a retalho, ao determinar a eficiência custo, com a preocupação de sintetizar as informações obtidas facilitando uma mais fácil compreensão e análise dos resultados. As empresas nacionais, especialmente as microempresas, enfrentam uma concorrência cada vez mais forte, fruto da conjuntura económica nacional. As questões da eficiência empresarial tornam-se cada vez mais importantes, uma vez que as empresas ineficientes poderão não sobreviver por muito mais tempo. Uma série de estudos têm aplicado a metodologia DEA para diversos setores: setor bancário [Miller & Noulas, 1996]; setor automóvel [Amirteimoori, 2011]; setor do comércio a retalho [Vaz et al., 2009]. Estes estudos à semelhança de muitos outros analisam empresas de grande dimensão. Atendendo às caraterísticas das empresas em estudo, considera-se o presente trabalho relevante, dado avaliar um grupo de empresas pouco estudado. Não se pretende ser exaustivo na análise da informação obtida, mas tão só evidenciar algumas das potencialidades da metodologia DEA como instrumento de gestão e avaliação de unidade homogéneas. O facto de a metodologia DEA considerar outputs e inputs observados em unidades homogéneas e estabelecer a medida simples da eficiência com base nos resultados e consumo de recursos verificados, revela-se de grande alcance em termos da gestão corrente e da motivação das pessoas envolvidas induzindo credibilidade ao modelo. Ao ser apresentada, para cada unidade ineficiente, a unidade ou unidades de referência, consideradas eficientes é importante para permitir adotar as melhores práticas de gestão. Verificou-se que as empresas em análise apresentam abundância de recursos, nomeadamente ativo não corrente. Uma possível explicação, para este facto, poderá estar relacionada com o tipo de gestão. Neste tipo de empresas a gestão normalmente não é profissional, sendo em muitos casos consideradas empresas familiares [Fernandes & Ussmane, 2012]. Conclui-se, tal como em outros trabalhos, (e.g. Vaz, 1995) que, em ambiente incerto de concorrência acrescida e global, as organizações públicas ou privadas, para garantirem a eficiência de cada uma das unidades e por esta via vantagens relativas, deverão estar munidas dos instrumentos analíticos que lhes possibilitem o diagnóstico correto que, de forma simples e objetiva, apontem alguns caminhos ou metas com vista a melhorarem o seu desempenho e concomitantemente a sua posição competitiva. Para tal, o recurso à metodologia DEA parece ser um instrumento de particular interesse, não deixando de deverem considerar outras abordagens do desempenho organizacional. Admite-se, no entanto, que a leitura dos resultados deva ser feita com cuidado, não permitindo generalizações uma vez que são apenas 35 empresas. Assim, para permitir uma generalização dos dados propõe-se a elaboração de um trabalho quantitativo, com uma amostra alargada de empresas.

Referências

[Amirteimoori, 2011] Amirteimoori, A. (2011). A DEA two-stage decision processes with shared resources. *Central European Journal of Operations Research*, 21(1), 141–151.

[Banker et al., 1984] Banker, R. D., Charnes, a., & Cooper, W. W. (1984). Some Models for Estimating Technical and Scale Inefficiencies in Data Envelopment Analysis. *Management Science*, 30(9), 1078–1092.

[Banker et al., 2004] Banker, Rajiv D., Cooper, W. W., Seiford, L. M., Thrall, R. M., & Zhu, J. (2004). Returns to scale in different DEA models. *European Journal of Operational Research*, 154(2), 345–362.

[Charnes et al. , 1978] Charnes, A., Cooper, W., & Rhodes, E. (1978). Measuring the efficiency of decision making units. *European Journal of Operational Research*, 2(6), 429–444.

[Correia et al., 2011] Correia, R., Teixeira, M., & Rebelo, J. (2011). Impacto do marketing e orientação para o mercado na performance económico-financeira nas PME do interior norte de Portugal. *In XXI Jornadas Hipano-Lusas de Gestión Científica*, 1–12.

[Fernandes & Ussmane, 2012] Fernandes, A. B., & Ussmane, A. M. (2012). Capacidade de Aprendizagem das Empresas Familiares. *XXII Jornadas Luso-Espanholas de Gestão Científica*, 1–21.

[Hu & Schive, 1998] Hu, M.-W., & Schive, C. (1998). The changing competitiveness of Taiwan's manufacturing SMEs. *Small Business Economics*, (11) 315–326.

[Jenkins, 2006] Jenkins, H. (2006). Small Business Champions for Corporate Social Responsibility. *Journal of Business Ethics*, 67(3) 241–256.

[Kao & Hwang, 2010] Kao, C., & Hwang, S. N. (2010). Efficiency measurement for network systems: IT impact on firm performance. *Decision Support Systems*, 48(3) 437–446.

[Miller et al., 1996] Miller, S. M., & Noulas, A. G. (1996). The technical efficiency of large bank production. *Journal of Banking & Finance*, 20(3) 495–509.

[Moreira, 2008] Moreira, S. (2008). Análise da eficiência dos hospitais-empresa: uma aplicação da data envelopment analysis. *Boletim Económico - Banco de Portugal*, (Primavera) 127–150.

[Norman & Stoker, 1991] Michael, N., & Stoker, B. (1991). Data Envelopment Analysis: The Assessment of Performance. *NY , EUA: John Wiley & Sons, Ltd.*

[Vaz, et al., 2009] Vaz, C. B., Camanho, A. S., & Guimarães, R. C. (2009). Avaliação de desempenho de lojas de retalho utilizando um modelo de Data Envelopment Analysis multi-nível. *14.o Congresso da APDIO*, 227–234.

[Vaz, 1995] Vaz, M. A. (1995). Eficiência Relativa e Output Potencial: Uma aplicação DEA (Data Envelopment Analysis). *Universidade do Porto.*

[Vaz, 2000] Vaz, M. A. (2000). Aplicação DEA à Análise da Eficiência de Estações de Correios. Casos de Aplicação da Investigação Operacional. *McGraw-Hil., pp. 365-388.*

[Vieira & Novo, 2010] Vieira, E. S., & Novo, A. J. A estrutura de capital das PME: evidência no mercado português. *Centro de Investigação GOVCOPP, pp. 1-19.*

Incorporação da resistência ao fogo na gestão florestal à escala da paisagem: uma aplicação à Mata Nacional de Leiria

Ferreira, L.*, Constantino, M.†, Borges, J. G.‡, Garcia-Gonzalo, J.‡

* *Instituto Politécnico de Leiria, Escola Superior de Tecnologia e Gestão, Centro de Investigação Operacional*
E-mail: liliana.ferreira@ipleiria.pt

† *Faculdade de Ciências, DEIO, Universidade de Lisboa, Centro de Investigação Operacional*
E-mail: mfconstantino@fc.ul.pt

‡ *Instituto Superior de Agronomia, Universidade Técnica de Lisboa, Centro de Estudos Florestais.*
E-mail: joseborges@isa.utl.pt; jordigarcia@isa.utl.pt

Resumo

Será apresentada uma proposta de um modelo de programação inteira mista para a gestão florestal à escala da paisagem. São incorporados índices de resistência ao fogo para cada povoamento, considerando as suas características e também o seu contexto espacial, refletindo assim a contribuição dos seus vizinhos para o aumento ou diminuição da sua resistência. O modelo pretende maximizar o valor esperado do solo da floresta e tentar assegurar um nível mínimo de resistência para a mesma, escolhendo para cada povoamento uma prescrição a aplicar durante um horizonte de planeamento. A Mata Nacional de Leiria foi utilizada como caso de estudo.

Palavras chave: risco, incêndios florestais, programação inteira mista, gestão florestal.

1 Introdução

Os incêndios florestais de grande dimensão têm aumentado substancialmente, durante as últimas décadas, em todos os países mediterrânicos, nomeadamente em Portugal (*e.g.* [Alexandrian *et al.*, 2000], [Pereira *et al.*, 2006]).

Devido às alterações climáticas os incêndios tendem a aumentar ainda mais, o que terá consequências avultadas nas perdas de madeira. Este cenário relembra a importância da integração do risco de incêndios florestais no processo de tomada de decisão na gestão das florestas portuguesas. No entanto, continuam a não existir estratégias disponíveis para ajudar os gestores florestais portugueses a incorporar o risco de incêndio no planeamento de medidas de gestão ([Borges and Uva, 2006]).

Os modelos de gestão florestal à escala da paisagem pressupõem uma floresta composta por diversos povoamentos, para os quais se pretendem escolher as prescrições mais adequadas a adotar, num determinado horizonte de planeamento, de modo a atingir os objetivos pretendidos. Ao incorporar o risco de incêndio nos modelos de gestão florestal à escala da paisagem, não se poderá ter apenas em conta o risco de cada povoamento arder isoladamente, uma vez que os povoamentos são contíguos e existe a possibilidade de um incêndio, que percorra um determinado povoamento, se propagar para outro povoamento da floresta, por força de inúmeros fatores como a adjacência, o vento, o declive, a altitude, entre outros.

Para integrar o risco na gestão florestal à escala da paisagem foi definido um indicador designado por *índice de resistência ajustada*, que pretende medir a resistência dos povoamentos ao fogo. Para calcular este indicador foram consideradas as probabilidades de ocorrência de incêndio para cada povoamento e a proporção esperada de árvores mortas. O nível de resistência ao fogo pretende ser uma aproximação da capacidade de sobrevivência dos povoamentos. Este indicador depende não só das características específicas de cada povoamento relacionadas, por exemplo, com a forma, com a dimensão ou com o seu estado, mas também da influência dos vizinhos no risco de propagação dos fogos. A ideia de que os vizinhos podem influenciar o nível de resistência de um povoamento é crucial para o desenvolvimento de um modelo de gestão florestal à escala da paisagem que inclui o risco de incêndio (*e.g.* [Agee *et al.*, 2000], [Acuna *et al.*, 2010], [González and Pukkala, 2010], [Konoshima *et al.*, 2010]). Se os vizinhos são muito resistentes podem contribuir para aumentar a resistência ajustada de um povoamento. Por outro lado, vizinhos com capacidade de sobrevivência baixa podem reduzir o valor da resistência ajustada de um povoamento.

Será apresentado um modelo de programação inteira mista que inclui o índice de resistência ajustada de forma a incorporar o risco de incêndio na gestão de uma floresta composta por diversos povoamentos. Este modelo permite escolher, para cada povoamento, uma prescrição composta por várias intervenções de gestão que são realizadas ao longo de diversos períodos de um determinado horizonte de planeamento. O contexto espacial será tido em conta, uma vez que se considera que as intervenções efetuadas em cada povoamento poderão afetar não só o risco de incêndio no próprio povoamento, mas também o risco de incêndio nos povoamentos vizinhos.

Serão ainda apresentados os principais resultados de um caso de estudo que traduz a aplicação do modelo proposto à Mata Nacional de Leiria.

2 Indicador de resistência ao fogo

Considerando uma floresta composta por I povoamentos, pretende-se determinar a prescrição mais adequada a adotar em cada povoamento $i,\ i \in \mathcal{I} = \{1, ..., I\}$, tendo em conta o contexto espacial da floresta. Neste sentido, medidas de gestão efetuadas num dado povoameno poderão afetar não só o risco de incêndio desse mesmo povoamento mas também o risco de incêndio dos povoamentos vizinhos. Fatores como a adjacência, o vento, o declive, a altitude, entre outros, poderão influenciar a propagação dos fogos de uns povoamentos para os outros. Medidas de proteção associadas ao tratamento de combustível também podem modificar o comportamento de potenciais incêndios, uma vez que podem alterar a distribuição espacial da biomassa (*e.g.* [Acuna *et al.*, 2010], [Konoshima *et al.*, 2010]).

A localização de povoamentos mais resistentes ou mais vulneráveis terá então um impacto na resistência global da floresta(*e.g.* [González and Pukkala, 2010]), razão que motivou o desenvolvimento de um índice de resistência. A resistência ao fogo, num dado período t, para o povoamento i, é entendida como uma aproximação da proporção de árvores do povoamento que sobrevivem ao fogo durante esse período. O índice de resistência para a floresta é calculado como uma média, ponderada pela área, dos índices de resistência dos diferentes povoamentos.

2.1 Indicador de resistência ajustada

A ideia de que os povoamentos vizinhos podem modificar o índice de resistência de um dado povoamento constituiu a base do desenvolvimento do indicador de resistência ajustada, apresentado de seguida.

Para cada povoamento i, consideram-se dois tipos de resistência: a resistência específica que depende apenas das características do próprio povoamento e a resistência ajustada que altera o valor da resistência específica de acordo com a resistência dos vizinhos e do risco de um incêndio se propagar para i.

A resistência específica do povoamento i, no período t, é denotada por R_{it}. Este indicador corresponde à percentagem esperada de árvores sobreviventes ao fogo, num dado povoamento, num determinado período, e é calculado com base nas probabilidades anuais de ocorrência de incêndio de cada povoamento, nas probabilidades de mortalidade em caso de incêndio e na proporção de árvores mortas em caso de mortalidade.

O índice de resistência ajustada do povoamento i, no período t, é designado por RA_{it}. Este indicador incorpora informação sobre a resistência ao fogo dos seus vizinhos e sobre o risco de propagação dos incêndios. A resistência ajustada para cada povoamento i, em cada período t, é calculada de acordo com a seguinte equação:

$$RA_{it} = R_{it} + (1 - w_i) \sum_{s \in \mathcal{V}\{i\}} \alpha_{is} \left(RA_{st} - R_{it} \right), \quad i \in \mathcal{I},\ t \in \mathcal{T} \tag{1}$$

com,

$t \in \mathcal{T} = \{1, ..., T\}$ e identifica o período do horizonte de planeamento, com T igual ao número total de períodos;

R_{it} = resistência específica do povoamento i no período t; $R_{it} \in [0, 1]$;

$(1 - w_i)$ = peso a atribuir à influência dos vizinhos na resistência do povoamento i; $(1 - w_i) \in [0, 1]$; w_i reflete a dimensão e a forma do povoamento i; assume-se que $w_i > 0$;

α_{is} = mede a possibilidade de um incêndio que ocorre no povoamento s se propague para o povoamento i. Fatores como a fronteira comum, o declive, a orientação, a existência de barreiras poderão ser considerados no cálculo destes pesos; $0 \leq \alpha_{is} \leq 1$ e $\sum\limits_{s \in \mathcal{V}\{i\}} \alpha_{is} \leq 1$, onde $\mathcal{V}\{i\}$ é o conjunto dos vizinhos do povoamento i.

A contribuição do próprio povoamento para a sua resistência ajustada é maior quanto maior for a sua área, uma vez que se assume que quanto maior é um povoamento menor é a influência dos vizinhos. A contribuição dos vizinhos também é mais elevada quanto maior o perímetro do povoamento i. Cada povoamento s, vizinho do povoamento i, tem uma contribuição diferente consoante a possibilidade de um fogo em s se propagar para i. Caso a contribuição dos vizinhos para a resistência do povoamento em estudo seja nula, a resistência ajustada será igual à resistência específica do povoamento.

A diferença $RA_{st} - R_{it}$ corresponde ao ajustamento que a resistência específica do povoamento i (R_{it}) sofre, de acordo com a contribuição do vizinho s para a resistência de i. Esta diferença pode ser positiva no caso da resistência ajustada de s ser superior à resistência específica do povoamento i, contribuindo para um aumento da resistência ajustada de i, e pode ser negativa no caso da resistência ajustada de s ser menor do que a resistência específica de i, diminuindo a resistência ajustada deste aos fogos.

A resistência da floresta aos incêndios depende da resistência de cada um dos povoamentos que a constituem. Deste modo, a cada povoamento i, a cada período de tempo t e a cada possível prescrição k, com $k \in \mathcal{K}_i = \{1, ..., K_i\}$, sendo K_i o número total de prescrições possíveis de adotar no povoamentoo i, corresponde um indicador de resistência (r_{ikt}) que será considerado no cálculo da resistência global da floresta.

2.2 Modelo de programação inteira mista

Foi utilizado um modelo de gestão em programação inteira mista (PIM) para avaliar o comportamento do índice de resistência ao fogo. Para além das restrições relacionadas com o índice de resistência, foram adicionadas restrições de regularidade de volume entre os diferentes períodos do horizonte de planeamento e restrições associadas à idade média da floresta em inventário final. Muitas outras restrições poderão ser incorporadas no modelo. O modelo foi definido de acordo com as seguintes condições:

$$Max\ Z = \sum_{i=1}^{I} \sum_{k=1}^{K_i} c_{ik} A_i x_{ik} \tag{2}$$

$$\sum_{k=1}^{K_i} x_{ik} = 1, \quad i \in \mathcal{I} \tag{3}$$

$$\sum_{i=1}^{I} \sum_{k=1}^{K_i} v_{ikt} A_i x_{ik} = V_t, \quad t \in \mathcal{T} \tag{4}$$

$$\begin{aligned} V_t &\geq (1-\lambda) V_{t-1}, \quad t \in \mathcal{T} \setminus \{1\} \\ V_t &\leq (1+\lambda) V_{t-1}, \quad t \in \mathcal{T} \setminus \{1\} \end{aligned} \tag{5}$$

$$\sum_{i=1}^{I} A_i \sum_{k=1}^{K_i} F_{ik} x_{ik} \geq A_F f \tag{6}$$

$$R_{it} = \sum_{k=1}^{K_i} r_{ikt} x_{ik}, \quad i \in \mathcal{I},\ t \in \mathcal{T} \tag{7}$$

$$RA_{it} = R_{it} + (1 - w_i) \sum_{s \in \mathcal{V}\{i\}} \alpha_{is} \left(RA_{st} - R_{it} \right), \quad i \in \mathcal{I},\ t \in \mathcal{T} \tag{8}$$

$$\frac{\sum\limits_{i=1}^{I} A_i RA_{it}}{A_F} \geq RES_t, \quad t \in \mathcal{T} \tag{9}$$

$$x_{ik} \in \{0, 1\}, \quad i \in \mathcal{I},\ k \in \mathcal{K}_i \tag{10}$$

onde,

conjuntos e índices:
i identifica o povoamento, com $i \in \mathcal{I} = \{1, ..., I\}$ e I representa o número total de povoamentos;
k identifica a prescrição adotada para o povoamento i, com $k \in \mathcal{K}_i = \{1, ..., K_i\}$ e K_i representa o número total de prescrições possíveis de adotar no povoamento i;
t identifica o período do horizonte de planeamento, com $t \in \mathcal{T} = \{1, ..., T\}$ e T representa o número total de períodos do horizonte de planeamento;
$\mathcal{V}\{i\}$ = conjunto dos povoamentos vizinhos de i;

parâmetros:
c_{ik} = valor do solo, por hectare, resultante da adoção da prescrição k para gerir o povoamento i;
A_i = área do povoamento i, em hectares;
A_F = área total da floresta, em hectares, $A_F = \sum_{i=1}^{I} A_i$;
r_{ikt} = indicador da resistência específica para o povoamento i no período t, caso seja gerido de acordo com a prescrição k; $r_{ikt} \in [0, 1]$;
RES_t = nível mínimo de resistência ajustada para a floresta, para o período t; $RES_t \in [0, 1]$;
$(1 - w_i)$ = peso atribuído à contribuição dos vizinhos para a resistência do povoamento i;
α_{is} = peso atribuído à contribuição da resistência do povoamento s, vizinho do povoamento i, para a resistência do povoamento i; $\alpha_{is} \in [0, 1]$;
v_{ikt} = volume realizado, por hectare, no período t, no povoamento i, se for adotada a prescrição k;
F_{ik} = idade do povoamento i, no final do horizonte de planeamento, caso seja adotada a prescrição k;
f = limite inferior para a idade média da floresta, no final do horizonte de planeamento;
λ = varição máxima para o volume total realizado entre dois períodos consecutivos;

variáveis:
$$x_{ik} = \begin{cases} 1, & \text{se o povoamento } i \text{ é gerido pela prescrição } k \\ 0, & \text{caso contrário} \end{cases};$$
R_{it} = resistência específica do povoamento i, no período t;
RA_{it} = resistência ajustada do povoamento i, no período t;
V_t = volume total realizado no período t.

A equação (2) exprime o objetivo do estudo de determinar, para cada povoamento, uma prescrição constituída por diversas intervenções, de forma a maximizar o valor esperado do solo da floresta. A garantia de que apenas uma prescrição é escolhida para cada povoamento é dada pelas equações (3).

As equações (4) determinam o nível de volume realizado em cada período, enquanto que a regularidade do volume realizado ao longo dos diferentes períodos é traduzida pelas inequações (5).

A preocupação com o estado da floresta, no final do horizonte de planeamento, é traduzida pela inequação (6) que permite garantir que a idade média do inventário final seja superior ou igual a um limite f previamente estabelecido.

O nível de resistência específica para cada povoamento, em cada período, de acordo com o seu estado e com a prescrição adotada é definido pelas equações (7).

A resistência ajustada é calculada através das equações (8). As inequações (9) impõem um nível mínimo de resistência ajustada para a floresta, para cada período do horizonte de planeamento, ponderando a resistência ajustada de cada povoamento de acordo com a sua área.

3 Caso de estudo

A fim de testar o modelo proposto, foi utilizada a Mata Nacional de Leiria (MNL) como área de estudo. A MNL é maioritariamente ocupada por povoamentos de pinheiro bravo e estende-se por uma área de aproximadamente 11000 hectares que totalizam 539 povoamentos. Destes 539, 393 estão classificados como parcelas de produção de madeira (cerca 8600 hectares).

A aplicação do modelo de PIM tem por base modelos de crescimento para o pinheiro bravo e para os matos e o modelo de silvicultura típico utilizado nesta região. As intervenções podem ser de vários tipos:

- corte raso (CLP), que engloba o corte propriamente dito, uma limpeza de matos e uma nova plantação;

- desbaste (DL), que, para além de uma redução de uma percentagem da área basal, tem também inerente uma limpeza de matos;
- limpeza (L), que pressupõe uma limpeza de matos;
- não intervenção (NI), quando a opção é não realizar qualquer intervenção.

Por exemplo, se se considerarem quatro períodos, quatro possíveis prescrições para um dado povoamento poderão ser as apresentadas na tabela 1.

Tabela 1: Possíveis prescrições, considerando um horizonte de planeamento com 4 períodos.

	Horizonte de planeamento			
Prescrições	$t = 1$	$t = 2$	$t = 3$	$t = 4$
1	CLP	NI	DL	L
2	NI	CLP	L	DL
3	L	NI	CLP	NI
4	DL	DL	DL	CLP

No caso de estudo, foram considerados 4 períodos de 10 anos para o horizonte de planeamento. Deste modo, admitiram-se desbastes obrigatórios de 10 em 10 anos, para povoamentos com idades entre 20 e 50 anos, acompanhados de limpezas de mato em cada um dos períodos. Para povoamentos com mais de 50 ou menos de 20 anos não são obrigatórios desbastes, sendo permitidas prescrições que envolvam intervenções só de limpeza ou, eventualmente, períodos sem intervenções. A idade mínima para realizar um corte raso é de 50 anos e a idade máxima de corte é de 80 anos. Após um corte raso, não é possível efetuar desbaste no período seguinte. Caso a prescrição adotada para o horizonte de planeamento inclua desbastes, estes retiram 20% de área basal, caso esta seja superior a $18m^2$/ha. Se a área basal do povoamento for inferior ou igual a $18m^2$/ha, não será permitida a realização de desbastes.

3.1 Cálculo do indicador de resistência específica (coeficientes $\mathbf{r}_{ikt}$)

Para determinar uma estimativa para a percentagem de árvores sobreviventes no povoamento i, durante o período t, quando é adotada a prescrição k (r_{ikt}), consideraram-se modelos de risco e danos de incêndio desenvolvidos para povoamentos portugueses puros ou mistos e que usam métodos de regressão logística ([Garcia-Gonzalo *et al.*, 2011a] e [Garcia-Gonzalo *et al.*, 2011b]). O modelo de risco permite estimar a probabilidade de ocorrência de incêndio em cada um dos povoamentos, de acordo com o seu estado (idade, área basal, número de árvores, biomassa existente, entre outros fatores). O modelo de danos permite estimar se a ocorrência de um dado incêndio origina ou não mortalidade, fornecendo informação sobre a probabilidade de ocorrência de mortalidade e sobre a proporção de árvores mortas, no caso de se verificar mortalidade.

3.2 Cálculo dos rendimentos associados às prescrições (coeficientes $\mathbf{c}_{ik}$)

O valor dos coeficientes c_{ik}, para cada prescrição k possível de aplicar em cada povoamento i, corresponde ao valor atual dessa prescrição repetida até à perpetuidade. Para determinar esse valor é necessário calcular o valor atual resultante de todas as intervenções que compõem a prescrição.

É necessário gerar as prescrições possíveis a aplicar nos povoamentos. Para cada prescrição é preciso estabelecer quais as intervenções a incluir, ou seja, para cada período, há que definir, por exemplo, se é realizado um corte raso, um desbaste, uma limpeza de mato ou se não é efetuada qualquer intervenção. Os dados biométricos de certos povoamentos, como a idade e/ou área basal, poderão colocar algumas restrições em relação a determinadas intervenções que compõem as prescrições. Deste modo, determinadas prescrições só poderão ser adotadas em alguns povoamentos, dependendo das condições destes no início do horizonte de planeamento.

O modelo de PIM apresentado é, essencialmente, um modelo determinístico onde se pretende maximizar o valor esperado do solo, com restrições de resistência ao fogo. O impacto de possíveis incêndios verifica-se ao nível da função objetivo, onde se assume que há uma percentagem das árvores que arde. Assim, o rendimento atual líquido associado a cada intervenção corresponde a um valor esperado que depende da probabilidade de ocorrência de incêndio, da probabilidade de ocorrer mortalidade, no caso de incêndio, e da proporção de árvores mortas, no caso de se verificar mortalidade. Este valor esperado é calculado recorrendo ao indicador de resistência associado a cada prescrição (r_{ikt}). No caso das intervenções serem do tipo desbaste ou corte raso, a receita gerada pela venda da madeira cortada só será obtida caso as árvores sobrevivam, logo, esta é multiplicada por r_{ikt}. Caso as árvores não sobrevivam - $(1 - r_{ikt})$- pressupõe-se que a madeira é vendida a um preço inferior (preço de salvados).

3.3 Cálculo dos pesos para a contribuição dos vizinhos na resistência ajustada

Para definir os pesos a atribuir à contribuição dos vizinhos para a resistência do povoamento i - $(1 - w_i)$ - teve-se em conta a dimensão e a forma do povoamento i, consideradno para o cálculo dos pesos w_i um quociente desigando por *shape index* já utilizado anteriormente para avaliar formas de regiões florestais (e.g. [Thomas, 1979], [Ohman e Lamas, 2005]), que depende da área e do perímetro do povoamento . Considerou-se que quanto maior o perímetro, maior é a importância dada à contribuição dos vizinhos para a sobrevivência do povoamento e que quanto maior a área do povoamento i, maior é o peso dado à contribuição do próprio povoamento i para a sua sobrevivência.

Para definir o peso da contribuição do povoamento vizinho s, na resistência do povoamento i - α_{is} - considerou-se:
i) a percentagem da fronteira do povoamento i partilhada com o povoamento s (quanto maior esta percentagem maior é a possibilidade de um incêndio passar de s para i);
ii) o declive do povoamento s (um maior declive de s facilita a propagação do fogo, uma vez que este, em geral, sobe);
iii) a posição relativa do povoamento s em relação a i (tendo em conta ii), a propagação do fogo também se intensifica quando s está a uma altitude inferior à de i e quando está orientado para Noroeste);
iv) a existência de barreiras entre os povoamentos - caminhos, estradas, rios,... (a existência de barreiras entre povoamentos dificulta a passagem do fogo).
Tendo em conta estes critérios, foram atribuídos pesos de forma intuitiva e empírica, de acordo com a opinião de especialistas. Por exemplo, considerando o povoamento $i = 9$ e os seus vizinhos $s_1 = 10$ e $s_2 = 27$, os valores atribuídos aos pesos $\alpha_{9,10}$ e $\alpha_{9,27}$ foram 0.2178 e 0.0820, respetivamente (Tabela 2).

Tabela 2: Pesos α_{is} , considerando o povoamento $i = 9$ e os seus vizinhos $s_1 = 10$ e $s_2 = 27$.

	Vizinhos do povoamento $i = 9$	
	$s_1 = 10$	$s_2 = 27$
Altitude (m)	11	27
Declive (°)	4	4
Orientação Noroeste	sim	não
Barreiras	nenhuma	10 metros
α_{is}	0.2178	0.0820

3.4 Resultados

O modelo de PIM foi implementado e resolvido através da linguagem algébrica ILOG OPL 6.3 e do *solver* CPLEX 12.1, com recurso a um computador i7-2600 CPU 3.40 GHz com 16GB de RAM.

Os resultados que serão apresentados consideram um horizonte de planeamento de 40 anos, decomposto em 4 períodos de 10 anos cada e um número máximo de 40 prescrições possíveis para cada povoamento. Utilizou-se ainda uma idade média mínima para a floresta, no final do horizonte de planeamento, de 40 anos ($f = 40$) e uma variação de volume permitida entre períodos de 25% ($\lambda = 0.25$). Quanto ao nível mínimo de resistência ajustada para a floresta, para cada período considerado, testaram-se diferentes valores desde $RES_t = 0.7$ a $RES_t = 0.89$, com $t = 1, ..., 4$.

O problema resolvido contém 1977 restrições e 4069 variáveis, das quais 2496 são variáveis binárias. Para uma resistência mínima de 0.89 obtém-se uma solução com o valor esperado do solo de 69, 127, 433.96€, com um *gap* para a relaxação linear de 0.015%, enquanto este em relação ao melhor nodo da árvore de pesquisa é de 0.010%. Esta solução foi obtida após 898 segundos e corresponde a uma resistência efetiva para os 4 períodos do horizonte de planeamento de 0.9444, 0.8944, 0.8900 e 0.8910, respetivamente (Tabela 3). Para este nível de resistência, a intervenção predominante em todos os períodos é o desbaste. Este é em número ainda mais elevado no último período, sendo efetuado em 333 povoamentos, enquanto é realizado em 270, 241 e 281 povoamentos nos 3 períodos anteriores. O número de cortes rasos é mais acentuado no primeiro período (realizado em 97 povoamentos). Por outro lado, é apenas no segundo período que as limpezas de mato ocorrem, como intervenção isolada, sendo levadas a cabo em 97 povoamentos.

A solução encontrada para o problema reflete o estado inicial considerado para a floresta, nomeadamente, no que respeita à estrutura inicial de idades ds povoamentos considerados. Por outro lado, a solução traduz também a silvicultura usual praticada na MNL, que considera desbastes e limpezas obrigatórias em determindaos períodos, de forma a obter resultados consistentes com o modelo de crescimento.

Tabela 3: Soluções do modelo de PIM, para os diferentes níveis de resistência mínima, para a floresta de 393 povoamentos.

RES_t					Solução					
Min.	Verificada				Função objetivo (€)				Tempo	Nodos
	$t=1$	$t=2$	$t=3$	$t=4$	Melhor inteira	Gaps (%)			(seg.)	analis.
						RL	1^o Nó	Final		
0.7	0.9444	0.8984	0.8443	0.8871	69,481,840.79	0.011	0.011	0.010	54	11901
0.8	0.9444	0.8984	0.8443	0.8878	69,482,155.42	0.010	0.010	0.010	16	7006
0.85	0.9444	0.8984	0.8503	0.8869	69,472,871.38	0.006	0.006	0.005	10	1685
0.86	0.9444	0.8984	0.8600	0.8872	69,444,763.90	0.012	0.012	0.010	424	694302
0.87	0.9444	0.8984	0.8700	0.8873	69,423,133.55	0.005	0.005	0.003	154	146680
0.88	0.9444	0.8984	0.8800	0.8872	69,389,985.81	0.011	0.011	0.010	47	11091
0.89	0.9444	0.8944	0.8900	0.8910	69,127,433.96	0.015	0.014	0.010	898	310817
0.9						Não há solução admissível				

Tabela 4: Soluções do modelo de PIM para uma resistência mínima de 0.89, considerando diferentes pesos para os vizinhos, para a floresta de 393 povoamentos.

Análise de Sensibilidade						
	Solução					
	Função objetivo (€)				Tempo	Nodos
Teste	Melhor inteira	Gaps (%)			(seg.)	analisados
		RL	1^o Nó	Final		
$\alpha_{is} \neq 0$	69,127,433.96	0.015	0.014	0.010	898	310817
$\alpha_{is} = 0$	69,052,043.29	0.013	0.011	0.001	38	27751

Um dos testes realizados teve como objetivo analisar a importância da influência dos vizinhos para a resistência ajustada de cada povoamento, comparando os casos em que o peso atribuído aos povoamentos vizinhos, no cálculo da resistência ajustada, é nulo ou diferente de zero. Quando se considera nula a contribuição dos vizinhos para a resistência ($\alpha_{is} = 0$) há alteração das prescrições de 30 povoamentos e o valor esperado do solo diminui em $75,390.68$€ (Tabela 4). Verifica-se assim que intervenções efetuadas num dado povoamento, com a finalidade de aumentar a sua resistência específica, podem refletir também um incremento na resistência dos povoamentos vizinhos. Intervenções de proteção levadas a cabo em determinados povoamentos, com o objetivo de diminuir a sua vulnerabilidade relativamente aos incêndios florestais, podem então também gerar benefícios em termos de resistência para os povoamentos vizinhos. Esta coordenação com os vizinhos das intervenções executadas, poderá diminuir o número total de medidas de proteção efetuadas, traduzindo-se numa redução dos custos totais suportados.

4 Conclusão

O modelo pretende otimizar a gestão de uma floresta, composta por diversos povoamentos, atribuindo prescrição para cada povoamento e maximizando o valor esperado do solo da floresta. Cada prescrição envolve a realização de diferentes intervenções ao longo do horizonte de planeamento considerado.

Uma inovação deste modelo é a tentativa de incorporar o risco de incêndio nos planos de gestão da floresta à escala da paisagem, incluindo intervenções de limpezas de mato, para além dos desbastes e cortes. O risco de incêndio foi incorporado no modelo através da utilização de um índice de resistência, que reflete a sobrevivência da floresta perante a ocorrência de eventuais incêndios florestais. Este índice incorpora a resistência específica do povoamento que tem por base informação proveniente de modelos de risco e danos de incêndios, desenvolvidos para as florestas portuguesas. Neste índice, é ainda englobada a contribuição dos povoamentos vizinhos para o aumento ou diminuição da resistência específica de um dado povoamento. Esta incorporação é fundamental em modelos à escala da paisagem que incluem o risco de incêndio, dada a possibilidade de propagação dos fogos de uns povoamentos para os outros.

Os resultados refletem que o modelo, apesar de combinatório, consegue ser resolvido em relativamente pouco tempo, sendo possível determinar uma prescrição a adotar em cada povoamento durante um determinado horizonte de planeamento.

A proteção de determinados povoamentos, com o objetivo de aumentar a sua resistência, nomeadamente através de limpezas de mato, é levada a cabo num número significativo de povoamentos. O modelo escolhe, várias vezes, prescrições que envolvem limpezas de mato em detrimento de prescrições semelhantes que não incluem essas limpezas. De notar ainda que, para além de intervenções que consideram apenas as limpezas de mato, quer as intervenções do tipo desbaste quer as intervenções do tipo corte raso pressupõem também a limpeza de matos e, consequentemente, contribuem também para diminuir a vulnerabilidade dos povoamentos, reduzindo o risco de incêndio. Estas medidas de proteção podem ser executadas, em cada povoamento, de forma coordenada com os seus vizinhos. Dado que um aumento da resistência dos vizinhos pode contribuir para aumentar a resistência ajustada de um dado povoamento, então intervenções adequadas em certos povoamentos podem permitir aumentar também a resistência dos povoamentos que os rodeiam, sendo possível efetivar uma poupança, uma vez que poderá não ser necessário efetuar medidas de proteção em todos os povoamentos. Ignorar a contribuição dos vizinhos no cálculo da resistência dos povoamentos pode implicar um aumento de custos resultante da realização de intervenções de proteção em cada povoamento de forma isolada.

Referências

[Alexandrian *et al.*, 2000] Alexandrian, D., Esnault, F. and Calabri, G. (2000). Forest fires in the Mediterranean area. *Unasylva*, 197(50): 35-41.

[Acuna *et al.*, 2010] Acuna, M., Palma, C., Cui, W., Martell, D., Weintraub, A. (2010). Integrated spatial fire and forest management planning. *Canadian Journal of Forest Research*, 40: 2370-2383.

[Agee *et al.*, 2000] Agee, J., Bahro, B., Finney, M., Omi, P., Sapsis, D., Skinner, C., Van Wagtendonk, J., and Weatherspoon, C. (2000). The use of shaded fuelbreaks in landscape fire management. *Forest Ecology Management*, 127: 55–66.

[Borges and Uva, 2006] Borges, J. G. e J. S. Uva. (2006). A prevenção e o combate a incêndios no âmbito do ordenamento e da gestão florestal. *J. S. Pereira , J. M. C. Pereira, F. C. Rego, J. N. Silva e T. P. Silva (Eds). Incêndios Florestais em Portugal: Caracterização, Impactes e Prevenção*, ISAPress, Lisboa: 313-335.

[Garcia-Gonzalo *et al.*, 2011a] Garcia-Gonzalo, J., Marques, S. ,Borges J. G., Botequim B., Oliveira M. M., Tomé, J., Tomé, M. (2011). A three-step approach to post-fire mortality modelling in Maritime pine (Pinus pinaster Ait) stands for enhanced forest planning in Portugal. *Forestry*, 84(2): 197-206.

[Garcia-Gonzalo *et al.*, 2011b] Garcia-Gonzalo,J., Botequim B., Zubizarreta-Gerendiain, A., Ricardo, A., Borges J. G., Marques, S., Oliveira M. M., Tomé, M. and Pereira, J.M.C. (2011). Modelling wildfire risk in pure and mixed forest stands in Portugal. *Allgemeine Forst und Jagdzeitung (AFJZ), German Journal of Forest Research*, 183 (11/12): 238-248.

[González and Pukkala, 2010] González-Olabarria, J.R. and Pukkala, T. (2010). Integrating fire risk considerations in landscape level forest planning. *Forest Ecology and Management*, 261: 278-287.

[Konoshima *et al.*, 2010] Konoshima, M., Albers, H., Montgomery, C., Arthur, J. (2010). Optimal spatial patterns of fuel management and timber harvest with fire risk. *Canadian Journal of Forest Research*, 40: 95-108.

[Ohman e Lamas, 2005] Ohman, K., Lamas, T. (2005). Reducing forest fragmentation in long-term forest planning by using the shape index. *Forest Ecology and Management*, 212, 346-357.

[Pereira *et al.*, 2006] Pereira, J. M. C., Carreiras, J.M.B., Silva, J. M. N., Vasconcelos, M. J. (2006). Alguns conceitos básicos sobre fogos rurais em Portugal. *Pereira, J. S., Pereira, J. M. C., Rego, F. C., Silva, J. M. N., Silva, T. P. (ed.), Incêndios Florestais em Portugal: Caracterização, Impactes e Prevenção*, ISAPress, Lisboa, 133: 161.

[Thomas, 1979] Thomas, J.W. (1979). Wildlife habitats in managed forests: the Blue Mountains of Oregon and Washington. *Agriculture Handbook* No. 553.

A tool to manage tasks of R&D projects

Joana Fialho*, Pedro Godinho†, João Paulo Costa†

* *Escola Superior de Tecnologia e Gestão de Viseu, Instituto Politécnico de Viseu, Portugal*
E-mail:`jfialho@estv.ipv.pt`

† *Faculdade de Economia da Universidade de Coimbra*
E-mail(Pedro Godinho):`pgodinho@fe.uc.pt`

E-mail (João Paulo Costa):`jpaulo@fe.uc.pt`

Abstract

We propose a tool for managing tasks of Research and Development (R&D) projects. We define an R&D project as a network of tasks and we assume that different amounts of resources may be allocated to a task, leading to different costs and different average execution times. The advancement of a task is stochastic, and the management may reallocate resources while the task is being performed, according to its progress. The operational cash flows depend on the task completion time, and their expected values follow a stochastic process. We consider that a strategy for completing a task is a set of rules that define the level of resources to be allocated to the task at each moment. We discuss the evaluation of strategies for completing a task, and we address the problem of finding the optimal strategy.

Key words: R&D task management and evaluation, real options, stochastic models, simulation, optimal decisions.

1 Introduction

Companies operating in dynamic markets, driven by technological innovation, need to decide, at each moment, which projects to carry out and the amount of resources to allocate to them. These decisions are crucial for the companies' success. R&D projects are characterized by several types of uncertainty and by the possibility of changing the initial plan of action, that is, R&D projects have two very important features that have to be taken into account: uncertainty and operational flexibility. The flexibility of a project leads to an increase in the project value that must be taken into account in its analysis or evaluation. When there is operational flexibility, it may be better to change the plan of action when new information arrives.

Traditional project evaluation methods, such as the ones based on discounted cash flows, are not adequate because they assume a pre-determined and fixed plan, which does not allow taking into account both uncertainty and flexibility (Yeo and Qiu, 2003).
The recognition that the financial options theory can be used to evaluate investment projects was made by Myers (1984), who used the expression real option to express the management flexibility under uncertain environments. Real options theory allows us to determine the best sequence of decisions to make in an uncertain environment, and provides the proper way to evaluate a project when such flexibility is present. The decisions are made according to the opportunities that appear along the project life time, which means that the optimal decision-path is chosen step by step, switching paths as events and opportunities appear (Cortazar *et al.*, 2008).
To evaluate or analyze an R&D project, it is important to evaluate the sequential real options that appear along the life time of the project. To evaluate these options, it is important to incorporate the associated risk. This risk may be related to prices, costs and technology, among others. There are several processes to model these variables, like Brownian motions (Cortazar *et al.*, 2003), mean reversing models (Copeland and Antikarov, 2001), controlled diffusion processes (Schwartz and Zozaya-Gorostiza, 2003), or even combinations between diffusion and Poisson processes (Pennings and Sereno, 2011). The Poisson processes are also widely used to model technological uncertainties (Pennings and Sereno, 2011) or catastrophic events that make it impossible to proceed with the project (Schwartz and Zozaya-Gorostiza, 2003). The revenues may also be uncertain, and it may be necessary to model them with stochastic processes (Schwartz and Moon, 2000).

We present a tool that can be applied to evaluate R&D projects, in order to help management make decisions concerning resource allocation. In developing the model, we were particularly concerned with the use of human resources, which are often the scarcer resources in this kind of projects. However, the model is flexible enough to handle other types of resources, like equipment or financial resources (e.g., for subcontracting some fractions of the tasks). The main condition to use the model is the ability to define a finite, discrete set of levels of resources that can be used at each instant, and to define the cost *per* time unit of each level of resources.
The model behind the tool presented in this paper follows Godinho *et al.* (2007), who proposed a real options model for the analysis of R&D projects from the telecommunications sector. The model and procedure evaluation we present take into account that the main limitation and difficulty is the human resources allocation. Management needs a tool that helps define the best allocation, in order to maximize the project value. So, the evaluation procedure should define a financial value for the project and the correspondent strategy to execute it, that is, we present a tool that gives a set of rules defining the level of resources to allocate, in order to maximize the project financial value.
We consider that an R&D project is composed by different tasks, and to evaluate a project, we must evaluate its tasks. The tool herein presented intends to evaluate those tasks. The output of this tool helps management to allocate resources to tasks that compose an R&D project. Although the tool presented evaluates a task of an R&D project, it implies that the evaluation of the tasks that compose a project leads to a financial evaluation of the project. The connection between tasks will be detailed in future work, because it is necessary to determine how the tasks are linked to each other and how codependent they are. Notice that some tasks can have precedents, that is, some tasks can only begin when others are completed or if others had obtained certain results.
For each task, we assume that different resource levels can be allocated, which have different costs and different average execution times. The advance of the task is stochastic and the project manager can reallocate resources while the task is in progress. The progression of a task defines, at each moment, which is the best level of resources to allocate to it. The difference between the resource levels can be quantitative or qualitative, that is, the levels can be different due the number of persons or due the qualities/specialization of those persons. Different strategies are analyzed, and the objective is to find the optimal strategy to execute an R&D task.

The procedure we used is based on Least Squares Monte Carlo - LSMC - proposed by Longstaff and Schwartz (2001): it constructs regression functions to explain the payoffs for the continuation of an option through the values of the state variables. A set of simulated paths of the state variables is generated. With the simulated paths, the optimal decisions are set for the last period. From these decisions, it is built, for the penultimate period, a conditional function that sets the expected value taking into account the optimal decisions of the last period. With this function, optimal decisions are defined for the penultimate period. The process continues by backward induction until the first period is reached. The use of simulation allows integrating different state variables in an easy way. We elected the LSMC process due to the fact that decisions are taken according to future expectations. However, some adaptations were necessary, for example in the way time is handled.

This paper is structured as follows: section 2 presents and characterizes the model, section 3 describes the analysis procedure, section 4 presents some examples, and section 5 concludes.

2 Proposed model

We assume that each task is homogeneous and needs a certain number of identical work units to be completed. These work units can be seen as small parts of the task and the set of these parts composes the task. The work units can be executed by different resource levels, which lead to different average times to finish the task and different costs *per* time unit.
We consider, in our model, that there is uncertainty in the time it takes to complete a task, and consequently, in the costs, because they depend directly on the time to complete the task.
The costs are deterministic, *per* unit of time, and depend on the level of resources. We also assume that there may be a cost inherent to switching between different resource levels, that is, reallocating persons to different tasks or to different work units may entail a cost.
In the model herein presented, we do not model the revenues, but the cash flows resulting from the completion of the task. The expected operational cash flows resulting from the exploration of the investment project follow a stochastic process and depends on the time it takes to complete each task.

Notice that a set of tasks composes a project. From now on, the present value of the cash flows resulting from the life time of each task (cash inflows and cash outflows) is denominated as task worth. These cash flows represent a portion of the total cash flows of the entire project. The concept of instantaneous task worth is used, which represents the present value of the task worth, assuming that the task was already finished. We also assume a penalty in the task worth according to its completion time, that is, the task worth is more penalized as the task takes longer to be completed. We admitted this penalty, because we assume that R&D projects can turn more profitable if a product or a service is launched earlier.

Before the presentation of the evaluation procedure, it is necessary to introduce the modelling of the state variables. Thus, the next subsections describe in detail how we handle the time to complete a task, the task worth, the costs and the net present value of the task.

2.1 Time to complete a task

The time to complete a task is not deterministic because it is impossible to know it with certainty, due to unpredictable delays or technical difficulties. Considering a specific level of resources k along the entire task, we define the time to finish the task as $T^{(k)}$. $T^{(k)}$ is a random variable and it is a sum of a deterministic term, the minimum time to finish the task, $M^{(k)}$, with a stochastic one. Let D be the necessary number of work units to complete the task. The time it takes to complete each work unit is composed by a constant part and a stochastic one, the latter being defined by an exponential distribution. This distribution is adequate because we assume that the average number of work units completed *per* unit of time is constant and there is no *a priori* expectation as to the nature of the distribution (Folta and Miller, 2002). We also assume that the time it takes to complete one work unit is independent of the time it takes to complete the other work units. This assumption comes from the fact the work units are well defined, separated and each one starts when the previous ends. Thus, it is immediate that the necessary time, $T^{(k)}$, to complete the task, using the level of resources k, is defined by

$$T^{(k)} = \sum_{i=1}^{D} \widehat{t}_i^{(k)} \tag{1}$$

where $\widehat{t}_i^{(k)}$ is the time that each work unit takes, considering the level of resources k. Each term $\widehat{t}_i^{(k)}$ can be written as

$$\widehat{t}_i^{(k)} = \frac{M^{(k)}}{D} + t_i^{(k)} \tag{2}$$

where $t_i^{(k)}$ follows an exponential distribution with average $1/\mu^{(k)}$. Replacing $\widehat{t}_i^{(k)}$ in (1)

$$T^{(k)} = M^{(k)} + \sum_{i=1}^{D} t_i^{(k)} \tag{3}$$

The time to finish the task is composed by a sum of a deterministic term, which represents the minimum time that is necessary to finish the task, with a stochastic one. This latter term is defined as the sum of D independent and identically distributed exponential variables.

2.2 The costs

The costs are deterministic *per* unit of time, and depend on the level of resources and on the necessary time to complete the task. We assume that, for each level of resources, the costs increase at a constant rate, possibly the inflation rate. Considering a specific level of resources k, let $C_x^{(k)}$ be the instantaneous cost, at instant x. The model for the costs can be defined by

$$dC_x^{(k)} = \rho C_x^{(k)} dx \tag{4}$$

where ρ is the constant rate of growth of the costs. Thus, the value of $C_x^{(k)}$ is

$$C_x^{(k)} = C_0^{(k)} e^{\rho x} \tag{5}$$

where $C_0^{(k)}$ is a constant dependent of the level of resources. The cost, $\overline{C}_j^{(k)}$, of a work unit j that uses the level of resources k, and that begins on instant x_j and ends on instant x_{j+1} is

$$\overline{C}_j^{(k)} = \int_{x_j}^{x_{j+1}} C_x^{(k)} dx = \int_{x_j}^{x_{j+1}} C_0^{(k)} e^{\rho x} dx = \left[\frac{1}{\rho} C_0^{(k)} e^{\rho x}\right]_{x_j}^{x_{j+1}} = \frac{C_0^{(k)}}{\rho} \left(e^{\rho x_{j+1}} - e^{\rho x_j}\right) \tag{6}$$

The present value of the cost of the work unit j with respect to an instant x_0 with discount rate r is $\widehat{C}_{j,x_0}^{(k)}$ and it is given by

$$\widehat{C}_{j,x_0}^{(k)} = \int_{x_j}^{x_{j+1}} C_x^{(k)} e^{-r(x-x_0)} dx = \frac{C_0^{(k)} e^{r x_0}}{\rho - r} \left(e^{(\rho-r)x_{j+1}} - e^{(\rho-r)x_j}\right) \tag{7}$$

The expression for $\widehat{C}_{j,x_0}^{(k)}$, given in (7), is valid when $\rho \neq r$. In the case $\rho = r$,

$$\widehat{C}_{j,x_0}^{(k)} = \int_{x_j}^{x_{j+1}} C_x^{(k)} e^{-r(x-x_0)} dx = C_0^{(k)} e^{r x_0} \left(x_{j+1} - x_j\right) \tag{8}$$

We also assume that costs related to changes of the level of resources can occur. That is, if there is a change in the level of resources from one work unit to another, it may be necessary to incur a cost. The cost for changing the level of resources from k_j in the work unit j to other level of resources, k_{j+1}, in the next work unit $j+1$ is given by $\gamma(k_j, k_{j+1})$. We also assume that these costs are deterministic, depend on the level of resources and grow with the same rate ρ. If the change occurs at moment x_{j+1}, that is, at the moment that work unit $j+1$ begins, the value of the respective cost is $\gamma(k_j, k_{j+1}) e^{\rho x_{j+1}}$ and the present value of this cost, with respect to an instant x_0 is

$$\gamma(k_j, k_{j+1}) e^{\rho x_{j+1}} e^{-r(x_{j+1} - x_0)}$$

In our evaluation procedure, it is necessary to calculate the present value of the total remaining costs, that is, it is necessary to determine the total costs from a certain work unit j until the last one, D. We assume that, for all work units, $j = 1, ..., D$, the present value of the remaining costs are determined with respect to x_j, which is the instant in which the work unit j starts, and it is denoted as $TotC(j, x_j)$. The expression of $TotC(j, x_j)$ can be given by

$$TotC(j, x_j) = \sum_{a=j}^{D-1} \left[\widehat{C}_{a,x_j}^{(k_a)} + \gamma(k_a, k_{a+1}) e^{\rho x_{a+1}} e^{-r(x_{a+1} - x_j)}\right] + \widehat{C}_{D,x_j}^{(k_D)} \tag{9}$$

where

- x_j is the instant in which work unit j starts;
- $\widehat{C}_{a,x_j}^{(k_a)}$ is the present value of the cost of the work unit a, relatively to the instant x_j. The work unit a begins at instant x_a and uses the level of resources k_a;
- $\widehat{C}_{D,x_j}^{(k_D)}$ is the present value of the cost of the work unit D with respect to instant x_j. The work unit D uses the level of resources k_D;
- $\gamma(k_a, k_{a+1})$ defines the value of the cost to change from the level of resources k_a used in work unit a to the level of resources k_{a+1} used in work unit $a+1$. Notice that if the level of resources is the same in the work unit a and in the work unit $a+1$, this cost is zero;
- r is the discount rate.

2.3 The task worth

We define the task worth as the present value of the cash flows resulting from completing the task (including both cash inflows and cash outflows). Thus, the task worth does not depend on the level of resources used to undertake the task, but on the time to complete it. We also define the related concept of instantaneous task worth (or instantaneous worth), which is the value of the task worth assuming that the task is completed at the instant being considered.

We assume that the instantaneous task worth changes according to a pre-defined rate and with some stochastic events. The rate can be positive or negative, depending on the nature of the project. In R&D projects, new information can arrive, or unexpected events can occur, that change the course of the project and, consequently, the expectations regarding to the instantaneous task worth. We also assume that a penalty in the instantaneous task worth may occur, due to the duration of the task. That is, we assume that it may be the case that the earlier the product is launched in the market, the bigger is the worth obtained. The task worth may be more penalized as the task takes longer to complete. The reason for such penalty may be related to the existence of competition: if a competitor is able to introduce,

earlier, a similar product in the market, the task worth might be lower.
Let the model of instantaneous worth, R, be defined by:

$$R\ dR = \alpha R\ dx + R\ dq \tag{10}$$

The parameter α included in the model represents the increasing or decreasing rate of the instantaneous worth, in each lapse dx The term dq represents a Jump process, that is

$$dq = \begin{cases} u, & \text{with probability } pdx \\ 0, & \text{with probability } 1 - pdx \end{cases} \tag{11}$$

with u defined by a uniform distribution, $u \sim U(u_{min}, u_{max})$, $u_{min} \leq u_{max}$.
Notice that, if the instantaneous worth would depend only on the rate α, it would be continuous and monotone increasing (assuming the rate positive). But, besides the rate α, there is the possibility of occurring jumps in the instantaneous worth, due the nature of these projects and/or the behavior of the market.

In order to handle the model, we assume the discrete version of the instantaneous worth. Ignoring the jump process, the solution of equation (10) would be $R_x = R_0 e^{\alpha x}$, and therefore we would have $R_{x+1} - R_x = R_0 e^{\alpha x}(e^{\alpha} - 1)$. Considering low values for α, we can assume $\alpha \approx e^{\alpha} - 1$, and the expression would become $R_{x+1} - R_x = \alpha R_x$. In order to incorporate the jump process, we assume that in a lapse of time that is not infinitesimal, more than one jump may take place. So, the discrete version of the model of instantaneous worth becomes

$$R_{x+1} = R_x + \alpha R_x + R_x \Delta q \tag{12}$$

where $\Delta q = \sum_{i=1}^{\nu} u_i$, with $u_i \sim U(u_{min}, u_{max})$ and ν is defined by a Poisson distribution with parameter p, that is $\nu \sim P(p)$.

The present value of the instantaneous worth in a given moment depends on the instantaneous worth of the previous moment. Thus, the instantaneous worth of the first period is $R_1 = R_0 + \alpha R_0 + R_0 \Delta q$. It is necessary to know the initial value R_0, which is an input parameter for the model. Assuming that the task is finished at the moment T, we define R_T as the task worth calculated according to the model previously presented.

The penalty mentioned initially can be expressed by a function $g(x)$, where x denotes the time. This function is positive, decreasing, and it takes values from the interval $[0, 1]$. Thus, assuming this feature, the final expected task worth is $R_T \times g(T)$. Notice that, if there is no penalty, $g(x) = 1$, $\forall x$.

2.4 The net present value of the task

For the model, it is necessary to calculate the expected value of the net present value of the task, for each work unit j, $j = 1, ..., D$, and with respect to the initial instant of that work unit, x_j. The net present value of the task in each work unit includes the present value of the expected task worth at the end of the task and the present value of the total remaining costs. These present values are calculated with respect to the instant x_j. Thus, assuming that the instant to finalize the task is T, the net present value of the task, at the beginning of the work unit j is $Val(j, x_j)$, and it is determined as follows:

$$Val(j, x_j) = R_T \times g(T) \times e^{-r(T - x_j)} - TotC(j, x_j) \tag{13}$$

3 Procedure for the task evaluation

This procedure uses a method similar to the Least Square Monte Carlo (Longstaff and Schwartz, 2001). In the beginning of the process, we build many paths with different strategies. A strategy is a set of rules that defines the level of resources to use in each work unit, possibly taking into account the way the task is progressing. The strategies used to build the paths include executing all work units with the same level of resources or using different resource levels to finish the task. For each strategy, and for each path, we simulated the values of the time to execute the task, through the model we presented in the previous section. With the time and the level of resources used, we can determine the costs using the model presented in section 2.2, and through the model for the task worth, we can simulate the values for

the instantaneous worth; finally, we can determine the net present value of the task, for each path, and for each work unit.
In this procedure we build, by backward induction and for all work units, regression functions for values previously calculated for the paths. These functions explain the net present value of the task in order to different state variables which are the elapsed time, the instantaneous worth and the number of work units already finished.

The evaluation procedure begins in the last work unit. For all paths initially simulated, it is considered the instantaneous worth observed in the beginning of the last work unit, as well as the time elapsed until the start of that work unit. For all paths, assuming a specific level of resources, k_D, to complete the last work unit, the time to complete the task is redefined, as well as the net present value of the task in the beginning of the last work unit. Taking the net present value of the task recalculated in all paths, $V|k_D$, the elapsed time until the beginning of the last work unit, $Y_{1,D}$, and the instantaneous worth observed in the beginning of the last unit, $Y_{2,D}$, a regression function, F_{D,k_D}, is built. This function explains the net present value of the task, in the last unit, as function of the elapsed time until the beginning of the last work unit and of the instantaneous worth observed in the beginning of that unit. We regress $V|k_D$ on a constant, $Y_{1,D}$, $Y_{2,D}$, $Y_{1,D}^2$, $Y_{2,D}^2$ and $Y_{1,D}Y_{2,D}$, that is,

$$F_{D,k_D} = a_0 + a_1 Y_{1,D} + a_2 Y_{2,D} + a_3 Y_{1,D} Y_{2,D} + a_4 Y_{1,D}^2 + a_5 Y_{2,D}^2$$

We assumed these basis functions for the regression, but other basis functions could be selected without interfering or altering the process (Stentoft, 2004).
This procedure is repeated assuming the other resource levels to perform the last work unit. Thus, considering that there are N resource levels, in the last work unit, for each level of resources $k_D, k_D = 1, ..., N$, we define a function, F_{D,k_D}, which explains the net present value of the task as function of the elapsed time (until the beginning of the last unit) and of the instantaneous worth observed in the beginning of that unit.

For the earlier work units, the procedure is based on the same principle: it is considered that the work unit under consideration, say j, is executed with a specific level of resources, k_j. Next, the net present value of the task in the beginning of that unit is recalculated, through the definition of the best strategy from the following unit until the last one. The definition of the best strategy is done using the regression functions already determined (Figure 1) and the costs for switching levels: for each of the following work units, the level of resources chosen is the one that leads to a higher value in the difference between the respective regression function and the cost of switching the level (if the level of resources is different from the level used in to the previous work unit), that is, for $a = j+1, ..., D$, the level of resources chosen k_a is

$$\max_{k_a=1,...,N} \{F_{a,k_a} - \gamma(k_{a-1}, k_a) e^{\rho x_a}\}$$

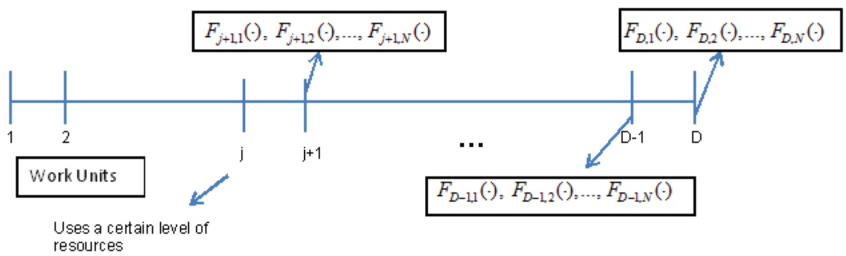


Figure 1: Functions that allow the definition of the best strategy from unit $j+1$ until the last unit D.

Assuming the specific level of resources used in the unit j, and with the best strategy defined from the work unit $j+1$ until the last one, we recalculate the net present value of the task in the beginning of the unit j. Taking the recalculated values of the net present value of the task, $V|k_j$, the values of the elapsed time until the beginning of the unit j, $Y_{1,j}$ and the values of the instantaneous worth observed in the same moment, $Y_{2,j}$, a regression function is defined. This regression function explains the net present value of the task as a function of the elapsed time until the beginning of work unit j and of the

instantaneous worth observed in the beginning of work unit j.
For this work unit j the procedure is repeated, assuming the other resource levels to execute it. In this way, we construct regression functions for all resource levels in the work unit j. These functions explain the net present value of the task as a function of the elapsed time and of the instantaneous worth.
The process proceeds by backward induction until the second work unit. This procedure allows having, for each work unit and for all resource levels, a regression function that estimates the net present value of the task, through the elapsed time and through the instantaneous worth observed (Figure 2).

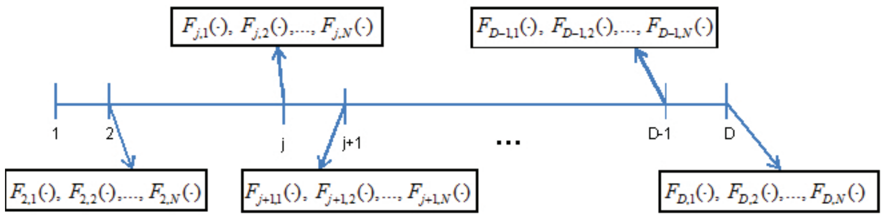


Figure 2: Regression functions, explaining the task value as function of elapsed time and instantaneous task worth.

For the first unit we do not construct the regression functions, due the fact of, for all paths, the instantaneous worth observed in the beginning of the first unit is R_0 and the elapsed time in that moment is 0, that is, the instantaneous worth observed and the elapsed time are constant. Thus, to evaluate the best strategy in the first work unit, a specific level of resources is assumed. Then, with the regression functions of the following work units, the best strategy is defined for all paths. With the best strategy in each path, the net present value of the task is calculated for the first work unit. The average of these values provides the task value, assuming that specific level of resources for the first unit. This evaluation is repeated, assuming the other resource levels for the first work unit. Notice that it is necessary to decide which level of resources may be used to begin the task. The level leading to a bigger average value of the task in the first unit, is chosen to initialize the task. After this procedure, the regression functions allow defining rules which can guide management in the decisions about the strategy to use. Thus, with the regression functions and knowing which level of resources was used, we can define rules, indicating management which is the level of resources to use next.

4 Numerical results

To test the evaluation procedure, we consider a project that is being executed. One of its tasks needs to be evaluated in order to be initialized. There are two different resource levels (level 1 and level 2) to execute the task and management needs to know which one will be used in the beginning of the task. For this specific task, $D = 20$ work units were defined, that is, the task is divided in 20 identical parts. We assume that, in average, level 1 can conclude 1.5 work units *per* unit of time, and level 2 can conclude 3 work units *per* unit of time. The costs increase at a rate of 0.5% *per* unit of time, and the discount rate is $r = 0.1\%$, *per* unit of time. We assume that the instantaneous task worth increases 1% *per* unit of time. The penalty function for the instantaneous worth punishes it up to 10%, if the task takes less than 10 units of time; if the task takes between 10 and 15 units of time, the task worth is penalized up to 30%; if the task takes longer than 15 units of time, the penalty is fixed: 30%. Thus, the penalty function, $g(x)$, where x represents time, is the following:

$$g(x) = \begin{cases} 1 - \frac{x}{10} \times 0.1, & \text{se } x \leq 10 \\ 0.9 - \frac{x-10}{15-10} \times 0.2, & \text{se } 10 < x \leq 15 \\ 0.7, & \text{se } x > 15 \end{cases}$$

The input parameters are listed in Table 1.
To run the evaluation procedure, 700 paths were considered and we used the following initial strategies: to execute the whole task with the first level of resources; to execute the whole task with the second level of resources; to execute half of the task with one level and the other half with the other level of resources. This led to a total of 2100 paths. The paths built, using only the level 1, led to an average time of 13.28,

Table 1: Input parameters for the numerical results.

Time	$M^{(1)} = M^{(2)} = 0$; $\mu^{(1)} = 1.5$; $\mu^{(2)} = 3$
Costs	$C_0^{(1)} = 10$; $C_0^{(2)} = 40$; $\gamma(k_j, k_{j+1}) = C_0^{k_{j+1}}$; $\rho = 0.5\%$
Task worth	$R_0 = 2000$; $\alpha = 1\%$; $\nu \sim P(0.4)$; $u_i \sim U(-0.2, 0.2)$

with a net present value of 1637.6. The paths built, using only the level 2, led to an average time of 6.64, with a net present value of 1708.5.
After running the procedure described in the previous section, the average time to execute the task is 8.4 and the net present value of the task is 1728.58.
Analyzing the results of the strategy used, level 2 is the only one chosen in the first units, but afterwards level 1 is chosen in many paths (Figure 3).

In order to analyze the procedure, we can assay which one is the indicated level of resources for the next work unit. If we know the level of resources used before and the regression functions of the next work unit, it is possible to choose the level of resources that should execute the next work unit, taking into account the state variables information. The best decisions can be plotted as regions in the two-dimension space defined by the instantaneous worth and by the elapsed time. Such plot can provide some intuition about the best choices concerning the level of resources to be used in the next work unit. If in a certain work unit d, level 1 was used, the equation to determine the frontier lines between "continuing with level 1" and "change to level 2" regions are defined by $F_{d+1,1}(\cdot) = F_{d+1,2}(\cdot) - \gamma(1,2)e^{\rho x_{d+1}}$. Similarly, if in a certain work unit d, level 2 was used, the equation to determine the frontier lines between the "continuing" and "change" regions is $F_{d+1,1}(\cdot) - \gamma(2,1)e^{\rho x_{d+1}} = F_{d+1,2}(\cdot)$.

For example, assume that unit 16 of the task is completed. Supposing that level 1 was used in unit 16, we can provide a plot that defines how the level of resources should be chosen for work unit 17. This plot defines two regions, "continue with level 1" and "change to level 2", with the frontier lines obtained through $F_{17,1}(\cdot) = F_{17,2}(\cdot) - \gamma(1,2)e^{\rho x_{17}}$. Besides these frontier lines, we also plotted choice of paths for work unit 17, when level 1 was used in work unit 16 (Figure 4). The little stars in Figure 4 correspond to the paths that used level 1 in unit 16 and continue with level 1 in unit 17. The little balls correspond to the paths that used level 1 in work unit 16 and changed to level 2 in work unit 17. According to the region in which the pair (elapsed time, instantaneous worth) is situated, it is possible to define the level of resources to use in unit 17.

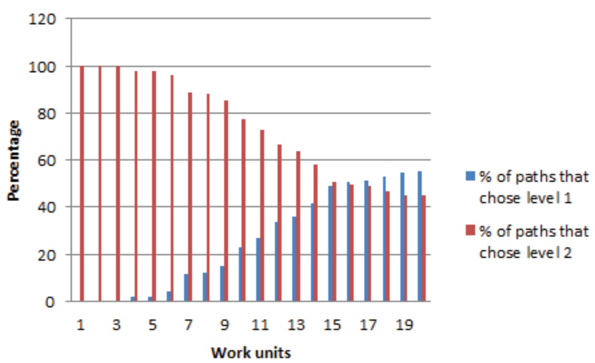


Figure 3: Percentage of the paths that choose each level of resources, after applying the evaluation process.

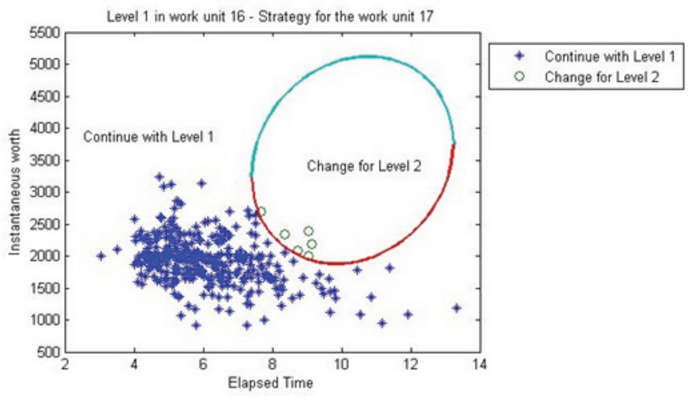


Figure 4: Strategy for work unit 17, when level 1 is used in the unit 16.

Now, we analyze the changes in the net present value of the task, when some parameters of the model are altered. To perform this analysis, we consider the example given above. The net present value in the beginning of the task can be different, depending on whether or not there are costs for switching the level of resources, and depending whether or on there is a penalty in the task worth. The Table 2 contains the net present value of the task for each case.

Table 2: Net present value of the task considering the existence or not of penalty in the task worth and/or costs to switch level.

Costs to switch level	Penalty	Net present value of the task in its beginning
Yes	Yes	1728.6
No	Yes	1733.0
No	No	2097.3
Yes	No	2097.7

The most significant increase in the net present value of the task occurs when we remove the task worth penalty. The removal of the cost of switching level might not change the net present value of the task very much, but changes the strategy to use. Notice that the inexistence of these costs leads to more changes of level, like shown in Figure 5. Notice that this happens if there is not a dominant level, that is, if there is not a level that leads to a higher net present value of the task in all work units.

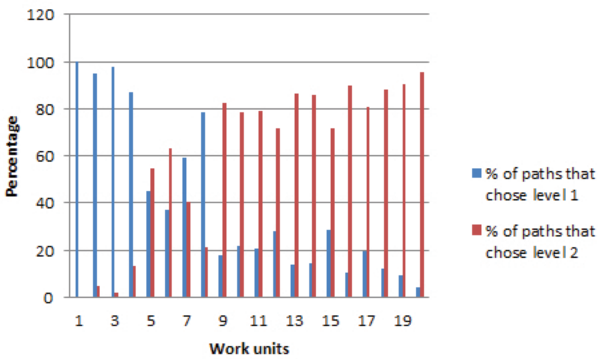


Figure 5: Percentage of the paths that choose each level of resources, after applying the evaluation process without costs to switch level of resources.

5 Concluding remarks

This work presents a tool for managing tasks of R&D projects, defining the strategy to execute a task which will lead to a higher net present value. This strategy consists in defining the level of resources to use, at each instant.
This approach can be improved by introducing an abandonment option, when the expected net present value is equal to or lower than a certain reference value. This option must be integrated and interpreted in the context of the project that contains the task.
Considering that an R&D project is a set of tasks, this evaluation procedure can be the basis to analyze the strategy to execute an R&D project, as well as the financial value associated. However, there are some aspects that must be taken into account: the result of the evaluation of a task influences the evaluation of the next task. On the other hand, the connections between the tasks and the way these connections influence the evaluation of an R&D project must also be taken into account.

References

[1] [Copeland and Antikarov, 2001] Copeland, T. and Antikarov, V. (2001). Real Options. Texere, New York.

[2] [Cortazar *et al.*, 2008] Cortazar, G., Gravet, M. and Casassus, J. (2008). The valuation of multidimensional American real options using the LSM simulation method. *Computers & Operations Research*, 35: 113-129.

[3] [Cortazar *et al.*, 2003] Cortazar, G., Schwartz, E. and Casassus, J. (2003). Optimal exploration investments under price and geological-technical uncertainty: a real options model. *R&D Management*, 31(2): 181-189.

[4] [Folta and Miller, 2002] Folta, T. B. and Miller, K. D. (2002). Real options in equity partnerships. *Strategic Management Journal*, 23(1): 77-88.

[5] [Godinho *et al.*, 2007] Godinho, P., Regalado, J. and Afonso, R. (2007). A Model for the Application of Real Options Analysis to R&D Projects in the Telecommunications Sector. *Global Business And Economics Anthology II*, 34: 63-76.

[6] [Longstaff and Schwartz, 2001] Longstaff, F. and Schwartz, E. (2001). Valuing American Options by Simulation: A Simple Least-Square Approach. *The Review of Financial Studies*, 14: 1113-1147.

[7] [Myers, 1984] Myers, S. C. (1984). Finance theory and finance strategy. *Interfaces*, 14(1): 126-137.

[8] [Pennings and Sereno, 2011] Pennings, E. and Sereno, L. (2011). Evaluating pharmaceutical R&D under technical and economic uncertainty. *European Journal of Operational Research*, 212: 374-385.

[9] [Rodrigues and Armada, 2006] Rodrigues, A. and Armada, M. (2006). The Valuation of Real Options with the Least Squares Monte Carlo Simulation Method. *Management Research Unit, University of Minho, Portugal, Available at http://ssrn.com/abstract=887953 (last access in 18-02-2013)*

[10] [Schwartz and Moon, 2000] Schwartz, E. and Moon, M. (2000). Rational Pricing of Internet Companies. *Financial Analysts Journal*, 56(3): 62-75.

[11] [Schwartz and Zozaya-Gorostiza, 2003] Schwartz, E. and Zozaya-Gorostiza, C. (2003). Investment Under Uncertainty in Information Technology: Acquisition and Development Projects. *Management Science*, 49(1): 57-70.

[12] [Stentoft, 2004] Stentoft, L. (2004). Convergence of the Least Squares Monte Carlo Approach to American Option Valuation. *Management Science*, 80(9): 1193-1203.

[13] [Yeo and Qiu, 2003] Yeo, K. T. and Qiu, F. (2003). The value of management flexibility - a real option approach to investment evaluation. *International Journal of Project Management*, 21: 243-250.

An optimisation model for the warehouse design and planning problem

Carla A. S. Geraldes*, Sameiro Carvalho†, Guilherme Pereira†

* Centro ALGORITMI and Polytechnic Institute of Bragança, Portugal
carlag@ipb.pt

† Centro ALGORITMI and University of Minho, Portugal
sameiro@dps.uminho.pt, gui@dps.uminho.pt

Abstract

In spite of the importance of warehouses in the field of the supply chain management, there is not a single decision model that integrates all the decisions that concerns the warehouse design and planning problem. A number of warehouse decision support models have been proposed in the literature but considerable difficulties in applying these models still remain, due to the large amount of information to be processed and to the large number of possible alternatives. In this paper we discuss a mathematical programming model aiming to support some warehouse management and inventory decisions. Our aim is to address the complexity related to the modeling of the warehouse design and planning problem. In particular a large mixed-integer nonlinear programing model (MINLP) is presented to capture the trade-offs among both inventory and warehouse costs in order to achieve global optimal design satisfying throughput requirements.

Keywords: Supply chain management, Warehouse design and planning, Mathematical modeling.

1 Introduction

In a supply chain network (see Figure 1) products need to be physically moved from one location to another. During this process, they may be stored or buffered at *warehouses* for a certain period of time for strategic or tactical reasons. Within this context, warehouses play an important role in supply chain management and may be considered a key aspect in a very demanding, competitive and uncertain market.

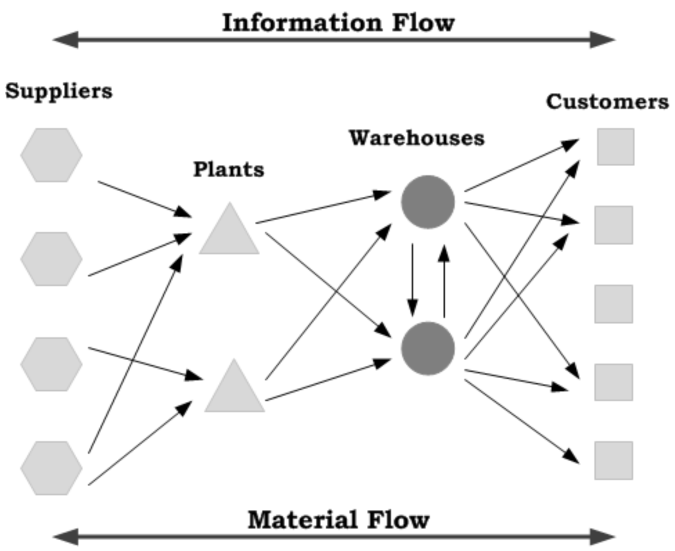


Figure 1: A typical supply chain network.

Modern supply chain principles compel organizations to reduce or eliminate inventory levels. Additionally warehouses require capital, labour, and information technologies, which are expensive resources. Although many companies examined the possibility of directly supply to customers, there are still many circumstances where this is not appropriate. So, why do we still need warehouses? According to Bartholdi and Hackman [2] there are four main reasons why warehouses are useful:

1. To consolidate products in order to reduce transportation costs and provide customer service;
2. To take advantage of economies of scale;
3. To provide value-added processing services, and
4. To reduce response time.

Thus, warehouses will continue to be an important node at the logistic network by the fact that if a warehouse cannot process the orders quickly, effectively, and accurately, then all the supply chain optimisation efforts will suffer (see Tompkins [14]).

In distribution logistic where market competition requires higher performances from the warehouses, organizations are compelled to continuously improve the design and planning of warehouses. Furthermore, the ever increasing variety of products, the constant changes in customer demands and the adoption of agile management philosophies also bring new challenges to reach flexible structures that provide quality, efficiency and effectiveness to the warehouse operations.

The primary functions of a warehouse include: (i) temporary storage of products; and (ii) providing of value-added services such as packing of products, after sales services, inspection, and assembly. To perform these functions warehouses are generally divided into different functional areas: receiving area, reserve storage area; picking or forward area, and the accumulation and shipping areas (see Figure 2).

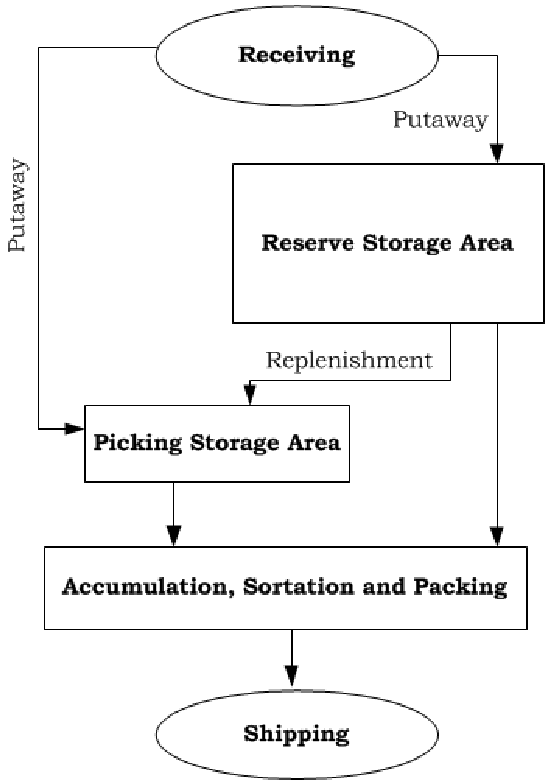


Figure 2: Typical warehouse functions and flows.

At the receiving area products are unloaded and inspected to verify any quantity and quality inconsistency. Afterwards, products are transferred to a storage zone. We can distinguish two types of storage areas: the reserve storage area and the forward or picking area. The reserve area is the place where the products stay until they are required by costumers' orders. The picking area is a relatively small area, typically used to store fast moving products. Most of the flows between these areas are the result of replenishment processes. Order picking is one of the most important functions in most warehouses. Stock Keeping Units (SKU) are retrieved from their storage positions based on customers' orders and moved to the accumulation and sorting area or directly to the shipment area. The picked units are then grouped by customer's order, packaged and stacked on the right unit load and transferred to the shipping area.

The design and planning of a warehouse is a very complex problem due to the large number of interrelated decisions. Some major decisions involved in the warehouse design and operation problem are illustrated in Figure 3 (see Gu et al. [5]).

Warehouse design decisions typically run from a functional description, through a technical specifica-

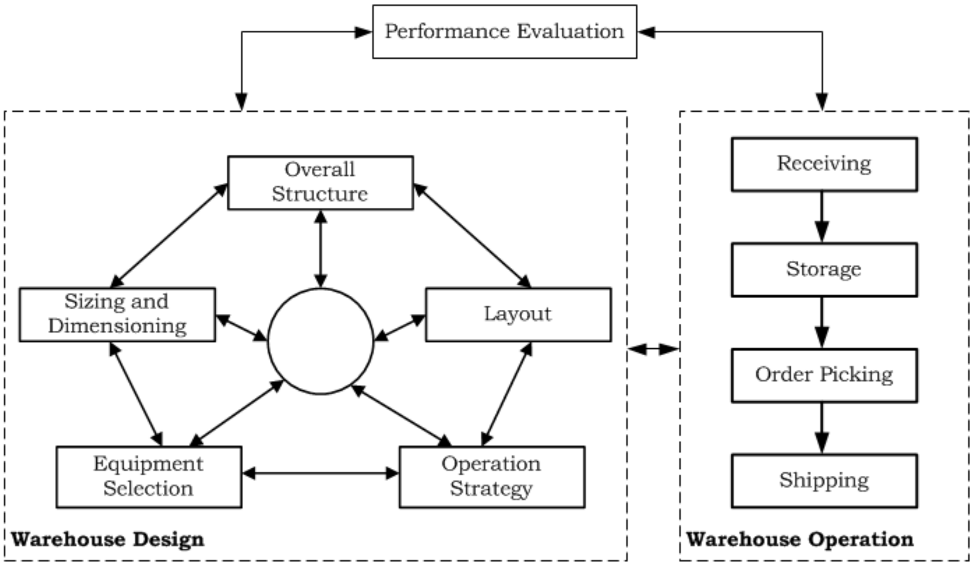


Figure 3: A framework for warehouse design and operation problems (adapted from Gu et al. [5]).

tion, to equipment selection and determination of the layout. The overall structure decision determines the material flow patterns within the warehouse, the specification of the functional areas and the flows between the areas. Sizing and dimensioning decisions determine the total size of the warehouse as well as the space allocation among functional areas. Layout definition is the detailed configuration within a functional area and equipment decisions define an automation level for the warehouse and identify equipment types. Finally operating policies refer to storage, picking and routing decisions.

Despite the importance of warehousing there is not a comprehensive systematic method for the warehouse design problem. Baker and Canessa [1] explored the literature on the overall methodology of warehouse design, together with the literature on tools and techniques used for specific areas of analysis. The output was a general framework of steps, with tools and techniques that can be of value to warehouse practitioners. Also Hassan [8] presented a framework for the design of a warehouse. The proposed framework accounts for several factors and operations of warehousing such as:

1. Specification of warehouse type and purpose;
2. Analysis and forecasting of the demand;
3. Definition of operating policies;
4. Establishment of inventory levels;
5. Class formation;
6. Definition of functional areas and general layout;
7. Storage partition;
8. Selection of equipment for handling and storage;
9. Design of aisles;
10. Determination of space requirements;
11. Location and number of Input/Output points;
12. Location and number of docks;
13. Arrangement of storage;
14. Zone formation.

Even though interrelated these decisions are dealt independently in a pyramidal top-down approach. Strategic decisions create limits to decisions taken at the tactical and operational levels and tactical decisions limits operational decisions. Also decisions taken at each level are handled independently and sequentially (see Van den Berg [15]).

The majority of warehouse decision models available in scientific literature addresses isolated or simplified problems. However, most of the real problems are unfortunately not well-defined and often cannot be reduced to multiple isolated sub-problems. Therefore, warehouse design often requires a mixture of analytical skills and creativity. Anyhow, research aiming an integration of various decisions models and methods is badly needed in order to develop a methodology for systematic warehouse design (see Rouwenhorst et al. [12]).

Furthermore, as the literature review in the next section shows, most research efforts have been dedicated to warehouse operations decisions instead of design decisions. This is not surprising since design decisions models are more difficult to develop and treat analytically once they require the integration of several and complex issues.

In this paper we present a warehouse and inventory mathematical model that jointly integrates issues concerning:

– The size of the warehouse and the size of the functional areas;
– The external storage additional capacity, if needed;
– The assignment and allocation of SKU to storage areas;
– The replenishment quantities and reorder points of the products to be stored.

Our aim is to test an integrated approach that takes into account inventory decisions and some warehouse design decisions.

In the Section 2 of this paper we will present a brief literature review on warehouse design and planning issues. The purpose of this section is not limited to the specific studied problem but also covers other important topics in warehouse literature. The Section 3 will present the proposed warehouse design and planning model formulation and a description of the methodology used to solve the model. Computational results of a preliminary study will be presented and summarized in Section 4, and finally some conclusions and future work directions are reported in Section 5.

2 Literature Review

2.1 Warehouse Design and Planning

Warehouse design can be defined as a structured approach of decision making at distinct decision levels in an attempt to meet a number of well-defined performance criteria. At each level, multiple decisions are interrelated and therefore it is necessary to cluster relevant problems that are to be solved simultaneously. According to Rouwenhorst et al. [12] a warehouse design problem is a “coherent cluster of decisions” and they define decisions to be coherent when a sequential optimization does not guarantee a globally optimal solution.

The design of a warehouse includes a large number of interrelated decisions involving warehouse processes, warehouse resources and the organization of the warehouse (see Heragu [9]). Rouwenhorst et al. [12] classified the management decisions concerning warehousing into strategic decisions, tactical decisions and operational decisions. Strategic decisions are long term decisions and always mean high investments. The two main issues involved are concerned with the design of the process flow and with the selection of the types of the warehousing systems. Tactical management decisions are medium term decisions based on the outcomes of the strategic decisions. The tactical decisions have a lower impact than the strategic decisions, but still require some investments and should therefore not be reconsidered too often. At the operational level, processes have to be carried out within the constraints set by the strategic and tactical decisions made at the higher levels. In this level, the concern includes the operational policies such as storage policies and picking and routing operations.

After determining warehouse location and its size, layout decisions must include areas definition and the dimension that should be allocated to each functional area. The forward-reserve problem (FRP) is the problem of assigning products to the functional areas. In this problem the critical decision concerns the choice of products that will be stored in the forward area. Van den Berg et al. [15] proposed a binary programming model to solve the FRP in the case of unit load replenishment, and presented efficient heuristics that provide tight performances guaranties. These replenishments can occur during busy or idle picking periods. The objective was to minimize the number of urgent or concurrent replenishments of the forward area during the busy periods. Although addressing this problem is a strategic decision problem, it is strongly associated upon some tactical problems such as how the items will be distributed among the functional areas. However, the approach usually adopted is to solve the problems sequentially by generating multiples alternatives for the functional area size problem and then determine how the products can be allocated for each of the alternatives.

Gray et al. [6] developed an integrated approach for the design and operation of a typical order-consolidation warehouse. This approach included warehouse layout definition, equipment and technology selection, product location, zoning, picker routing, pick generation list and order batching. Due to the complexity of the overall problem, they developed a multi-stage hierarchical decision approach. This hierarchical approach used a sequence of coordinated mathematical models to evaluate the major economic trade-offs and to reduce the decision space to a few number of alternatives. They also used simulation technique for validation and fine tuning of the resulting warehouse design and operating policies.

Heragu et al. [9] developed a mathematical model and a heuristic algorithm that jointly determines the size of the functional areas and the allocation of the product in a way that minimizes the total material handling and storage costs. The proposed model uses real data readily available to a warehouse manager and considers realistic constraints.

Geraldes et al. [3] adapted the mixed-integer programming model proposed by Heragu et al. [9] to tackle the storage allocation and assignment problems during the redesign process of a Portuguese company warehouse.

Liu [10] applied clustering techniques to extract the correlated information from the customer orders, and then stock locations were optimised. The author proposed a binary programming model to group products or customers and simulations results demonstrated the potential benefits of the clustering technique to solve the stock location problem.

More recently Strack and Pochet [13] presented a robust approach that integrates aspects such as: (i) the size of the functional areas; (ii) the assignment and allocation of products to storage locations in the warehouse; (iii) the replenishment decision in the inventory management. This is probably the most integrated decision model found in this area, nevertheless still assumes fixed and known capacity for the warehouse.

2.2 Inventory Decisions

The adoption of new management philosophies compels companies to eliminate or reduce inventory levels. In addition to warehouse management decisions, an appropriate inventory policy may result in a reduction of the total warehousing costs and can also improve the efficiency of the operating policies within the warehouse. The aim of inventory management is to minimize total operating costs satisfying customer service requirements (see Ghiani et al. [4]). To accomplish this, an optimal ordering policy must answer questions, for each SKU, such as when to order and how much to order.

Two different inventory policies arise (see Hadley and Whitin [7]): continuous review policy and the periodic review policy. The first policy implies that the stock level will be monitored continuously. Whenever the inventory on hand decreases to a predetermined level, referred to as the *reorder point*, a new order is placed to replenish the inventory level. The placed order is a fixed quantity that minimizes the total inventory costs and is normally called the *economic order quantity*. In the second policy, the inventory level is checked at specific fixed time periods and an order decision is made to complete the stock to a desired upper limit. In this system the inventory level is not monitored at all during the time interval between orders, so it has the advantage of little or no required record keeping. The disadvantage is less direct control. Such system also requires that a new order quantity must be determined each time a periodic order is made. The operating costs taken into account in both inventory policies are the acquisition cost, the holding cost and the shortage cost.

These basic policies can be adapted to take into account special situations such as single or multi-item models with or without a constraint on the total storage space, deterministic or stochastic demands, lost sales, etc. For more details and examples see Ghiani et al. [4] and Nahmias [11].

3 Mathematical Programming Model

Some fundamental questions occur during the design and planning of a warehouse: (i)

1. How much inventory should be kept in the warehouse?
2. How frequently and at what time should the inventory be replenished?
3. What should be the size of the warehouse?
4. What should be the size of the functional areas?
5. Which product must be stored within each functional area and how much must be stored?
6. Etc.

The first two decisions lead to the traditional inventory management problem and the others are some of the decisions concerning warehouse design and planning. Among these decisions the size of the warehouse is probably one of the most important aspects because once warehouse size is determined, it will act as a constraint that may last for a long period of time.

In this section we present a mathematical programming model that integrates both warehouse management decisions and inventory decisions in a way that minimizes the expected inventory and warehouse costs.

3.1 Model assumptions

This work considers a general warehouse configuration that include the following four functional areas: receiving, reserve, forward or picking and shipping. Thus, the following pattern flows are possible:

1. Receiving → Reserve → Shipping
2. Receiving → Reserve → Forward → Shipping
3. Receiving → Forward → Shipping

Flow 1 refers to a pattern that characterises a typical warehouse operation. Products are stored in a reserve area and picking operation is performed as required. Usually, it is assumed that only those products that remain for long periods of time or product quantities used for replenishment of the forward area will be allocated in this area.

Flow 2 is also a typical warehouse operation. Products with this pattern flow are initially stored in the reserve area and then moved to the forward area. This pattern flow is considered for fast picking operations, order consolidation or even to perform value-added operations.

Flow 3 refers to products that go directly to the forward area. This pattern flow is usually seen when there is a need to consolidate large orders.

The next section presents a mathematical model that determines the inventory parameters, the size of the warehouse, the flow to which each product must be assigned and, as result, the size of the functional areas within the warehouse. In developing the model we also assume:

- An inventory control policy based on continuous review policy (reorder point system);
- The forward storage area will be handled through a dedicate storage policy and the reserve area will assume a random storage policy;
- The inventory costs are known;
- The warehouse operation costs are known;
- Customers demand rates are known;
- It is possible to rent external storage area, if necessary.

3.2 Model formulation

The following notation, adapted from Strack and Pochet [13], is used:

Parameters:

i	: Product number ($i = 1, ..., I$)
j	: Number of locations in the forward area ($j = 1, ..., J$)
$CostRepA$	: Cost of advanced replenishment
$CostRepC$	: Cost of concurrent replenishment
$CostR$	: Reception cost for the reserve area
$CostF$	: Reception cost for the forward area
$PickF$	: Picking cost in the forward area
$PickR$	: Picking cost in the reserve area
$CostCapaW$	: Capacity cost of the private warehouse
$CostCapaFW$	: Private warehouse capacity fixed cost
$CostCapaS$	: External capacity cost
$CostCar$	: Inventory carrying cost
$CostAcqu_i$	: Acquisition cost of product i
$CostShort$	: Shortage cost
α_i	: Number of units of product i that can be stored in a single location of the forward area
u_{ij}	: Increase in the expected number of replenishment if an additional location in the forward area is allocated for product i
$E(U_i)$	: Expected value of the demand of product i
$E(p_i)$	: Expected value of the number of picks of product i
L	: Supply lead time
μ_i^L	: Average demand of product i during L
σ_i^L	: Standard deviation of demand of product i during L
d_i^L	: Demand of product i during L
M	: Large positive number

Decision variables:

$$x_{ij} = \begin{cases} 1 & \text{if product } i \text{ has a Flow 2 pattern with at least} \\ & j \text{ locations allocated in the forward area} \\ 0 & \text{otherwise} \end{cases}$$

$$y_i = \begin{cases} 1 & \text{if product } i \text{ has a Flow 3 pattern} \\ 0 & \text{otherwise} \end{cases}$$

$$z_i = \begin{cases} 1 & \text{if product } i \text{ has a Flow 1 pattern} \\ 0 & \text{otherwise} \end{cases}$$

$$w = \begin{cases} 1 & \text{if we have a private warehouse} \\ 0 & \text{otherwise} \end{cases}$$

$CapaW$	: Capacity of the private warehouse
$CapaF$	: Capacity of the forward area
$CapaR$	: Capacity of the reserve
$CapaS$	: External additional storage capacity
Q_i	: Replenishment quantity of product i
r_i	: Reorder point of product i

The general formulation of the model can be stated as:

$$
\begin{aligned}
minimize \quad & \sum_{i=1}^{I} CostRepA \times x_{i1} + \sum_{i=1}^{I}\sum_{j=1}^{J} CostRepC \times u_{ij} \times x_{ij} \\
& + \sum_{i=1}^{I} CostR \times \frac{E(U_i)}{Q_i} \times (z_i + x_{i1}) + \sum_{i=1}^{I} CostF \times \frac{E(U_i)}{Q_i} \times y_i \\
& + \sum_{i=1}^{I} PickF \times E(p_i) \times (x_{i1} + y_i) + \sum_{i=1}^{I} PickR \times E(p_i) \times z_i \\
& + CapaW \times w \times (CostCapaW + CostCapaFW) + CostCapaS \times CapaS \\
& + \sum_{i=1}^{I} Costcar \times \left(\frac{Q_i}{2} + r_i - \mu_i^L\right) + \sum_{i=1}^{I} CAcqu_i \times E(U_i) \\
& + \sum_{i=1}^{I} CostShort \times \frac{E(U_i)}{Q_i} \times \int_{r_i}^{\infty} (d_i^L - r_i) f(d_i^L)\, \mathrm{d}d_i^L \,,
\end{aligned}
\tag{1}
$$

subject to:

$$x_{ij} \leq x_{ij-1} \quad \forall_{ij} : j \geq 2\,, \tag{2}$$

$$x_{i1} + y_i + z_i = 1 \quad \forall_i\,, \tag{3}$$

$$\sum_{i=1}^{I}\left[\left(\sum_{j=1}^{J} x_{ij}\right) + \left(\frac{Q_i + r_i - \mu_i^L}{\alpha_i}\right) \times y_i\right] \leq CapaF\,, \tag{4}$$

$$\sum_{i=1}^{I}\left[\left(\frac{Q_i}{2} + r_i - \mu_i^L\right) \times (z_i + x_{i1}) - \sum_{j=1}^{J} \alpha_i x_{ij}\right] \leq CapaR + CapaS\,, \tag{5}$$

$$CapaF + CapaR \leq CapaW\,, \tag{6}$$

$$CapaW \leq Mw\,, \tag{7}$$

$$CapaW \geq w\,, \tag{8}$$

$$Q_i, r_i \geq 0\,, \tag{9}$$

$$CapaW, CapaF, CapaR, CapaS \geq 0\,, \tag{10}$$

$$x_{ij}, y_i, z_i, w \in \{0,1\}\,. \tag{11}$$

The objective function (1) minimizes the warehouses operating costs and inventory costs per period. Concerning the inventory costs we have taken into account: carrying cost, acquisition cost and shortage cost. The warehouse costs are composed by the reception costs, picking costs, the additional external storage capacity cost (from a public warehouse) and the costs of the owned (private) warehouse. Constraints (2) are sequencing constraints that specify that a j^{th} location can only be assigned to product i if $j-1$ locations have already be assigned. In addition, constraint (3) ensures that each product is assigned to only one of the three flow patterns of the warehouse. Constraints (4)-(5) ensure that the space for the forward and reserve areas are met, and constraint (6) guarantees that the total area in the warehouse is not exceeded. Constraints (7)-(8) serve to include the costs of the private warehouse. Finally, a set of variables must be nonnegative (9)-(10) and another is considered binary (11). Comparatively to the original model, proposed by Strack and Pochet [13], this formulation allows to determine the size and type (own or rented) of storage space required. Depending on which is least expensive a warehouse manager may use only one type or adopt a mixed strategy - both owned and rented storage space. More decision variables ($w, CapaW, CapaF, CapaR$) and warehouse costs ($CostCapaW, CostCapaFW$) were added since the size of the warehouse and of the functional areas are now unknown, and new constraints (6)-(8) were considered.

3.3 Model analysis and methodology

The above model considers inventory and warehouse decisions since it integrates both issues supporting decision makers defining warehouse design and planning. It is a mixed-integer nonlinear programming model (MINLP) with a large number of variables when real cases are considered. To demonstrate the complexity involved in solving the model to optimality, LINGO 12.0 solver was used on an Intel Core 2Duo 1.4 GHz CPU and 3GB RAM. For a randomly generated instance with only 5 products (SKU), a very small size instance, we only were capable of find local optimums within three hours of CPU time. Given the complexity of solving this model, such as Strack and Pochet [13], we used a sequential solution procedure. We decompose our model in two: (i) an inventory management sub model and (ii) a warehouse management sub model. These two sub models are solved sequentially. First we solve the inventory sub model and then the optimal values of the inventory variables are fixed and used to solve the warehouse management sub model.

3.3.1 Inventory management sub model

To obtain the inventory management sub model we have to eliminate costs and constraints related to the warehouse management problem. We obtain a mixed-integer nonlinear programming model under inventory and storage constraints defined as:

$$
\begin{aligned}
minimize \quad & \sum_{i=1}^{I} Costcar \times \left(\frac{Q_i}{2} + r_i - \mu_i^L\right) + \sum_{i=1}^{I} CAcqu_i \times E(U_i) \\
& + \sum_{i=1}^{I} CostShort \times \frac{E(U_i)}{Q_i} \times \int_{r_i}^{\infty} (d_i^L - r_i) f(d_i^L)\, \mathrm{d}d_i^L \\
& + \sum_{i=1}^{I} CostRecp \times \frac{E(U_i)}{Q_i} + CostCapaS \times CapaS \\
& + CapaW \times w \times (CostCapaW + CostCapaFW)\,,
\end{aligned}
\tag{12}
$$

subject to:

$$
\sum_{i=1}^{I} (Q_i + r_i - \mu_i^L) \leq CapaW + CapaS\,, \tag{13}
$$

$$
CapaW \leq Mw\,, \tag{14}
$$

$$
CapaW \geq w\,, \tag{15}
$$

$$
Q_i, r_i \geq 0\,, \tag{16}
$$

$$
CapaW, CapaS \geq 0\,, \tag{17}
$$

$$
w \in \{0, 1\}\,. \tag{18}
$$

To render this model independent of the warehouse decision variables an approximation and a relaxation were performed. First the objective function (12) was approximated using $CostRecp$ as the reception cost which is independent on the pattern flow taken by the products (defined as an average of the reserve and forward reception costs). Secondly, both reserve and forward areas were relaxed into one global capacity constraint (13) for the entire warehouse. The objective function and the global capacity constraint so obtained are independent of the flow patterns taken by the different products. Nevertheless, the inventory variables will be dependent of the storage capacity and, in a certain way, we can consider that this sub model integrates an inventory decision and a warehouse size decision.

3.3.2 Warehouse management sub model

Warehouse management sub model is obtained fixing the inventory variables (Q_i) and the capacity warehouse variable ($CapaW$), based on the solution of the inventory sub model and eliminating costs and constraints related with the inventory problem. The resulting model is a mixed-integer problem defined as follows:

$$
\begin{aligned}
\text{minimize} \quad & \sum_{i=1}^{I} CostRepA \times x_{i1} + \sum_{i=1}^{I}\sum_{j=1}^{J} CostRepC \times u_{ij} \times x_{ij} \\
& + \sum_{i=1}^{I} CostR \times \frac{E(U_i)}{Q_i} \times (z_i + x_{i1}) + \sum_{i=1}^{I} CostF \times \frac{E(U_i)}{Q_i} \times y_i \\
& + \sum_{i=1}^{I} PickF \times E(p_i) \times (x_{i1} + y_i) + \sum_{i=1}^{I} PickR \times E(p_i) \times z_i \\
& + CostCapaS \times CapaS
\end{aligned} \tag{19}
$$

subject to:

$$x_{ij} \leq x_{ij-1} \quad \forall_{ij} : j \geq 2\,, \tag{20}$$

$$x_{i1} + y_i + z_i = 1 \quad \forall_i\,, \tag{21}$$

$$\sum_{i=1}^{I}\left[\left(\sum_{j=1}^{J} x_{ij}\right) + \left(\frac{Q_i + r_i - \mu_i^L}{\alpha_i}\right) \times y_i\right] \leq CapaF\,, \tag{22}$$

$$\sum_{i=1}^{I}\left[\left(\frac{Q_i}{2} + r_i - \mu_i^L\right) \times (z_i + x_{i1}) - \sum_{j=1}^{J} \alpha_i x_{ij}\right] \leq CapaR + CapaS\,, \tag{23}$$

$$CapaF + CapaR \leq CapaW\,, \tag{24}$$

$$CapaF, CapaR, CapaS \geq 0\,, \tag{25}$$

$$x_{ij}, y_i, z_i \in \{0, 1\}\,. \tag{26}$$

In the warehouse sub model the flow pattern variables and the size of both reserve and forward functional areas will be optimized and the optimal value of the external additional storage capacity ($CapaS$) is re-optimised. The mixed-integer model will be solved using a Branch-and-Bound procedure.

4 Computational results

In this section, numerical results of a preliminary study are presented. Table 1 shows parameter values used to generate the testing problems. Instances for different scenarios (see Table 2) were randomly generated to assess the behaviour of the model when the number of products increases.

Table 1: Parameter values for the numerical examples.

Parameter	Value
$CostRepA$	5
$CostRepC$	20
$CostR$	5
$CostF$	8
$PickF$	3
$PickR$	10
$CostRecp$	5
$CostCar$	3
$CostShort$	50
$CostCapaS$	20
$CostCapaW$	3
$CostCapaFW$	10
$E(U_i)$	Uniform $[1, 50]$
$E(p_i)$	Uniform $[1, 5]$
d_i^L	$N(\mu_i^L, \sigma_i^L)$

Table 2: Analysed scenarios.

Scenario	I	II	III	IV	V
SKU [units]	10	100	500	1000	5000

Table 3: Inventory sub model computational results.

	Scenario				
	I	II	III	IV	V
Total variables	13	103	503	1003	5003
Nonlinear variables[a]	12	102	502	1002	5002
Iterations	203	647	1762	4020	16021
CPU time $[mm : ss]$	00 : 03	00 : 11	00 : 41	01 : 20	14 : 01
State	Global Opt.	Global Opt.	Global Opt.	Global Opt.	Global Opt.

[a] Variables involved in the nonlinear relationships of the model.

Table 4: Warehouse management sub model computational results.

	Scenario				
	I	II	III	IV	V[b]
Total variables	43	2203	38503	152003	3760003
Binary variables	40	2200	38500	152000	3760000
No. of constraints	24	2004	37504	150004	3750004
Iterations	6	20	849	2798	–
CPU time $[mm : ss]$	00 : 02	00 : 05	05 : 03	11 : 03	–
State	Global Opt.	Global Opt.	Global Opt.	Global Opt.	–

[b] Due to the size of the generator matrix, the computer did not have sufficient memory.

The computational results for the different testing cases are shown in Table 3 and Table 4. As it can be seen, with exception of Scenario V it was possible to solve to optimality all the others four scenarios in a very satisfactory time. Nevertheless, the computational time of LINGO solver rises as the problem sizes increases. For large instances (Scenario V) the number of variables and constraints of the warehouse sub model increases considerably. Consequently using the branch-and-bound algorithm takes significant computational time and memory. In our tests it was not possible to obtain the details for Scenario V due to the insufficient memory of the used computer. It appears that more sophisticated techniques can be explored, e.g. decomposition techniques, to circumvent the memory problem encountered with the branch-and-bound algorithm and solve very large instances.

5 Conclusions and future work

Most of the times warehouse design and planning decisions are taken independently. In fact, having a single decision model capable of integrate several decisions is a very complex task due to the enormous amount of existing alternatives, and to the existence of various and often conflicting objectives through and out of the warehouse. In this work a mathematical model for the warehouse design and planning problem was extended. The proposed model jointly integrates: (i) the size of the warehouse, including the size of both reserve and forward functional areas; (ii) the external storage additional capacity, if needed; (iii) the assignment and allocation of SKU to storage areas and; (iv) the replenishment quantities and reorder points of products to be stored.

Due to the complexity of the obtained analytical model, an optimal global solution was definitely difficult to achieve. For this reason our global model was decompose in two sub models, which were solved using a hierarchical sequential approach. First a nonlinear inventory model was solved taking into account the inventory variables and a global capacity constraint for the warehouse. Secondly, a warehouse management sub model was obtained fixing the solution of the inventory sub model. This sub model allowed us to determine the pattern flow for each product, the sizes of the functional areas and to re-optimise the value of the external additional storage capacity.

Computational results of the preliminary study suggest that it is possible to solve to optimality both sub models. Nevertheless, one must note that the computational time rises considerably as the problem sizes increases.

Even though the presented model integrates important decisions concerning the design and planning of a warehouse, many other decisions were not included, such us the storage policy, the picking and routing strategies, etc. However gathering in a single model several decisions leads us to very complex models difficult to treat and analytically solve. For that reason we believe that simulation technique can be used to validate the models and to incorporate dynamic aspects not yet included. For example, we can use the solution of the analytical model and then simulation can be used to introduce demand fluctuations or operational decisions related with the storage, picking or routing policies.

In summary, despite some advances in integrated approaches, further research focusing integrated models where different processes in the warehouse are jointly considered (and its corresponding dynamic nature), is still required. Given the prevalence of warehouses in the supply chain networks we believe that such research achievements can have a significant impact in the supply chain performance.

References

[1] Baker, P. and Canessa, M.: Warehouse design: A structured approach. European Journal of Operational Research 193, 425-436 (2009).

[2] Bartholdi, J. and Hackman, S.: Warehouse & Distribution Science (2006). Release 0.76. (http://www.warehouse-science.com)

[3] Geraldes, C., Carvalho, S. and Pereira, G.: A warehouse design decision model. Proceedings of the International Engineering Management Conference 2008, IEMC-Europe 2008, Estoril, Portugal. ISBN: 978-1-4244-2289-0. IEEE Catalog Number: CFP08EMS (2008).

[4] Ghiani, G., Laporte, G. and Musmanno, R.: Introduction to Logistics Systems Planning and Control. John Wiley & Sons Ltd, England (2004).

[5] Gu, J., Goetschalckx, M., McGinnis, L. F.: Research on warehouse operation: A comprehensive review. European Journal of Operational Research 177, 1-21 (2007).

[6] Gray, A., Karmarkar, U. and Seidman, A.: Design and operation of an order-consolidation warehouse: models and application. European Journal of Operational Research 58, 14-36 (1992).

[7] Hadley, G. and Whitin, T.: Analysis of Inventory Systems. Prentice-Hall. Englewood Cliffs, New Jersey (1963).

[8] Hassan, M.: A framework for the design of warehouse layout. Facilities 20, 13/14, 432-440 (2002).

[9] Heragu, S. Du, L., Mantel, R.J. and Schuur, P.C.: Mathematical model for warehouse design and product allocation. International Journal of Production Research 43, 2, 432-440 (2005).

[10] Liu, C.: Clustering techniques for stock location and order-picking in a distribution center. Computers & Operations Research 26, 989-1002 (1999).

[11] Nahmias, S.: Production and Operation Analysis, third edition. McGraw-Hill International Editions (1997).

[12] Rouwenhorst, B., Reuter, B., Stokrahm, V., van Houtum, G.,Mantel, R. and Zijm, W.: Warehouse design and control: Framework and literature review. European Journal of Operational Research 122, 515-533 (2000).

[13] Strack, G., Pochet, Y.: An integrated model for warehouse design and planning. European Journal of Operational Research 204, 35-50 (2010).

[14] Tompkins, J.A., White, J.A., Bozer, Y.A., Tanchoco, J.M.A.: Facilities Planning. John Wiley & Sons, New York (2003).

[15] Van den Berg, J.: A literature survey on planning and control of warehousing systems. IEEE Transactions 31, 751-762 (1999).

Assessing residential building sustainability in the operation phase

I. M. Horta*, A. S. Camanho*, T. G. Dias*

* *Faculdade de Engenharia da Universidade do Porto, Rua Dr. Roberto Frias, 4200-465 Porto, Portugal*
imhorta@fe.up.pt, acamanho@fe.up.pt, tgalvao@fe.up.pt

Abstract

The purpose of this paper is to assess residential buildings' sustainability during the operation phase, focusing on environmental aspects concerning the consumption of resources. The assessment is carried out at a municipality level, enabling decision makers to know the relative position of their municipalities compared to others. In addition, the study identifies the factors associated with better levels of municipality performance, and quantifies the extent of their effects. The study uses an enhanced stochastic frontier panel model to evaluate municipalities' performance over time. The analysis is based on data of energy, water and materials consumption in Lisbon municipalities between 2003-2009.

Key words: Stochastic Frontier Analysis, Sustainability, Construction Industry

1 Introduction

The construction industry (CI) is one of the largest sectors in terms of Gross Domestic Product and employment in the world, and it highly contributes to socio-economic development. The CI produces residential and nonresidential buildings (e.g. industrial, commercial or educational), as well as the infrastructures for several economic activities (e.g. transport, communications, energy, water or solid waste management). Due to the considerable adverse impact of CI activity on the environment, sustainable construction practices have attracted increased interest in recent years.

The buildings sector represents the most critical segment in terms of CI sustainability practices. The life cycle of buildings is wide (around 50 years) and includes various phases (e.g. production or extraction of materials, design, construction, operation and maintenance). All phases may have a severe impact on the environment. According to [EuroACE, 2012], the 210 million buildings across Europe are responsible for 40% of Europe's total energy consumption and 36% of CO_2 emissions. In addition, the construction of buildings requires approximately 50% of total raw material consumption and 40% of total materials' waste.

Traditionally, the buildings sector was mostly engaged with the improvement of energy efficiency. For instance, the European directives (2002/91/CE and 2010/31/EU) oblige buildings to be certified in terms of energy efficiency, which involves regular inspections of air conditioning and heating systems. More recently, the buildings sector is addressing the challenge of sustainability in other areas, such as water, materials, indoor air quality, or waste. The broader sustainability certification of buildings is currently carried out on a voluntary basis. However, due to the growing interest on sustainability, overall buildings' sustainability certification may become mandatory in the near future. Therefore, it is essential to develop robust methods to investigate the best practices and performance of buildings in terms of sustainability.

The purpose of this research is to assess residential buildings' sustainability during the operation phase, focusing on environmental aspects concerning the consumption of resources. This assessment is carried out at a municipality level, enabling to inform decision makers on the position of their municipalities compared to others. In addition, the study identifies the factors associated with better levels of municipality performance and quantifies the extent of their effects. From a methodological perspective, this research uses the Stochastic Frontier Analysis (SFA) technique to evaluate municipalities' performance over time. SFA provides a single overall measure of performance for each municipality in a given year by comparing it with the best observed performance.

The performance assessment described in this paper used data from the Lisbon region (Lisbon city and 17 surrounding municipalities) covering the time period 2003-2009. The model included both endogenous and exogenous indicators to evaluate adequately municipalities in terms of buildings' sustainability.

The endogenous variables concern the major resources used in buildings' operation phase, i.e. the consumption of energy, water, and materials for maintenance in each municipality. The exogenous variables control for contextual conditions and refer to the buildings' average age, the quality of buildings, the environmental policy expenditure, and the buildings' sustainability in the construction phase, all measured at the municipality level.

The remainder of this paper is organized as follows. Section 2 reviews the literature concerning the assessment of sustainability in construction. Section 3 explains the method used in the empirical part of the paper. Section 4 describes the variables and the sample used in the study. Section 5 discusses the results obtained. The last section concludes and suggests topics for future research.

2 Background

As mentioned by [Vatalis et al., 2011], sustainable construction relies on four main pillars: environmental, social, economic, and technical. Environmental sustainability concerns environmental protection, encouraging, for example, the efficient use of resources. Social sustainability concerns the quality of human life and the human living environment in terms of safety, comfort, and health. Economic sustainability is related to the financial return of the project taking into account the perspectives of all stakeholders involved in the process, such as clients, construction players or government. Technical sustainability is related to building components and construction technologies in terms of durability, quality and usability.

The assessment of environmental sustainability in construction has attracted particular attention in recent years (see [Forsberg and Malmborg, 2004] and [Ding, 2008] for a literature review). From the literature review, it can be observed that the research conducted on environmental sustainability mostly focuses on the evaluation of buildings.

Traditionally, the environmental evaluation of buildings was based on a single criterion, such as energy, water or materials usage. For instance, [Mwasha et al., 2011] investigated the most appropriate energy performance indicators for modeling the performance of the residential building envelope. [Ilha et al., 2009] reviewed the main issues in terms of water conservation to be considered on environmental evaluations of buildings. Other studies focused on the environmental impact of specific building materials, such as marble [Traverso et al., 2010], facade materials [Kim, 2011], concrete [Bjorklund and Tillman, 1997], clay bricks [Koroneos and Dompros, 2007], and insulating stone wool [Schmidt et al., 2004]. However, sustainability issues became more demanding and monitoring buildings performance merely based on a single criteria resulted in a limited assessment.

More recently, comprehensive building assessment models appeared in the literature. These models typically cover a broad set of environmental criteria using performance indicators. The scope of analysis usually covers various phases of the building life cycle (e.g. design, construction or operation phase), and different types of buildings (e.g. residential or industrial buildings). The first building assessment model was launched in 1990, in the United Kingdom, and is called the BRE Environmental Assessment Method (BREEAM). The BREEAM system includes the assessment of new and existing buildings of any type (offices, supermarkets, houses, light and heavy industrial buildings). Nowadays, the BREEAM serves as a support to the development of many other building assessment systems in different countries, such as in Canada, Australia, Hong Kong. A few other models to assess building environmental performance appeared later. For instance, the Leadership in Energy and Environmental Design Green Building Rating System (LEED) was developed in 2002 in the United States, the Comprehensive Assessment System for Building Environmental Efficiency (CASBEE) was developed in 2004 in Japan, and the Sustainable Building Tool (SBTool) was developed in 1995 aiming to carry out an international tool for building assessment. In addition to the aforementioned models, there are some studies in the literature that focus on the development of models to assess specific types of buildings, such as industrial buildings [San-Jose et al., 2007], apartment buildings [Kim et al., 2005], or intelligent buildings [ALwaer and Clements-Croome, 2010].

Nowadays, the literature on environmental assessment in the CI has attempted evaluations at an urban scale. This is a consequence of today's urbanization context, which requires adequate infrastructure, buildings and utilities leading to enhancements in quality of life, urban governance, and environmental quality. The most well known tools developed to assess CI in an urban context are typically extensions of the building environmental assessment models (BREEAM Communities, LEED for Neighborhood Development, CASBEE for Urban Development, and SBTool Generic). There are other studies in the literature worth mentioning as they emphasize the need of models to enable an urban evaluation [Conte and Monno, 2012, Huang and Hsu, 2011, Jones et al., 2007]. However, as mentioned by [Haapio, 2012], the development of these tools is recent and thus publications related to this topic are still scarce.

This paper is to the best of our knowledge the first to propose the use of frontier methods, specifically the SFA technique, to evaluate residential buildings' sustainability in terms of the consumption of energy, water and materials during the operation stage. This work intends to provide the foundations for benchmarking exercises at a municipality level.

3 Stochastic Frontier Analysis

Stochastic Frontier Analysis was first proposed by [Aigner et al., 1977] and [Meeusen and van den Broeck, 1977]. SFA is a parametric technique, such that the production function is specified using a mathematical form, usually the Cobb-Douglas or translog functional forms. These are specified a-priori and their parameters are estimated from empirical data. One of the major strengths of SFA is to assume that deviations from the estimated frontier are composed by inefficiency and random error, such that it is possible to account for random noise or measurement errors in the data.

In this paper, the SFA technique is used to evaluate municipalities' performance in terms of residential buildings sustainability. Due to the availability of data from different time periods and the need of including exogenous factors to control for contextual conditions, we used a panel data stochastic frontier model accounting for exogenous variables as proposed by [Battese and Coelli, 1995]. To specify the model, we use a Cobb-Douglas function due to its simplicity in terms of parameters to be estimated. This is an important feature, as the number of observations is limited (only 18 municipalities), and we believe this functional form is suitable for an exploratory study. The [Battese and Coelli, 1995] model, using a Cobb-Douglas function, is specified as follows:

$$y_{jt} = \beta_0 + \sum_{i=1}^{s} \beta_i \; x_{ijt} + v_{jt} - u_{jt} \tag{1}$$

y_{jt} represents the logarithm of the output of the j^{th} unit in time period t ($t = 1, \ldots, l$), x_{ijt} is a vector of the logarithm of the input variables i ($i = 1, \ldots, s$) used by the j^{th} unit in time period t. β_0 is an intercept and β_i is a vector of unknown parameters to be estimated. v_{jt} is a random variable which is assumed to account for statistical noise and to be independently and identically distributed $N(0, \sigma_v^2)$. u_{jt} is a non-negative random variable, which is assumed to account for technical inefficiency. It is obtained by truncation at zero of the $N(u_{jt}, \sigma_u^2)$ distribution. u_{jt} is assumed to be independent of v_{jt}. The technical inefficiency, u_{jt}, is specified as follows:

$$u_{jt} = z_{jt}\delta + w_{jt} \tag{2}$$

z_{jt} is a vector of exogenous variables, δ is a vector of unknown parameters to be estimated, and w_{jt} is a random variable that is obtained by truncation of the $N(0, \sigma^2)$, such that u_{jt} is non-negative (i.e. $w_{jt} \geq -z_{jt}\delta$). The technical efficiency for the j^{th} unit in period t is defined by equation (3).

$$TE_{jt} = exp(-u_{jt}) = exp(-z_{jt}\delta - w_{jt}) \tag{3}$$

To estimate the parameters of the stochastic frontier model (1) and the technical inefficiency model (2), we used the computer program FRONTIER version 4.1 [Coelli, 1996], which applies the maximum likelihood method. For more details on the likelihood function and its partial derivatives see [Battese and Coelli, 1993].

4 Variables and sample

4.1 Input and output variables

The input variables used in the SFA model represent the major natural resources consumed during residential buildings' operation phase. Based on the review of the building environmental assessment models (BREEAM, LEED, CASBEE, and SBTool), it can be concluded that the most critical natural resources to evaluate residential buildings' sustainability include energy, water, and materials. Therefore, the SFA model specified includes inputs representing these resources. Energy corresponds to the electricity and natural gas consumed by residential buildings in each municipality and year. Materials includes fossil fuels, metals, non-metallic minerals and biomass materials used for rehabilitation and maintenance of residential buildings in each municipality and year. Water is the quantity of water consumed by residential buildings in each municipality and year. On average, Lisbon municipalities consumed approximately 17080 K Nm^3 of energy, 30387 tonnes of materials and 10400 m^3 of water during the period analyzed.

The output variable of the SFA model correspond to the total number of residential buildings in each municipality and year. Note that the number of floors by building is relatively similar among Lisbon municipalities (mean equals 3.79 and standard deviation equals 0.64) according to the 2001 census. This enables to undertake a fair comparison between municipalities. According to sample analyzed, on average, there are approximately 23400 residential buildings in Lisbon region, which corresponds to 8 residential buildings per km^2 during the period analyzed. With this specification of the inputs and output for the SFA model, a higher performance score means that the municipality is able to consume less resources per building than its peers.

4.2 Exogenous variables

Concerning the factors that may impact buildings's sustainability in each municipality, we analyzed with particular detail the effect of four exogenous variables: the buildings' average age, the quality of buildings, the environmental policy expenditure, and the buildings' sustainability in the construction phase, measured at the municipality level.

Buildings' average age was included in the model to check if municipalities with buildings constructed more recently tend to be more resource efficient. In Portugal, 56% of the residential buildings were constructed between 1961-1991. Rehabilitation has a small expression in the CI activity, corresponding to about 10% of all construction projects (www.ine.pt). Older and degraded buildings tend to consume more resources, such as energy for heating, cooling and lighting. Therefore, it is expected that buildings' average age strongly influences the municipalities' performance. The buildings' average age in Lisbon region is 36 years old, which corresponds to an advanced stage of the building life cycle. In addition, the municipalities exhibit similar buildings' average age, given the small value of the standard deviation (7.7 years).

Building quality was analyzed to test if municipalities with buildings exhibiting better quality also have better performance levels. Building quality is typically associated with the use of superior materials, equipments and technologies. High quality buildings include modern appliances aiming to reduce the consumption of natural resources. For example, many new buildings are designed with the maximum exposure to the sun, incorporating large south-facing windows to capture the energy in winter, and overhangs to shade the windows from the sun in summer. Windows are also strategically placed around the buildings to make use of natural light, reducing the need for artificial lighting during the day. Some buildings have included green roofs for stormwater management, or equipments such as a dual plumbing, ultra-low flush toilets and low-flow shower heads to minimize the waste of water. As the data available did not allow a more detailed quantification of building quality we used as a proxy the average value of bank evaluation. Municipalities with high average values of bank evaluation, and thus with more expensive buildings are likely to accommodate buildings with higher construction quality. The average value of bank evaluation was around 1400 euros per m^2 in the Lisbon area.

Buildings' sustainability in construction phase was included in the model to test if municipalities accommodating buildings with higher performance in the construction phase, i.e. buildings constructed with sustainable principles (e.g. using 'green' materials), tend to be good performers in the operation phase. The buildings' sustainability in the construction phase measures the quantity of resources consumed (metals, non metallic minerals, fossil fuels, biomass) per area of completed buildings in each municipality. This variable was defined in the study developed by [Horta et al., 2012] concerning the assessment of CI sustainability in Lisbon municipalities. Such variable varies between 0% and 100%, where 100% indicates an efficient municipality in terms of buildings' sustainability in the construction phase.

Environmental policy expenditure was included to test if municipalities that invest in environmental policies are able to be more resource efficient. The environmental projects promoted by municipalities include wastewater management, protection of groundwater and surface water, protection of biodiversity and landscape, and research and development activities. Municipalities with environmental concerns are also likely to have environmentally friendly and well-informed inhabitants. For example, homeowners can reduce energy consumption by insulating water heaters, repairing leaky faucets and toilets, sealing windows, attics and walls, by closing the drapes on windows during sunny summer days and after sunset in winter to maintain a more constant temperature, by using the energy-saving setting on appliances, or by replacing incandescent bulbs with new bulb models. The average expenditure in environmental policies was about 64 euros per inhabitant in Lisbon municipalities.

4.3 Sample

This paper analyzed a sample of 18 municipalities, including Lisbon city (the Portuguese capital) and the 17 surrounding municipalities. The longitudinal assessment reported covers the time period between 2003 and 2009. The Lisbon region is a medium-sized European metropolitan area, with more than 2.8 million residents according to the 2011 census. It assumes a central importance in the national economy. The data on input and output variables came from the National Statistics Institute (INE) database, with the exception of materials consumption that came from the study carried out by [Niza et al., 2009] concerning the assessment of urban metabolism of Lisbon. The data related to exogenous variables come from the INE database, with the exception of data on buildings' sustainability in the construction stage that was gathered from [Horta et al., 2012].

5 Empirical Results

The first step of the assessment was to estimate the parameters of the Stochastic frontier model (1) and the inefficiency model (2). Table 1 reports the maximum-likelihood estimates of the parameters of the models, the corresponding standard errors, and the test statistics. The total number of observations included was 126, corresponding to the 18 municipalities in the 7 years analyzed. The estimates confirm that the inclusion of the inefficiency effects is highly significant in the analysis of buildings' sustainability, as the estimate for the variance parameter, $\gamma \equiv \sigma_u^2/(\sigma_v^2+\sigma_u^2)$ of the likelihood function is close to one (γ=0.999). This indicates that deviations from the frontier are due to inefficiency u_{jt} rather than random errors v_{jt}.

Table 1: Estimates of the Stochastic frontier and inefficiency model

Variable	Coef.	Std. Error
Stochastic frontier model		
ln(water)	0.072*	0.041
ln(energy)	0.612**	0.038
ln(materials)	0.014**	0.003
constant	3.918**	0.166
Inefficiency model		
Building age	0.019**	0.002
Building sust. construction stage	-0.425**	0.128
Building quality	-0.142*	0.081
Environmental expenditure	-0.003**	0.000
constant	0.335**	0.133

*Significant at 5%; **Significant at 1%.

From the results of Table 1, it is possible to observe that the coefficients related to the input variables of the SFA model are positive and significant at a 5% level. This means that the number of buildings goes up as the consumption of water, energy and materials increase, as expected.

Concerning the inefficiency model results reported in the lower part of Table 1, we can conclude that building average age has a significant impact on the buildings' sustainability. The impact is positive meaning that municipalities with older buildings tend to be less resource efficient. This confirms that older and more degraded buildings impact adversely on building sustainability in the operation phase.

Building sustainability in the construction phase affects significantly the building sustainability in the operation phase. The coefficient is negative meaning that municipalities with high number of sustainable buildings in the construction stage are likely to have more efficient buildings in the operation stage. This result is particularly important to encourage the use of sustainable principles since the initial phases of construction projects.

Building quality is an influential factor to explain buildings' sustainability. The significant and negative coefficient associated to building quality means that municipalities with high average value of bank evaluation tend to have resource efficient buildings. This is as expected, since high quality buildings tend to use recent appliances and technologies fostering efficient operation.

Environmental policy expenditure has also an impact on buildings' sustainability. The negative and significant coefficient means that municipalities investing in environmental policies tend to be more resource efficient. This finding suggests that municipalities should focus on the implementation of environmental friendly plans and on the awareness of their inhabitants in order to operate efficiently.

The second step of the assessment was to assess the technical efficiency of the municipalities in terms of their buildings' sustainability. Table 2 reports the efficiency scores obtained using expression (3), the grand mean and the standard deviation for each year and municipality.

Table 2: Efficiency results

Municipality	2003	2004	2005	2006	2007	2008	2009	Average	Std. Dev.
Alcochete	0.50	0.47	0.43	0.44	0.43	0.45	0.48	0.46	0.03
Almada	0.78	0.76	0.75	0.80	0.79	0.78	0.73	0.77	0.02
Amadora	0.40	0.40	0.39	0.39	0.40	0.42	0.40	0.40	0.01
Barreiro	0.45	0.45	0.43	0.46	0.45	0.49	0.54	0.47	0.04
Cascais	0.64	0.62	0.59	0.60	0.62	0.64	0.60	0.62	0.02
Lisboa	0.46	0.46	0.44	0.46	0.46	0.47	0.45	0.46	0.01
Loures	0.65	0.63	0.62	0.61	0.69	0.72	0.70	0.66	0.04
Mafra	0.99	0.97	0.92	1.00	0.88	0.89	0.87	0.93	0.06
Moita	0.56	0.55	0.53	0.54	0.53	0.52	0.54	0.54	0.01
Montijo	0.73	0.72	0.67	0.65	0.63	0.65	0.62	0.67	0.04
Odivelas	0.44	0.43	0.43	0.44	0.46	0.46	0.43	0.44	0.01
Oeiras	0.38	0.42	0.40	0.40	0.42	0.39	0.41	0.40	0.02
Palmela	0.89	0.87	0.85	0.82	0.77	0.81	0.79	0.83	0.05
Seixal	0.66	0.65	0.71	0.65	0.64	0.68	0.63	0.66	0.03
Sesimbra	0.96	0.85	0.82	0.94	0.91	0.92	0.81	0.89	0.06
Setubal	0.65	0.64	0.62	0.63	0.62	0.64	0.61	0.63	0.02
Sintra	0.76	0.74	0.70	0.69	0.68	0.71	0.70	0.71	0.03
Vila Franca Xira	0.55	0.48	0.47	0.53	0.46	0.46	0.47	0.49	0.04
Average	0.64	0.62	0.60	0.61	0.60	0.62	0.60		
Std. Dev.	0.19	0.17	0.17	0.18	0.16	0.17	0.14		

Analyzing the performance results over the years, we can observe that there are no considerable differences in terms of municipalities' performance during the period analyzed. On average, the municipalities' performance is 0.61, with very small variations from year to year. This may suggest that in the last decade none of the municipalities was able to establish a fruitful strategy for performance improvement in residential buildings' sustainability. Comparing Lisbon municipalities, it is possible to verify that they are relatively heterogenous in terms of buildings' sustainability. This result indicates that there is a large potential for improvement in terms of buildings' sustainability practices in the operation phase.

Next, we explored with particular detail the characteristics of the best municipalities that may show how the remaining ones can improve their performance. For this purpose, we analyzed in detail the consumption of water, energy and materials of benchmark municipalities (i.e. municipalities with a performance score in the 75^{th} percentile). The municipalities in the 75^{th} percentile (Almada, Mafra, Palmela, and Sesimbra) exhibit an average performance score higher than 0.72 (see the penultimate column of Table 2). Figure 1 shows the average values of the input variables for these municipalities and for the other municipalities in the sample. Note that in Figure 1 the average values of the inputs are divided by the number of buildings (output variable), and normalized using the average values of the worst performing municipalities as a reference.

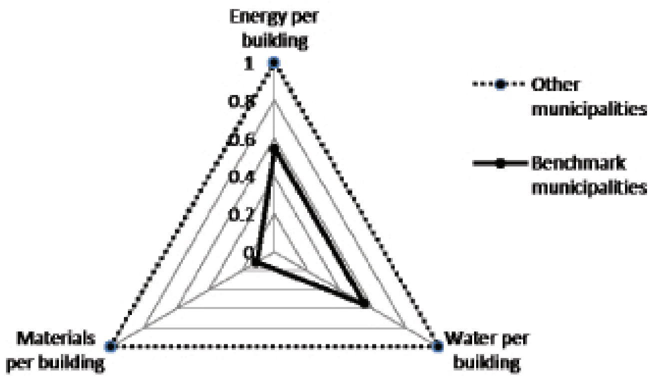


Figure 1: Comparison of resources consumption between best and worst performing municipalities

Figure 1 provides an intuitive picture of the municipalities' performance in terms of resource consumption per building, showing how the best performing and the remaining municipalities stand against each other. In particular, the largest scope for improvement in the worst performing municipalities concerns materials usage, followed by energy and water consumption. This information can be very relevant to drive decision makers' strategies. In particular, municipalities should redefine their strategies towards a more efficient consumption of materials used for rehabilitation and maintenance purposes, as they are responsible for constructing and maintaining a large number of buildings.

6 Conclusions

The purpose of this research is to develop a quantitative methodology to evaluate residential buildings' sustainability during the operation stage. The assessment focuses on a resource consumption perspective at a municipality level. This perspective of construction sustainability is a critical issue in today's society, and has not yet been addressed in the literature. In addition, to the best of our knowledge, this paper is the first to propose the use of the SFA technique to measure buildings' sustainability at an urban level. The development of advanced methodologies is of particular interest in the field of construction sustainability in order to assist the design of realistic strategies to implement and disseminate sustainable practices. The ability to demonstrate and quantify issues that decision makers' might only know in a general and qualitative way makes the methodology developed particularly valuable.

The methodology proposed was applied to a real world context by analyzing data of energy, water and materials consumption in Lisbon municipalities between 2003-2009. The empirical results indicate that there is a large potential to improve the sustainability of residential buildings during the operation phase. In particular, the results of this paper suggest that the majority of municipalities should focus their attention to reduce the consumption of materials, followed by energy and water consumption to improve buildings' sustainability. The results point that the municipalities' performance tend to improve with the quality of buildings, the environmental policy expenditure, the buildings' sustainability in the construction phase, and the newness of buildings.

Finally, it would be interesting to complement the methodology developed in this paper with a detailed analysis of the benchmark municipalities in order to identify the best practices dealing with water, energy, materials, and waste in the building sector. This would contribute to promote performance improvement at a municipality level. In addition, other interesting avenues for future research could include the application of the methodology developed in other metropolitan areas, or even for assessments at regional or country levels, enabling international comparisons and dissemination of best practices worldwide.

7 Acknowledgments

The authors are grateful to Leonardo Rosado and Samuel Niza from Instituto Superior Técnico, in Lisbon, for enabling access to data used in this study. The authors also acknowledge the financial support of the Portuguese Foundation for Science and Technology (FCT) through iTEAM project (MIT-Pt/SES-SUES/0041/2008) and MeSUr project (PTDC/SEN-ENR/111710/2009).

References

[Aigner et al., 1977] Aigner, D., Lovell, C., and Schmidt, P. (1977). Formulation and estimation of stochastic frontier production models. *Journal of Econometrics*, 6(1):21–37.

[ALwaer and Clements-Croome, 2010] ALwaer, H. and Clements-Croome, D. (2010). Key performance indicators (KPIs) and priority setting in using the multi-attribute approach for assessing sustainable intelligent buildings. *Building and Environment*, 45(4):799–807.

[Battese and Coelli, 1993] Battese, G. E. and Coelli, T. J. (1993). A stochastic frontier production function incorporating a model for technical inefficency effects. Working papers in econometrics and applied statistics; no. 69, Department of Econometrics, University of New England, Armidale.

[Battese and Coelli, 1995] Battese, G. E. and Coelli, T. J. (1995). A model for technical inefficiency effects in a stochastic frontier production function for panel data. *Empirical Economics*, 20(2):325–32.

[Bjorklund and Tillman, 1997] Bjorklund, T. and Tillman, A. M. (1997). LCA of building frame structures environmental impact over the life cycle of wooden and concrete frames. Technical environmental planning report, Chalmers University of Technology, Uppsala, Sweden.

[Coelli, 1996] Coelli, T. J. (1996). A guide to FRONTIER version 4.1: A computer program for stochastic frontier production and cost function estimation.

[Conte and Monno, 2012] Conte, E. and Monno, V. (2012). Beyond the buildingcentric approach: A vision for an integrated evaluation of sustainable buildings. *Environmental Impact Assessment Review*, 34(0):31–40.

[Ding, 2008] Ding, G. K. (2008). Sustainable construction-the role of environmental assessment tools. *Journal of Environmental Management*, 86(3):451–464.

[EuroACE, 2012] EuroACE (2012). Energy efficiency directive. Technical report, The European Alliance of Companies for Energy Efficiency in Buildings, Brussels, Belgium.

[Forsberg and Malmborg, 2004] Forsberg, A. and Malmborg, F. V. (2004). Tools for environmental assessment of the built environment. *Building and Environment*, 39(2):223–228.

[Haapio, 2012] Haapio, A. (2012). Towards sustainable urban communities. *Environmental Impact Assessment Review*, 32(1):165–169.

[Horta et al., 2012] Horta, I. M., Camanho, A. S., and Galvao, T. D. (2012). Towards sustainability in construction: a DEA assessment at an urban level. In *25th European Conference on Operational Research (EURO 2012)*, Vilnius, Lithuania.

[Huang and Hsu, 2011] Huang, R. Y. and Hsu, W. T. (2011). Framework development for state-level appraisal indicators of sustainable construction. *Civil Engineering and Environmental Systems*, 28(2):143–164.

[Ilha et al., 2009] Ilha, M. S. O., Oliveira, L. H., and Goncalves, O. M. (2009). Environmental assessment of residential buildings with an emphasis on water conservation. *Building Services Engineering Research & Technology*, 30(1):15–26.

[Jones et al., 2007] Jones, P., Patterson, J., and Lannon, S. (2007). Modelling the built environment at an urban scale - energy and health impacts in relation to housing. *Landscape and Urban Planning*, 83(1):39–49.

[Kim, 2011] Kim, K. H. (2011). A comparative life cycle assessment of a transparent composite facade system and a glass curtain wall system. *Energy and Buildings*, 43(12):3436–3445.

[Kim et al., 2005] Kim, S. S., Yang, I. H., Yeo, M. S., and Kim, K. W. (2005). Development of a housing performance evaluation model for multi-family residential buildings in korea. *Building and Environment*, 40(8):1103–1116.

[Koroneos and Dompros, 2007] Koroneos, C. and Dompros, A. (2007). Environmental assessment of brick production in greece. *Building and Environment*, 42(5):2114–2123.

[Meeusen and van den Broeck, 1977] Meeusen, W. and van den Broeck, J. (1977). Efficiency estimation from cobb-douglas production functions with composed error. *International Economic Review*, 18(2):435–44.

[Mwasha et al., 2011] Mwasha, A., Williams, R. G., and Iwaro, J. (2011). Modeling the performance of residential building envelope: The role of sustainable energy performance indicators. *Energy and Buildings*, 43(9):2108–2117.

[Niza et al., 2009] Niza, S., Rosado, L., and Ferrao, P. (2009). Urban metabolism: Methodological advances in urban material flow accounting based on the lisbon case study. *Journal of Industrial Ecology*, 13(3):384–405.

[San-Jose et al., 2007] San-Jose, J., Losada, R., Cuadrado, J., and Garrucho, I. (2007). Approach to the quantification of the sustainable value in industrial buildings. *Building and Environment*, 42(11):3916–3923.

[Schmidt et al., 2004] Schmidt, A. C., Jensen, A. A., Clausen, A. U., Kamstrup, O., and Postlethwaite, D. (2004). A comparative life cycle assessment of building insulation products made of stones wool, paper wool and flax. *International Journal of Life Cycle Assessment*, (9):122–129.

[Traverso et al., 2010] Traverso, M., Rizzo, G., and Finkbeiner, M. (2010). Environmental performance of building materials: life cycle assessment of a typical sicilian marble. *The International Journal of Life Cycle Assessment*, 15(1):104–114.

[Vatalis et al., 2011] Vatalis, K. I., Manoliadis, O. G., and Charalampides, G. (2011). Assessment of the economic benefits from sustainable construction in greece. *International Journal of Sustainable Development & World Ecology*, 18(5):377–383.

Optimization of a Humanoid Robot gait: multilocal optimization approach

José Lima*, Ana I. Pereira†, Paulo Costa‡, José Gonçalves*

* *INESC TEC (formerly INESC Porto) and Polytechnic Institute of Bragança, Portugal*
E-mail:`{jllima,goncalves}@ipb.pt`

† *Polytechnic Institute of Bragança, Portugal & Algoritmi, University of Minho, Portugal.*
E-mail:`apereira@ipb.pt`

‡ *INESC TEC (formerly INESC Porto) and Faculty of Engineering, University of Porto, Portugal*
E-mail:`paco@fe.up.pt`

Abstract

The humanoid robot gait planning presents a large number of unknown parameters that should be found to make the humanoid robot to walk. There are several approaches to achieve the gait but an accurate simulation can be used to compute it. A stable joint model of a humanoid robot is used in simulation to optimize the gait parameters. The optimization is based on the stretched simulated annealing with the multilocal algorithm approach. Final results prove the benefits of the presented optimization algorithm.

Keywords: Humanoid robot. Optimization. Simulation.

1 Introduction

Recently, humanoid robots have been enjoying great popularity and are now used as a research tool in many groups worldwide. They are a challenging task due to its hard control. Moreover, gait generation and optimization still remain a challenge for such a high-order highly-coupled nonlinear dynamical system [Hu et al., 2006]. The humanoid robot gait planning presents a large number of unknown parameters that should be found to make the humanoid robot to walk. It can be approached in two ways: the online (done in real-time that requires high computational effort) and the offline gait generation methods. This offline approach, brings some advantages such as the ability to use complex algorithms to find an optimal solution. This is the main topic of this paper.

The optimization of the humanoid robot gait is a good area to apply optimization methods over the simulated robot. It is prudent to avoid time-consuming optimization runs that wear out the robot hardware. Several approaches have been presented that aim at optimizing properties such as speed [Faber and Behnke, 2007, Hemker et al, 2007, Niehaus et al., 2007] or torso stability [Chalodhorn et al., 2007, Huang et al., 2008, Zhe et al, 2005] of a humanoid's walk.

In order to generate walking patterns for different locomotion kinematics, the common way of most existing approaches is to precompute reference trajectories [Sakka and Yokoi, 2005]. Similarly and based on previous authors works, using precomputed reference trajectory, this paper presents an overview to optimize the gait planning methods for humanoid robots [Lima et al., 2010]. For that purpose, the Stretched Simulated Annealing method was used for the optimization technique. It is a generic probabilistic method for the global optimization problem. It tries to find the global minimum of a given function in a large search space (as it happens in humanoid gait planning) and can escape from local minima. Besides, there must be used an optimizing method because it is unthinkable to perform a complete search once the search space is huge (i.e. dimension 100).

Before the optimization on the real robot, several iterations were evaluated using a simulated model of the humanoid robot. The simulations were conducted in *SimTwo* [SimTwo, 2009], a physical robot simulator that is capable to simulate user-defined robots in three-dimensional space since it includes a physical model based on rigid body dynamics (the *ODE* - *Open Dynamics Engine* [Smith, 2000]).

The paper is organized as follows: Section 2, shortly point out the developed Simulator (*SimTwo*) and the humanoid robot modeling with its parameters. Then, in Section 3, it is addressed the optimization technique based on the stretched simulated annealing well known algorithm where the gait optimization is applied. Section 4 presents a discussion of results and finally, Section 5 concludes this work and gives a future work on this topic.

2 Simulation Environment

Studying the robot's behaviour without real hardware is possible due to a physics based simulator implementation. The physics engine is the key to making simulations useful in terms of high performance robot control [Browning and Tryzelaar, 2003]. The dynamic behaviour of the robot (or multiple robots) is computed by the ODE (Open Dynamics Engine [Smith, 2000]), which is a free library for simulating rigid body dynamics.

There are several simulators with humanoid simulation capability. SimTwo, as a developed simulator, is a generic simulator that allows the access to the low level behaviour, such as dynamical model, friction model, servomotor model and sensors model in a way that can be mapped to the real robot, with a minimal overhead. The developed simulator, based on ODE, allows to build several robots. SimTwo was developed having in mind the full access to all control levels and the possibility of adding several sensors and its modeling. Besides, it owns a new precise and stable joint model that allows to simulate a robot with a high number of joints in a chain architecture without instability and noise (previously presented in [Lima et al., 2013]). SimTwo also allows to run simulations faster than real time: a very useful property to accelerate the research task. A snapshot of the developed simulator is presented in Figure 1.

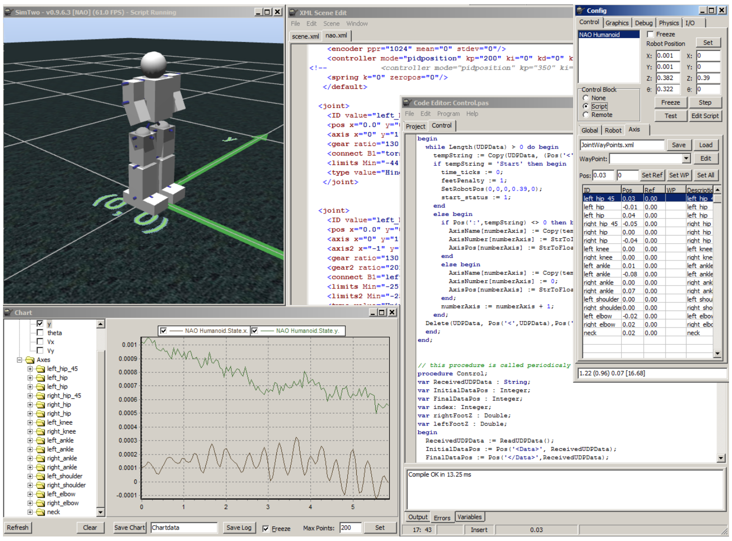


Figure 1: SimTwo simulator environment.

The dynamics realism in SimTwo is obtained by decomposing a robot in rigid bodies and electric motors (joints). Each body behaviour is numerically simulated using its physical characteristics such as shape, mass and moments of inertia, surface friction and elasticity. It is also possible to define standard joints such as socket, hinge and slider which can be coupled with an actuator or a sensor. Previous works validates the realism of the SimTwo simulator [Costa et al., 2011].

2.1 Servomotor Model

The servomotor can be based on a DC motor model with its main characteristics with an embedded state space feedback closed-loop controller. The block diagram of the servomotor model can be seen in Figure 2 where θ^i_{Rev} is the reference of the angle, ω^i_{Ref} is the reference of angular speed, Ua' is the Supply voltage of DC motor (with nonlinearities) and T_S is the available torque that will be applied to the gearbox (Electrical torque). The motor model parameter estimation has already been presented in previous work [Lima et al., 2005], as well as the friction constants that were computed based on an optimization method with validation of real hardware. The low level controller models the closed loop controller, in the real robot, which is implemented by the servomotor manufacturer (AX12 from Dynamixel).

2.2 Gearbox Model

A standard model would present some instability because an inertia moment of the motor shaft should appear multiplied by the square of gearbox ratio (in a chain of joints as it happens in a humanoid robot). The implemented model allows to increase the stability of the simulation with a spring damper

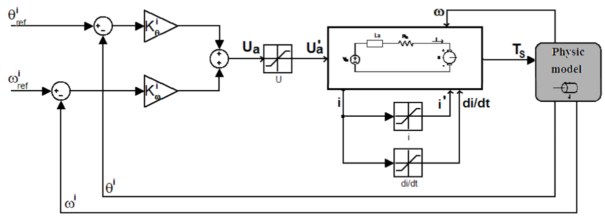


Figure 2: Motor model.

transmission approach. Moreover, the backlash of the gearbox can be approximated by this model. The improved model is implemented over the *physic model* presented in Figure 2, thus keeping the servo motor model and the controller unchanged.

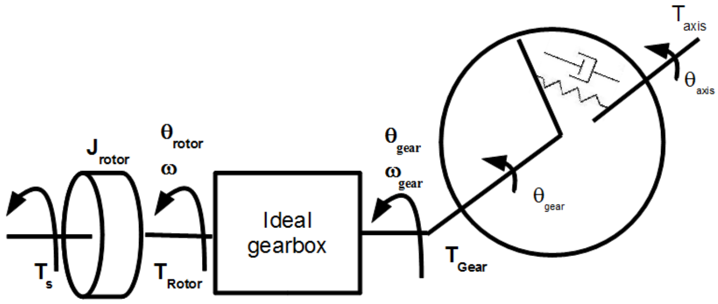


Figure 3: Gearbox model.

The motor inertia plus a spring damper gearbox model, $MIPSD$ is detailed in [Lima et al., 2013] and presented in Figure 3 where the T_{axis} is the available torque at the gearbox output and T_{Rotor} is the available torque seen by the rotor.

2.3 Gait parameters

The gait-planning is one of the fundamental problems in humanoid mobile robots. The problem of gait planning for humanoid robots is fundamentally different from the path planning for wheeled robots due to the inherent characteristics of legged locomotion. The main challenge of gait planning is to find constraint functions and their associated gait parameters. However, finding repeatable gait when the constraint equations involve higher order differential equations still remained unsolved [Zhou et al., 2004]. There are the online and the offline generation methods [Zhou et al., 2004]. The first one, should be done in real time and requires a high computational effort. On the other hand, a popular way to solve this problem is to resort to offline optimization techniques. In this paper it is used the model parameters and actuator inputs that lead to fully open-loop stable walking motions (Figure 4).

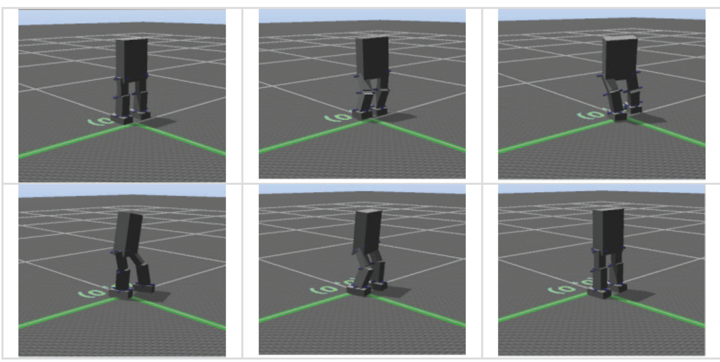


Figure 4: SimTwo Humanoid Simplified model walking

Having in mind the most basic locomotion of a humanoid robot (a simple step) it is desired to obtain the gait planning that allow joints to move the robot. It should be optimized according to the criteria:

maximizing the step distance. For the optimization technique, it was used the stretched simulated annealing. The search space in the present problem is composed by 10 degrees of freedom (each joint axis) and for 10 time instants: the optimization should be applied to a dimension (D) of 100. The optimization algorithm is implemented in Matlab that connects the simulator through network UDP packets and xml file. The simulator computes the walk distance and sends it to the Matlab so that to analyze the objective function, $f(x)$. Matlab optimization algorithm generates a new solution to be tested in simulator and shares it by a xml file and a start event of simulation by network, as presented in Figure 5.

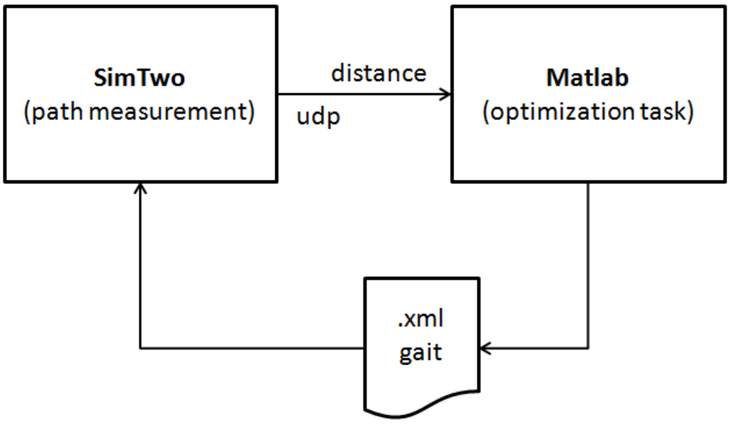


Figure 5: SimTwo and Matlab interface.

3 Stretched simulated annealing

The stretched simulated annealing (SSA) method belong to the class of the multilocal optimization methods. The SSA method solves, in each iteration, a global optimization problem using the simulated annealing (SA) algorithm. The SA is a point-to-point stochastic algorithm that does not require derivative information and is able to guarantee convergence to a global solution with probability one. In each iteration of the SSA method, the global problem is transformed using a function stretching technique [Pereira and Fernandes, 2009].

The function stretching technique aims to prevent the convergence of the SA algorithm to a previously computed global solution. Let x^j be that particular solution. Thus, the function stretching technique is applied only locally, in order to transform the objective function $f(x)$ in a neighbourhood of x^j, say $V_\varepsilon(x^j)$, $\varepsilon > 0$. So, $f(x)$ is reduced only on the region $V_\varepsilon(x^j)$ leaving all the other maxima unchanged. The mathematical formulation of the stretching function is given as follows.

$$\max_{l \le x \le u} f_j(x) \equiv \begin{cases} \bar{f}(x) - \dfrac{\delta_2[\text{sign}(f(x^j) - f(x)) + 1]}{2\tanh(\kappa(\bar{f}(x^j) - \bar{f}(x))} & \text{if } x \in V_{\varepsilon^l}(x^j),\, l \in \{1, \ldots, M\} \\ f(x) & \text{otherwise} \end{cases} \tag{1}$$

where $\bar{f}(x)$ is defined as

$$\bar{f}(x) = f(x) - \frac{\delta_1}{2}\|x - x^j\|[\text{sign}(f(x^j) - f(x)) + 1] \tag{2}$$

with δ_1, δ_2 and κ positive constants and M is the number of solutions founded. The value of these parameters depend on the problem to solve. Transformations on the original objective function stretch the neighbourhood of x^j, with ray ε^j, downwards assigning smaller function values to those points to prevent the convergence of the global optimization method to that previously computed solution [Parsopoulos and Vrahatis, 2004, Pereira and Fernandes, 2009].

Each global optimization problem is solved using SA algorithm, see [Pereira and Fernandes, 2009] for details. The multilocal procedure terminates when for a predefined set of consecutive iterations no other solution is encountered [Pereira and Fernandes, 2008, Pereira and Fernandes, 2009].

The main steps of the SSA algorithm can be defined as follows.

Algorithm 1: (SSA algorithm)

Given: x^0 and the initial control parameter values for SA algorithm. Set $j = 1$.

While the stopping condition is not verified **do**

1. Find the solution of problem, x^j, using SA method.
2. Apply the stretching function in the point x^j.
3. Set $j = j + 1$

End while

End Algorithm

In the optimization of the humanoid robot gait problem the objective function $f : \mathbb{R}^{n\times n} \longrightarrow \mathbb{R}$ gives the distance, in meters, of the humanoid robot. The variable $x \in \mathbb{R}^{n\times n}$ represents the gait for the humanoid robot to walk.

4 Numerical Results

The numerical results were obtained using a Inter Core i7-2600 CPU 3.4 GHz with 8.0 GB of RAM.

For the optimization procedure it was considered $n = 10$ and the initial approximation is defined as

$$x_0 = \begin{bmatrix} -26 & 15 & 0 & -25 & 15 & 0 & 10 & 0 & 10 & 0 \\ -26 & 15 & -10 & -55 & 25 & -10 & 10 & 10 & 35 & 10 \\ -26 & 15 & -10 & -30 & -10 & -10 & 10 & 10 & 45 & 10 \\ -26 & 25 & 10 & -30 & 0 & 10 & 10 & -10 & 45 & -10 \\ -40 & 40 & 10 & -15 & 5 & 10 & 15 & -10 & 25 & -10 \\ -55 & 25 & 10 & -15 & 5 & 10 & 45 & -10 & 25 & -10 \\ -30 & -10 & 10 & -15 & 5 & 10 & 45 & -10 & 15 & -10 \\ -30 & -10 & 10 & -40 & 30 & 10 & 45 & -10 & 15 & -10 \\ -30 & -5 & -10 & -40 & 30 & -10 & 40 & 10 & 15 & 10 \\ -30 & 0 & -10 & -45 & 33 & -10 & 35 & 10 & 20 & 10 \\ -30 & 0 & -10 & -55 & 10 & -10 & 35 & 10 & 50 & 10 \end{bmatrix} \quad (3)$$

that was obtained through empirical analysis. The matrix (3) presents the initial gait for 10 joints (lines) and for 10 time instants (columns).

The upper and lower limits are defined using x_0, and are defined as $l = x_0 - \alpha 1_{10}$ and $u = x_0 + \alpha 1_{10}$, where $1_{10} \in \mathbb{R}^{10\times 10}$ is a matrix with coefficient one in all positions. The walking distance of the humanoid robot when considered the input variable x_0 is $f(x_0) = 0.089202$ m.

The parameters δ_1, δ_2, κ and α were fixed as 10^2, 1, 10^{-3} and 2, respectively.

The SSA method was capable to identify only one solution for the proposed problem. The following figure presents optimization results. The solution was obtained after 2942 iterations, where it was identified the optimum value as 0.174153. This solution was obtained after 3030 seconds.

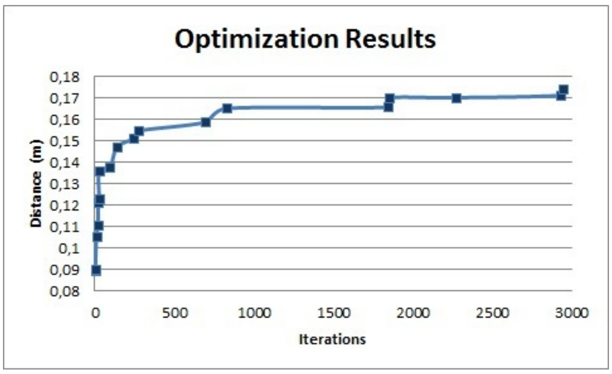


Figure 6: Optimization results

It is possible to observe that was possible to have 95% of improvement when it compared with the initial value.

5 Conclusions and future work

In this paper an optimization method based on the stretched simulated annealing method for the gait planning of humanoid robots was presented. For that purpose, tests were conducted on the personalized simulator SimTwo, a robot simulator that is capable to simulate user-defined robots in three-dimensional space with physical model based on rigid body dynamics (ODE). The new model of joints (motor inertia plus a spring damper gearbox model, $MIPSD$) was used since it allows to obtain more stable results. Using approximately 3000 iterations it was possible to find the optimum solution of the problem (maximize the walking distance), the stretched simulated annealing method was capable of increase 95%. These results validate the optimizations method even in a noisily system. As future work, the implementation of different optimization techniques could be tested and evaluated and the humanoid get up movements could be determined and optimized through this approach.

Acknowledgment

This work is financed by the ERDF - European Regional Development Fund through the COMPETE Programme (operational programme for competitiveness) and by National Funds through the FCT - Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) within projects FCOMP - 01 - 0124-FEDER-022701 and FCOMP-01-0124-FEDER-022674.

References

[Browning and Tryzelaar, 2003] Browning, B., Tryzelaar, E. (2003) Ubersim: A realistic simulation engine for robotsoccer. *In Proceedings of Autonomous Agents and Multi-Agent Systems*, Australia.

[Chalodhorn et al., 2007] Chalodhorn, R., Grimes, D., Grochow, K. and Rao, R. (2007) Learning to walk through imitation. *Proc. of the Int. Conf. on Artificial Intelligence (IJCAI)*, India.

[Costa et al., 2011] Costa, P., Gonçalves, J., Lima, J. Malheiros, P. (2011) SimTwo Realistic Simulator: A tool for the development and validation of robot software. *Theory and Applications of Mathematics and Computer Science*, 1 (1): 17-33.

[Faber and Behnke, 2007] Faber, F. and Behnke, S. (2007) Stochastic optimization of bipedal walking using gyro feedback and phase resetting. *Proc. of the IEEE-RAS Int. Conf. on Humanoid Robots (Humanoids)*, Estados Unidos.

[Hemker et al, 2007] Hemker, T., Stelzer, M., Stryk, O. and Sakamoto, H. (2009) Efficient walking speed optimization of a humanoid robot. Int. Journal of Robotics Research, 28 (2): 303-314.

[Hu et al., 2006] Hu, L., Zhou, C., Sun, Z. (2006) Biped Gait Optimization Using Spline Function Based Probability Model. *Proceedings of the 2006 IEEE International Conference on Robotics and Automation*, Estados Unidos.

[Huang et al., 2008] Huang, Q., Yu, Z., Zhang, W., Duan, X., Huang, Y. and Li, K. (2008) Generation of humanoid walking pattern based on human walking measurement. *Proc. of the IEEE-RAS Int. Conf. on Humanoid Robots (Humanoids)*, Coreia do Sul.

[Lima et al., 2013] Lima, J., Gonçalves, J., Costa P., Moreira, A. (2013) Humanoid Gait Optimization Resorting to an Improved Simulation Model. *International Journal of Advanced Robotic Systems*,10 (67): 1-7.

[Lima et al., 2010] Lima, J., Gonçalves, J., Costa, P., Moreira, A. (2010) Humanoid Robot Gait Planning Resorting to an Adaptive Simulated Annealing Algorithm. *Proceedings of 10th Conference on Autonomous Robot Systems and Competitions*, Portugal.

[Lima et al., 2005] Lima, J., Gonçalves, J., Costa, P., Moreira, A. (2005) Humanoid Realistic Simulator: The servomotor joint modeling. *Proceedings of 6th International Conference on Informatics in Control, Automation and Robotics*, Italia.

[Niehaus et al., 2007] Niehaus, C. and Röfer, T. and Laue, T. (2007) Gait optimization on a humanoid robot using particle swarm optimization. *Proceedings of the Second Workshop on Humanoid Soccer Robots*, Estados Unidos.

[Parsopoulos and Vrahatis, 2004] Parsopoulos, K., Vrahatis, M. (2004) On the computation of all global minimizers through particle swarm optimization,*IEEE Transaction on Evolutiona y Computation*, 8(3): 211–224.

[Pereira and Fernandes, 2009] Pereira, A., Fernandes, E.M.G.P. (2009) A reduction method for semi-infinite programming by means of a global stochastic approach. *Optimization*, 58: 713–726.

[Pereira and Fernandes, 2008] Pereira, A., Fernandes, E.M.G.P. (2008) On a reduction line search filter method for nonlinear semi-infinite programming problems. *Euro Mini Conference "Continuous Optimization and Knowledge-Based Technologies"*, L. Sakalauskas, G.W. Weber, E.K. Zavadskas (eds.), 174–179.

[Sakka and Yokoi, 2005] Sakka, S., Yokoi, K. (2005) Humanoid Vertical Jumping based on Force Feedback and Inertial Forces Optimization. *Proceedings of the 2005 IEEE International Conference on Robotics and Automation*, Spain.

[SimTwo, 2009] "*SimTwo*" (2009) http://www.fe.up.pt/∼paco/wiki.

[Smith, 2000] Smith, R. (2000) Open Dynamics Engine, http://www.ode.org.

[Zhe et al, 2005] Zhe, T., Changjiu, Z. and Zengqititle, S. (2005) Humanoid walking gait optimization using GA-based neural network. *Advances in Natural Computation*, 3611: 252–261 .

[Zhou et al., 2004] Zhou, C., Yue, P. K., Ni, J., Chan, S. (2004) Dynamically Stable Gait Planning for a Humanoid Robot to Climb Sloping Surface. *Proceedings of the 2004 IEEE Conference on Robotics, Automation and Mechatronics*, Singapore.

Um método híbrido de pontos interiores e *branch-and-bound* aplicado ao modelo multiobjetivo de custo de colheita, coleta e aproveitamento de resíduos da cana-de-açúcar

Camila de Lima[1], Antonio Roberto Balbo[2], Helenice de Oliveira Florentino Silva[3], Thiago Pedro Donadon Homem[4]

[1] *Programa de Pós-graduação em Engenharia Elétrica, Faculdade de Engenharia de Bauru, UNESP, Bauru-SP, Brasil*
*E-mail:*cadlima@yahoo.com.br
[2] *Departamento de Matemática, Faculdade de Ciências, UNESP, Bauru-SP, Brasil*
*E-mail:*arbalbo@fc.unesp.br
[3] *Departamento de Bioestatística, Instituto de Biociências, UNESP, Botucatu-SP, Brasil*
*E-mail:*helenice@ibb.unesp.br
[4] *Instituto Federal de Educação, Ciência e Tecnologia de São Paulo, Boituva-SP, Brasil*
*E-mail:*thiagopdh@feb.unesp.br

Resumo

Neste trabalho desenvolveu-se um método híbrido envolvendo os métodos previsor-corretor primal-dual de pontos interiores e *branch-and-bound*, o qual foi explorado à resolução do modelo multiobjetivo relativo à colheita, à coleta e ao aproveitamento de resíduos da cana-de-açúcar. Este modelo visa determinar a escolha das variedades de cana-de-açúcar que minimizam o custo de colheita e coleta de seus resíduos e/ou que maximizam o balanço de energia resultante do aproveitamento destes resíduos, considerando áreas mecanizáveis e semi-mecanizáveis, e respeitando as restrições do problema. O modelo multiobjetivo é transformado em uma classe de problemas mono-objetivo através das estratégias da soma ponderada e do ε-restrito. O método previsor-corretor primal-dual de pontos interiores é utilizado, explorando o parâmetro de barreira em ambos os procedimentos, para se obter a solução ótima relaxada de cada subproblema. A partir desta, utiliza-se o método *branch-and-bound* para integralizar variáveis fracionárias, e obter a solução ótima inteira 0-1 para cada ponderação imposta. Os testes foram realizados através de uma implementação no software Borland C++ Builder 6.0 e os resultados foram comparados àqueles do aplicativo Solver do software Excel, demonstrando que o procedimento tem um bom desempenho computacional, determina as soluções eficientes dos subproblemas e a curva de Pareto ótimo para o modelo multiobjetivo, em um tempo computacional pequeno para um problema realístico com 10 variedades e 16 talhões.

Palavras chave: Método Previsor-Corretor Primal Dual de Pontos Interiores, Método *Branch-and-Bound*, Modelo Multiobjetivo, Cana-de-Açúcar, Biomassa Residual.

1 Introdução

De acordo com [Pellegrini, 2002], a maior parte da energia do Brasil é gerada por usinas hidrelétricas, porém a geração proveniente de fontes alternativas têm se desenvolvido. Dentre elas destacam-se a energia nuclear, a solar, a eólica e a energia cogerada pela biomassa residual. Segundo [Lima, 2009], a cogeração consiste em um processo de produção simultânea e sequenciada de duas ou mais formas de energia a partir de um único combustível, que podem ser convertidas para consumo próprio ou venda. Além disso, destaca-se pela baixa produção de micro poluentes. [Ripoli e Ripoli, 2004] afirmam em seus estudos que o bagaço e o palhiço da cana-de-açúcar são as biomassas que possuem maior poder calorífico.

Conforme os dados do [IBGE, 2012], o Brasil é o maior produtor de cana-de-açúcar do mundo, e as safras aumentam a cada ano. Alguns problemas são resultados deste crescimento acelerado, como o aumento da queima da cana-de-açúcar, utilizadas no processo de colheita semi-mecanizada, o que gera poluição e impactos ao meio ambiente, e ainda, a grande quantidade de resíduos no solo, como folhas, palhas, ponteiros e frações de colmo, ocasionada pela colheita mecanizada, que favorecem o aparecimento de pragas, a contaminação do solo, e o comprometimento da próxima safra [Tolentino, 2007]. A fim de resolver este problema, tem-se investido no reaproveitamento desta biomassa residual através da cogeração de energia, para utilização desta no setor sucroalcooleiro ou comercialização no mercado de energia.

Assim, muitos estudos têm sido propostos visando otimizar o custo de coletar e transferir a cana-de-açúcar e a biomassa residual, do campo para o centro de processamento, bem como ao uso deste resíduo para a geração de energia, considerando-se os sistemas de colheita mecanizável e semi-mecanizável. Nos trabalhos de [Florentino, 2006], [Tolentino, 2007] e [Homem, 2010], são discutidos modelos matemáticos para a escolha de variedades de cana-de-açúcar que buscam otimizar o custo de coleta da biomassa residual e/ou a geração de energia. [Ramos, 2010] e [Silva, 2011] apresentam modelos de minimização de custo da colheita da cana-de-açúcar, considerando áreas mecanizáveis e semi-mecanizáveis em sua formulação.

De modo geral, o problema investigado visa minimizar o custo da colheita da cana-de-açúcar, bem como minimizar o custo de coleta e transporte da biomassa residual e/ou maximizar de geração de energia da biomassa, determinando qual tipo de variedade será plantada em cada talhão, respeitando as restrições de produção de sacarose e fibra de cana-de-açúcar, uso total da área destinada ao plantio (mecanizável ou semi-mecanizável) e o plantio de apenas uma variedade de cana-de-açúcar por talhão. Assim, cada variedade de cana-de-açúcar pode ou não ser plantada em determinado talhão. Este aspecto caracteriza um problema de programação inteira zero-um, e dependendo de sua dimensão, pode ser de difícil resolução. Além disso, inseriu-se uma restrição com o intuito de limitar o plantio de uma mesma variedade no conjunto de talhões.

Desta forma, desenvolveu-se um procedimento híbrido envolvendo métodos previsor-corretor primal-dual de pontos interiores e *branch-and-bound* para a resolução de um modelo multiobjetivo relativo à cana-de-açúcar referente a uma usina do estado de São Paulo, que envolve 10 variedades de cana-de-açúcar e 16 talhões para plantio. O procedimento híbrido usado é uma excelente ferramenta computacional para a resolução destes problemas, determinando soluções eficientes para o modelo em um tempo computacional pequeno. Este modelo é tratado através dos métodos da soma ponderada e ε-restrito, que são estratégias de resolução que transformam o modelo multiobjetivo em uma classe de problemas mono-objetivo. Assim, será utilizado o método previsor-corretor primal-dual de pontos interiores para se obter a solução ótima relaxada do modelo e, a partir desta, determinar a solução ótima inteira relacionada às restrições de integralidade do problema, relativas à escolha das variedades a serem plantadas, através do método *branch-and-bound*.

2 Modelo multiobjetivo

2.1 Formulação da função objetivo

De acordo com [Ramos, 2010] e [Silva, 2011], na colheita de cana-de-açúcar queimada têm-se os custos de aceiro, queima, corte manual, carregamento da cana para o caminhão e transporte da cana do campo para a usina. Na colheita mecanizada, têm-se os custos de corte e transporte da cana do campo para a usina. Estes dados estão apresentados nas tabelas 1, 3, 4 e 5 da seção 4. O custo de transporte da variedade i plantada no talhão j (Ct_{ij}) a uma distância (Dj) do talhão j para a usina:

$$Ct_{ij} = c_{med_i}.D_j \tag{2.1}$$

Em que: $i = 1, 2, ..., n$ são os índices que representam as variedades; $j = 1, 2, ..., k$ são os índices que representam os talhões; c_{med_i} é o custo médio do transporte da cana por km; e D_j é a distância do talhão j do centro de processamento, em talhões.

O custo C_{ij}^{SM} de colheita e transporte da cana-de-açúcar de variedade i plantada no talhão j no sistema semi-mecanizado é calculado da seguinte forma:

$$C_{ij}^{SM} = (Ca_i + Cq_i + Cco_i + Cca_i + Ct_{ij}).L_j \tag{2.2}$$

Em que: Ca_i é o custo de aceiro da variedade i (R\$.ha^{-1}); Cq_i é o custo da queima da variedade i (R\$.ha^{-1}); Cco_i é o custo de corte da variedade i (R\$.ha^{-1}); Cca_i é o custo de carregamento da variedade i (R\$.ha^{-1}); Ct_{ij} é o custo de transporte da variedade i plantada no talhão j (R\$.ha^{-1}), calculado em (2.1); e L_j é área do talhão j, em hectare.

No sistema mecanizado o custo, C_{ij}^{M}, de colheita e transporte da cana de variedade i plantada no talhão j, é calculado da seguinte forma:

$$C_{ij}^{M} = (Cco_i + Ct_{ij}).L_j \tag{2.3}$$

Em que: Cco_i é o custo de corte da variedade i (R\$.ha^{-1}); Ct_{ij} é o custo de transporte da variedade i plantada no talhão j (R\$.ha^{-1}), calculado em (2.1); e L_j é área do talhão j, em hectare.

Além disso, no sistema mecanizado, o corte é realizado por máquinas, sem a queima da cana-de-açúcar, provocando a permanência de resíduos no solo. Segundo [Florentino, 2006] e [Tolentino, 2007] o recolhimento dos resíduos resultantes deste tipo de colheita é feito da seguinte forma: primeiramente, o palhiço é enleirado, em seguida é passado em uma máquina para compactação, depois é carregado no caminhão e finalmente transportado para o centro de processamento.

Assim, o custo de coleta do palhiço da cana-de-açúcar da variedade i plantada no talhão j (CC_{ij}) é calculado pela soma do custo envolvido no processo de enleirar, compactar e carregar o caminhão com o palhiço da variedade i (C_i) e o custo de transporte do palhiço até a usina (CT_{ij}), multiplicada à área do talhão j (L_j), calculados através das tabelas 1, 2 e 5, apresentadas na seção 4. Abaixo, segue a expressão do custo referente ao processo de coleta dos resíduos:

$$CC_{ij} = (C_i + CT_{ij})\, L_j \tag{2.4}$$

Desta forma, a função objetivo que representa o custo total do processo de colheita da cana-de-açúcar e de coleta de resíduos é expresso por:

$$CT = \sum_{i=1}^{n}\sum_{j=1}^{l}\left(C_{ij}^{M} + CC_{ij}\right) X_{ij} + \sum_{i=1}^{n}\sum_{j=l+1}^{k} C_{ij}^{SM} X_{ij} \tag{2.5}$$

em que: CT é o custo total gasto pela usina na colheita da cana-de-açúcar e na coleta de seus resíduos, considerando as áreas mecanizáveis e semi-mecanizáveis; C_{ij}^{M} é o custo da colheita mecanizada, e está definido em (2.3); CC_{ij} é o custo da coleta de resíduos, e está definido em (2.4), lembrando que somente as áreas mecanizáveis são consideradas devido ao processo de colheita da cana-de-açúcar com a utilização de máquinas; C_{ij}^{SM} é o custo da colheita semi-mecanizada, e está definido em (2.2); l é número de talhões em que se considera o sistema mecanizado; $k - l$ é o número de talhões em que se considera o sistema semi-mecanizado; e X_{ij} são as variáveis de decisão.

Em se tratando de energia, de acordo com [Florentino, 2006] e [Florentino et al., 2011], o balanço de energia para o aproveitamento do palhiço é obtido pela diferença entre a energia proveniente do palhiço da variedade i plantada no talhão j (EB_{ij}) e a energia gasta na transferência palhiço da variedade i plantada no talhão j (ET_{Bij}) que é a soma das energias gastas para enleirar e compactar (EEC_{ij}), carregar (EC_{ij}) e transportar esta biomassa (ET_{ij}), calculadas através dos dados das tabelas 1, 2 e 5 apresentados na seção 4. A fórmula do balanço de energia é apresentada a seguir:

$$BE_{ij} = EB_{ij} - ET_{Bij} \tag{2.6}$$

Assim, define-se a função objetivo referente ao balanço de energia no aproveitamento de resíduos de cana-de-açúcar:

$$BET = \sum_{i=1}^{n}\sum_{j=1}^{l} BE_{ij}X_{ij} \tag{2.7}$$

em que: BET é o balanço de energia total no aproveitamento de resíduos da cana-de-açúcar, e BE_{ij}^{M} é o balanço de energia para o aproveitamento do palhiço, e está definido em (2.6).

Neste trabalho, os cálculos de custo de coleta e do balanço de energia referentes aproveitamento de resíduos das áreas semi-mecanizáveis são considerados nulos, visto que a colheita nesta área não gera resíduos. Pretende-se inserir no modelo, dados para a cogeração de energia a partir do bagaço da cana-de-açúcar, e estes serão considerados tanto em áreas mecanizáveis quanto em áreas semi-mecanizáveis.

2.2 Modelo multiobjetivo

O modelo em questão consiste em determinar quais das n variedades i de cana-de-açúcar devem ser plantadas nos k talhões j de área L_j (ha) e distância D_j (Km) da usina, que investigue, simultaneamente, o mínimo custo total de colheita da cana-de-açúcar e coleta de resíduos e o máximo balanço de energia no aproveitamento de resíduos resultantes da colheita em áreas mecanizáveis, os quais são objetivos conflitantes, levando em consideração restrições como quantidade de produção de sacarose e fibra de cana-de-açúcar, uso total da área destinada ao plantio e o plantio de apenas uma variedade de cana-de-açúcar por talhão. O modelo matemático é definido por:

$$\text{Minimizar } \ (CT; (-1), BET) \tag{2.8}$$

$$
\begin{gathered}
\text{Subject to: } \sum_{i=1}^{n}\sum_{j=1}^{k} A_i L_j X_{ij} \geqslant T\bar{A} \\
T\bar{F}_I \leqslant \sum_{i=1}^{n}\sum_{j=1}^{k} F_i L_j X_{ij} \leqslant T\bar{F}_S \\
\sum_{i=1}^{n} X_{ij} = 1, \text{ for all } j \\
X_{ij} = 0 \text{ or } 1, \; i = 1,2,\ldots,n \text{ and } j = 1,2,\ldots,k \\
\sum_{j=1}^{k} X_{ij} \leqslant M
\end{gathered}
\tag{2.9}
$$

Em que: CT está defido em (2.5) e BET está definido em (2.7) , $i = 1,2,\ldots,n$ são os índices que representam as variedades, $j = 1,2,\ldots,k$ são os índices que representam os talhões; A_i é a estimativa de produção de sacarose da variedade i (t/ha); $\bar{A}$ é a quantidade mínima estabelecida para a POL da cana; T é o número total de talhões; F_i é a estimativa do teor de fibra da variedade i; $\bar{F}_I$ e$\bar{F}_S$ são as quantidades mínimas e máximas estabelecidas para a fibra da cana; $X_{ij} = 1$ implica que a cana de variedade i deve ser plantada no talhão j e em caso contrário $X_{ij} = 0$; M é o número máximo que cada variedade i pode ser plantada.

2.3 Estratégias de resolução do modelo multiobjetivo

Geralmente, os modelos multiobjetivo são de difícil resolução e na maioria das vezes exigem a intervenção do usuário para a determinação de soluções satisfatórias (eficientes). Assim, para a resolução do modelo apresentado em (2.2), será utilizada a estratégia de otimização conhecida por Otimalidade de Pareto, através dos métodos da soma-ponderada e do ε-restrito, baseando-se em [Deb, 2004], que transformam o modelo em uma classe de subproblemas mono-objetivos, buscam determinar o conjunto de soluções eficientes para o modelo e possiblitam a construção da curva de Pareto ótimo.

2.3.1 Estratégia da soma ponderada

Esta estratégia consiste na resolução do problema através de um balanceamento entre as funções objetivo, através de um parâmetro $\alpha \in [0,1]$, que é apresentado em (2.10). Através desta, é possivel selecionar as soluções que retornam os melhores valores para a função objetivo balanceada para um dado α, as quais são chamadas de soluções eficientes. A seguir, essa estratégia de resolução é apresentada:

$$\text{Minimizar } (\alpha CT - (1-\alpha)BET) \tag{2.10}$$

$$\text{Sujeito a: Restrições (2.9)}$$

em que $0 \leqslant \alpha \leqslant 1$.

2.3.2 Estratégia do ε-restrito

Esta estratégia consiste em manter um dos objetivos como função, restringindo-se os demais objetivos, com valores delimitados pelo usuário, neste caso, de interesse das usinas. Essa estratégia investiga soluções do modelo multiobjetivo através do seguinte problema mono-objetivo:

$$\text{Minimizar } CT \tag{2.11}$$

$$\text{Sujeito a: Restrições (2.9)}$$

$$\sum_{i=1}^{n}\sum_{j=1}^{k} BET X_{ij} \geqslant \varepsilon \tag{2.12}$$

3 Método previsor-corretor primal-dual de pontos interiores e *branch-and-bound*

Neste trabalho, utilizou-se o método primal-dual com procedimento previsor-corretor baseando-se em [Homem et al., 2011], variante daquele proposto por [Mehrotra, 1992] diferenciando-se deste por já utilizar no passo previsor informações do parâmetro de barreira μ_k, o que melhora a eficiência do método por evitar que os pontos definidos por este no passo previsor, aproximem-se da fronteira do problema, podendo, inclusive, inviabilizá-los. Enquanto que, no passo corretor, este reajusta as direções com informações dos aproximantes de segunda ordem referentes às condições de complementaridade, possibilitando que, o procedimento de centragem do passo previsor mais o ajuste feito no passo corretor, acelerem a convergência do processo, para a determinação da solução ótima do problema contínuo. Além disso, o modelo investigado enquadra-se como um problema de programação inteira binária ou programação inteira zero-um. Desta forma, o método *branch-and-bound*, associado a uma estratégia encontrada em [Borchers e Mitchell, 1992] é utilizado no procedimento híbrido proposto. Neste, inicialmente realiza-se a busca da solução ótima relaxada do problema utilizando o método previsor-corretor primal-dual de pontos interiores e a seguir, utiliza-se o método *branch-and-bound* para a geração de soluções inteiras, de modo a obter a solução ótima e eficiente dos problemas mono-objetivos, visto em (2.10) e (2.11).

Neste sentido, considera-se o seguinte problema de programação linear (PPL) primal, com restrições lineares e variáveis canalizadas:

$$\begin{array}{l}\text{Minimizar } c^Tx\\ \text{Sujeito a: } \begin{cases} Ax = b \\ l \leqslant x \leqslant u \end{cases}\end{array} \Leftrightarrow \begin{array}{l}\text{Minimizar } c^Tx\\ \text{Sujeito a: } \begin{cases} Ax = b \\ x \geqslant l \text{ e } x \leqslant u \end{cases}\end{array} \Leftrightarrow \begin{array}{l}\text{Minimizar } c^Tx\\ \text{Sujeito a: } \begin{cases} Ax = b \\ x - r = l \\ x + z = u \\ r \geqslant 0 \text{ e } z \geqslant 0 \end{cases}\end{array} \tag{3.1}$$

em que $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, $x, c, l,\ u \in \mathbb{R}^n$, A com posto m, $r \geqslant 0$ é a variável de folga e $z \geqslant 0$ é a variável de excesso.

Considerando o PPL com restrições lineares de igualdade e variáveis canalizadas (3.1), este é redefinido através de um problema de programação não-linear (PPNL) primal-dual irrestrito expresso a partir da função lagrangiana barreira logarítmica $L_\mu(x, w, z, r, y, s)$:

$$L_\mu(x, w, z, r, y, s) = c^Tx + w^T(b - Ax) + s^T(l + r - x) + y^T(x + z - u) - \mu \sum\nolimits_{i=1}^{n} \ln(z_i) - \mu \sum\nolimits_{i=1}^{n} \ln(r_i) \tag{3.2}$$

em que: $w \in \mathbb{R}^m$ e $y,\ s \in \mathbb{R}^n$; $s \geqslant 0$, $y \geqslant 0$, são as variáveis duais do problema e $\mu > 0$ é o parâmetro de barreira ou parâmetro de centragem.

A partir de (3.2) e considerando a restrição $x - r = l$ de (3.1), nota-se que quando $l = 0$, temos que $x = r$. Desta forma, temos as seguintes condições de otimalidade de Karush-Kuhn-Tucker (KKT) para este problema:

$$A^Tw + s - y = c \tag{3.3}$$
$$Ax = b \tag{3.4}$$
$$x + z = u \tag{3.5}$$
$$XSe - \mu e = 0 \tag{3.6}$$
$$ZYe - \mu e = 0 \tag{3.7}$$

em que: X, Z, S e Y são matrizes diagonais, respectivamente com r_i, z_i, s_i e y_i como elementos diagonais e $e = (1, \ldots, 1)^T$.

Se um ponto $(x^k, z^k, w^k, s^k, y^k)$ de uma iteração corrente k não satisfaz as equações de KKT apresentadas de(3.3) a (3.7), então gera o resíduo dual $g^k = c - A^Tw^k - s^k + y^k$, relacionado à equação (3.3) , os resíduos primais $t^k = b - Ax^k$ e $f^k = u - x^k - z^k$, referentes às equações (3.4) e (3.5), respectivamente, e as folgas complementares $v^k = \mu_k e - X_kS_ke$ e $q^k = \mu_k e - Z_kY_ke$, relacionadas às equações (3.6) e (3.7), respectivamente. Estes resíduos estão definidos no passo 3, do algoritmo da seção 3.1, e serão utilizados no critério de parada do método apresentado no passo 2 desse algoritmo.

No critério de parada são feitos os testes de factibilidade primal, factibilidade dual e folgas complementares, expressos respectivamente por (3.8) com o objetivo de garantir que $(x^k, z^k, w^k, s^k, y^k)$ seja a solução ótima do problema:

$$\frac{\|t^k\|}{\|b\| + 1} \leqslant \varepsilon_1;\ \frac{\|g^k\|}{\|c\| + 1} \leqslant \varepsilon_2;\ \|\tilde{v}^k\| < \varepsilon_3;\ \|\tilde{q}^k\| < \varepsilon_4 \tag{3.8}$$

em que $\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4 > 0$ são pequenas tolerâncias positivas. Além destes, utilizou-se o seguinte teste de integralidade (3.9), proposto por [Borchers e Mitchell, 1992], para um melhor desempenho do algoritmo híbrido proposto.

$$\left|\frac{x_i^k}{x_i^{k-1}} - 1\right| < 0,01; \ \left|\frac{s_i^k}{s_i^{k-1}} - 1\right| < 0,01; \ \frac{y_i^k}{y_i^{k-1}} < 0,9; \ \frac{z_i^k}{z_i^{k-1}} < 0,9; \tag{3.9}$$

A definição de um novo ponto depende diretamente das direções de movimento e comprimento de passo nesta direção, sendo definido através de:

$$(x^{k+1}, z^{k+1}, w^{k+1}, s^{k+1}, y^{k+1}) = (x^k + \alpha_k^P d_x^k, z^k + \alpha_k^P d_z^k, w^k + \alpha_k^D d_w^k, s^k + \alpha_k^D d_s^k, y^k + \alpha_k^D d_y^k) \tag{3.10}$$

em que $\alpha_k^P > 0$ é o comprimento de passo das variáveis primais, $\alpha_k^D > 0$ é o comprimento de passo das variáveis duais, e d_x^k, d_z^k, d_w^k, d_s^k e d_y^k são as direções de busca.

As direções do passo previsor são determinadas utilizando o método de Newton, considerando uma aproximação linear de primeira ordem por série de Taylor das condições de KKT, apresentadas de (3.3) a (3.7), sobre o novo ponto definido em (3.10), sem considerar ainda o comprimento do passo. Assim, tem-se as seguintes direções do passo previsor, que serão apresentadas no passo 4 do algoritmo da seção 3.1:

$$\begin{gathered} d_x^k = \theta_k \left(A^T d_w^k + p^k - u^k\right); d_z^k = -d_x^k + f^k; d_w^k = \left(A\theta_k A^T\right)^{-1} \left[A\theta_k \left(-p^k + u^k\right) + t^k\right]; \\ d_s^k = X_k^{-1} \left(v^k - S_k d_x^k\right); d_y^k = Z_k^{-1} \left(q^k - Y_k d_z^k\right); \end{gathered} \tag{3.11}$$

em que $\theta_k = \left(X_k^{-1} S_k + Z_k^{-1} Y_k\right)^{-1}$ e $p^k = Z_k^{-1}\left(Y_k f^k - q^k\right) + X_k^{-1} v^k$.

As direções do passo corretor, por sua vez, consideram aproximações de segunda ordem sobre os resíduos relacionados às condições de complementaridade, q^k e v^k, os quais são redefinidos utilizando as direções de busca $\left(d_x^k, d_z^k, d_w^k, d_s^k, d_y^k\right)$ determinadas no passo previsor do método, para obter os novos resíduos do passo corretor $\tilde{q}^k$e $\tilde{v}^k$, vistos no passo 5 do algoritmo 3.1, e calculados da seguinte forma:

$$\tilde{v}^k = \mu_k e - X_k S_k e - D_x^k D_s^k; \tilde{q}^k = \mu_k e - Z_k Y_k e - D_z^k D_y^k e \tag{3.12}$$

em que: $D_x^k = Diag(d_{x_i}^k)$, $D_s^k = Diag(d_{s_i}^k)$, $D_z^k = Diag(d_{z_i}^k)$, e $D_y^k = Diag(d_{y_i}^k)$.

Desta forma, obtemos através do método de newton, como no passo previsor, as novas direções do passo corretor $(\tilde{d}_x^k, \tilde{d}_z^k, \tilde{d}_w^k, \tilde{d}_s^k, \tilde{d}_y^k)$, que podem ser vistas no passo 6 do algoritmo 3.1:

$$\begin{gathered} \tilde{d}_x^k = \theta_k \left(A^T \tilde{d}_w^k + \tilde{p}^k - u^k\right); \tilde{d}_w^k = \left(A\theta_k A^T\right)^{-1} \left[A\theta_k \left(-\tilde{p}^k + u^k\right) + t^k\right]; \tilde{d}_z^k = -\tilde{d}_x^k + f; \\ \tilde{d}_s^k = X_k^{-1} \left(\tilde{v}^k - S_k \tilde{d}_x^k\right); \tilde{d}_y^k = Z_k^{-1} \left(\tilde{q}^k - Y_k \tilde{d}_z^k\right); \end{gathered} \tag{3.13}$$

em que $\tilde{p}^k = Z_k^{-1}\left(Y_k f^k - \tilde{q}^k\right) + X_k^{-1} \tilde{v}^k$.

O comprimento do passo, apresentado no passo 8 do algoritmo citado, referente às variáveis primais e duais do problema, são calculados da seguinte maneira baseando-se em [Granville, 1994]:

$$\begin{aligned} \alpha_k^P &= \min\left\{1, \min\left\{\frac{-\alpha x_i}{\tilde{d}x_i} / \tilde{d}x_i < 0\right\}, \min\left\{\frac{-\alpha z_i}{\tilde{d}z_i} / \tilde{d}z_i < 0\right\}\right\}; \\ \alpha_k^D &= \min\left\{1, \min\left\{\frac{-\alpha s_i}{\tilde{d}s_i} / \tilde{d}s_i < 0\right\}, \min\left\{\frac{-\alpha y_i}{\tilde{d}y_i} / \tilde{d}y_i < 0\right\}\right\}; \end{aligned} \tag{3.14}$$

em que $0 < \alpha < 1$.

Desta forma, definem-se os passos de 1 a 9 do algoritmo previsor-corretor primal-dual. Este algoritmo é complementado no passo 10 pelo método *branch-and-bound*, utilizado para integralizar as soluções obtidas pelo método primal-dual, baseando-se em [Bazaraa e Shetty, 1979], [Borchers e Mitchell, 1992] e [Homem et al., 2011].

3.1 Algoritmo previsor-corretor primal-dual e *branch-and-bound* (PDBB)

Passo 1: Ajustar k = 0 e encontrar uma solução inicial $(x^0, z^0, w^0, s^0, y^0) \in P \times D$, ou seja, uma solução inicial factível. Seja $\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4 > 0$ pequenas tolerâncias positivas auxiliares ao passo 2 do algoritmo.

Passo 2: Testar a otimalidade de solução: Se o critério de parada definido em (3.8) é atingido então vá para o passo 10, pois a solução relaxada x^k, z^k, w^k, s^k, y^k obtida é ótima. Caso contrário, continue.
Passo 3: Fazer os cálculos intermediários do passo previsor.
Passo 4: Calcular as direções $d_x^k, d_z^k, d_w^k, d_s^k$ e d_y^k do passo previsor, definidas em (3.11).
Passo 5: Fazer os cálculos intermediários do passo corretor, atualizando os termos de segunda ordem das folgas complementares vistos em (3.12)
Passo 6: Atualizar as direções $\tilde{d}_x^k, \tilde{d}_z^k, \tilde{d}_w^k, \tilde{d}_s^k$ e $\tilde{d}_y^k$ do passo corretor, definidas em (3.13).
Passo 7: Testar a ilimitariedade: Se $t^k = 0, f^k = 0$, $\tilde{d}_x^k, \tilde{d}_z^k > 0$, e $c^t\tilde{d}_x^k < 0$, então o problema primal é ilimitado. Se $g^k = 0$, $\tilde{d}_w^k, \tilde{d}_s^k, \tilde{d}_y^k > 0$ e $b^t\tilde{d}_w^k > 0$, então o problema dual é ilimitado. Se ambos os casos acontecem, então PARE e vá para o passo 10. Se $\tilde{d}_x^k, \tilde{d}_z^k, \tilde{d}_w^k, \tilde{d}_s^k, \tilde{d}_y^k = 0$, então também PARE, x^k, z^k, w^k, s^k, y^k são soluções ótimas dos problemas primal e dual, respectivamente. Caso contrário ir para o passo 8.
Passo 8: Calcular os comprimentos dos passos primal e dual, através de (3.8).
Passo 9: Determinar uma nova solução: $x^{k+1} = x^k + \alpha_k^P\tilde{d}_x^k$, $z^{k+1} = z^k + \alpha_k^P\tilde{d}_z^k$, $w^{k+1} = w^k + \alpha_k^D\tilde{d}_w^k$, $s^{k+1} = s^k + \alpha_k^D\tilde{d}_s^k$, e $y^{k+1} = y^k + \alpha_k^D\tilde{d}_y^k$. Atualizar $k \leftarrow k+1$ e ir para o Passo 2.
Passo 10: Método *Branch-and-Bound*
10.1 Avaliando o problema: Se a solução ótima encontrada do problema relaxado for inteira, então esta é solução ótima do problema original. Senão, inicie a lista de subproblemas a serem avalidos.
10.2 Ramificação: Para as componentes que não satisfazem o teste de integralidade apresentada no passo 2 do algoritmo proposto, baseando-se em Borches e Mitchell (1992), utiliza-se a seguinte estratégia: para cada x_i, se $x_i \geq 0,99$ (ou tão próximo de 1 quanto for preciso) assuma $x_i = 1$ e faça $x_j = 0$ para todos i restantes $(i = j + h)$ e $j \neq i$. Senão faça $x_i = 0$. As variáveis restantes, diferentes de 0 ou 1, que ainda não atenderam o critério de integralidade, e satisfazem o critério apresentado dos autores citados, devem ser ramificadas, ou seja, divididas em dois subproblemas: $x_i = 1$ e $x_i = 0$, para o início do processo de separação e ramificação.
10.3 Seleção do nó: Selecione um subproblema da lista e resolva a relaxação corrente utilizando o método previsor-corretor primal dual de pontos interiores.
10.4 Avaliando os nós: Percorra todos os nós, verificando

a. *Viabilidade* - Se o problema é inviável, o nó é podado. Caso contrário, esta é considerada solução ótima do problema, e o valor da função objetivo para esta, torna-se um limite para a otimalidade.

b. *Integralidade* - Se a solução assume valor inteiro (0 ou 1), o nó é podado. Lembrando que, somente uma componente pode assumir o valor 1 para cada nível da árvore. Armazene sempre o menor valor da função objetivo encontrado.

c. *Otimalidade* - Se o valor da função objetivo obtido através da solução for maior que um limite encontrado, então o nó é podado. Observação: a cada poda, é escolhido um novo subproblema em seleção do nó.

d. Se o valor obtido é fracionário, uma nova ramificação é feita, e todos os passos são feitos novamente.

10.5 Obtendo a solução: Após percorrer todos os nós, a solução que retorna o melhor valor factível da função objetivo, atendendo os critérios de integralidade, viabilidade e otimalidade, é a solução ótima do problema.

O algoritmo previsor-corretor primal-dual e *branch-and-bound* PDBB é definido através de um procedimento envolvendo os métodos primal-dual e *branch-and-bound* e é proposto para a resolução do modelo definido na seção 2.2, da seguinte forma:

i) Os passos de 1 a 9 resolvem o modelo relaxado para as variáveis limitadas superiormente $0 \leqslant x_i \leqslant u_i$;

ii) O passo 10, que utiliza o método *branch-and-bound*, integraliza a variável xi ($x_i = 0$ ou $x_i = u_i = 1$) para cada nível da árvore. A variável $x_i = 1$ implica que a variedade i deverá ser plantada em um talhão j, fazendo-se as demais $x_p = 0$, para $p = 1, \ldots, k$, tal que $p \neq i$, em que k é o número de talhões considerados.

4 Resultados

Para a obtenção de resultados, o algoritmo previsor-corretor primal-dual e *branch-and-bound* (PDBB), visto na seção 3.1, foi implementado no software Borland C++ Builder 6.0. Assim, foi possível fazer a aplicação do método PDBB ao modelo multiobjetivo investigado referente à minimização dos custos de colheita de cana-de-açúcar e coleta de resíduos, e à maximização de energia relativa ao aproveitamento de resíduos, apresentado na seção 2.2, para obter o conjunto de soluções eficientes desse modelo. Com o intuito de validar estas soluções, comparamos com os resultados obtidos com o aplicativo Solver do software Excel. Para os cálculos referentes ao modelo multiobjetivo investigado, utilizou-se os dados das tabelas de 1 a 5, apresentadas por [Florentino, 2006], [Lima, 2009], [Homem, 2010], [Ramos, 2010] e [Silva, 2011].

A tabela 1 apresenta as estimativas por tipo de variedades, considerando 10 variedades, em que V_i representa a estimativa do volume do palhiço em toneladas da variedade i; Pb_i , a produtividade de palhiço da variedade i; EC_{Bi}, o poder calorífico útil do palhiço produzido pela variedade i; A_i , a produtividade de açúcar fermentescível (POL) da variedade i; Q_i , a estimativa do volume do palhiço por unidade de área plantada da variedade i; F_i , a produtividade de fibra da variedade i; e P_c é a produtividade da cana-de-açúcar da variedade i.

Tabela 1: Estimativas de valores por variedade

i	Variedade	V_i	Pb_i	EC_{Bi}	A_i	Q_i	F_i	P_c
1	SP80-1816	7,964	33,360	2671,990	16,420	354,200	13,940	100,000
2	RB72454	8,610	37,580	2649,950	20,400	299,280	12,900	186,000
3	SP80-3280	9,369	36,720	2602,140	18,460	316,180	12,630	158,000
4	SP81-3250	10,619	34,250	1947,850	18,380	320,850	11,320	179,000
5	RB855536	9,780	26,430	2211,950	17,050	258,460	12,510	165,000
6	RB855113	10,870	29,380	2310,370	17,540	319,380	10,910	155,000
7	SP79-1011	8,910	24,090	1977,470	15,800	214,720	10,330	158,000
8	RB835486	9,560	21,530	2444,200	12,840	205,770	9,280	155,000
9	RB711406	12,320	33,200	2008,830	20,770	410,290	16,120	183,000
10	SP70-1143	7,050	22,140	1924,800	15,010	155,980	11,590	155,000

A tabela 2 apresenta os custos, consumos, e recomendações referentes às variedades em que $Cecc$ representa o custo para enleirar, compactar e carregar o palhiço; Co, o consumo de combustível do caminhão usado no transporte do palhiço; P, o preço de um litro de combustível; V_c, a capacidade de carga do caminhão a ser usado no transporte do palhiço; Ec_EC , a energia consumida pelas máquinas para enleirar e compactar uma tonelada de resíduo; Ec_C , a energia consumida pela máquina para carregar o caminhão com uma tonelada do resíduo; Ec_T, a energia consumida pelo caminhão para o transporte do resíduo; $\bar{A}$, a quantidade mínima recomendada de POL; $\bar{F_I}$ e $\bar{F_S}$ a produção mínima e máxima de fibra.

Tabela 2: Custos, consumos e recomendações

Ce_{CC}	Co	P	Ec_{EC}	Ec_C	Ec_T	V_c	$\bar{A}$	$\bar{F_I}$	$\bar{F_S}$
US\$.t^{-1}	L.km^{-1}	US\$.L^{-1}	MJ.t^{-1}	MJ.t^{-1}	MJ.km^{-1}	m^{-3}	t.ha^{-1}	t.ha^{-1}	t.ha^{-1}
7,03	0,125	0,85	7,56	57,54	5,25	54,57	14	11	15

A tabela 3 apresenta os custos referentes ao processo de transporte. A tabela 4 apresenta os custos referentes ao processo de colheita. E por fim, a tabela 5 apresenta a área e a distância dos talhões à usina, considerando 16 talhões.

Tabela 3: Custos envolvidos no processo de colheita

Transporte	Custo
Cana crua	6,42
Cana queimada	5,35

Tabela 4: Custos envolvidos no processo de transporte

Operação	Custo
Aceiro	0,14
Queima	0,17
Corte cana queimada	7,03
Corte cana crua	10,5
Carregamento	1,62

Tabela 5: Área e distância dos talhões até a usina

Talhão j	1	2	3	4	5	6	7	8
L_j	8,490	4,520	58,180	4,220	5,740	6,610	30,410	5,080
D_j	3,490	2,490	16,080	3,490	2,590	2,590	15,330	8,300
Talhão j	9	10	11	12	13	14	15	16
L_j	12,010	54,950	38,660	3,780	10,430	6,150	8,790	57,790
D_j	9,240	12,630	16,430	8,250	7,800	8,590	2,250	17,200

E ainda, para a aplicação do método PDBB ao modelo apresentado na seção 2.2, foram definidos os talhões 3 e 11 para áreas semi-mecanizáveis, e os demais para as áreas mecanizáveis.

4.1 Análise de resultados

Encontar uma solução ótima para o modelo multiobjetivo não é possível, visto que os objetivos de minimizar o custo total e maximizar o balanço de energia total da colheita da cana-de-açúcar, coleta e aproveitamento dos resíduos são conflitantes. Desta forma, utiliza-se as estratégias da soma ponderada e do ε-restrito, definido pelos problemas (2.10) e (2.11), respectivamente, que transformam o modelo multiobjetivo (2.8) em problemas mono-objetivo. Assim, a partir da aplicação do método PDBB visto na seção 3 a estes problemas, pode-se obter um conjunto de soluções eficientes, as quais determinam a curva de Pareto.

Os resultados referentes ao problema (2.10) foram obtidos através da ponderação das funções objetivo a partir da variação de $\alpha \in [0, 1]$. O tempo computacional gasto para a determinação do conjunto de soluções eficientes destes subproblemas foi de 13,9 segundos. A figura 1 apresenta a curva de Pareto referente à aplicação do método PDBB a este problema, construído de acordo com os valores das funções objetivo apresentados na tabela 6.

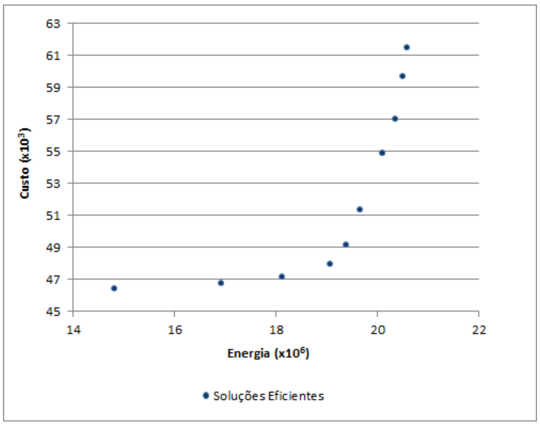


Figura 1: Gráfico que representa as soluções eficientes obtidas pela estratégia da soma ponderada

Tabela 6: Valores de α que geraram a curva de Pareto ótimo

α	Custo (CT)	Energia (BET)
0	615158,02	20597468,47
0,1	597149,49	20502583,59
0,22	570449,91	20374393,59
0,24	548832,61	20112221,03
0,243	548744,44	20111042,03
0,245	513690,82	19672598,63
0,25	491706,68	19391530,67
0,7	479792,4	19085493,7
0,8	471663,28	18130372,76
0,9	467847,69	16942898,79
1	464043,93	14819555,52

Fixando valores diferentes para o limitante superior ε da restrição 2.12, é possível obter os resultados referentes ao problema (2.11). Para a determinação do conjunto de soluções eficientes destes subproblemas, o método gastou um tempo computacional de 1 minuto e 22,5 segundos. A figura 2 apresenta a curva de Pareto referente à aplicação do método PDBB a este problema, construído de acordo com os valores das funções objetivo apresentados na tabela 7.

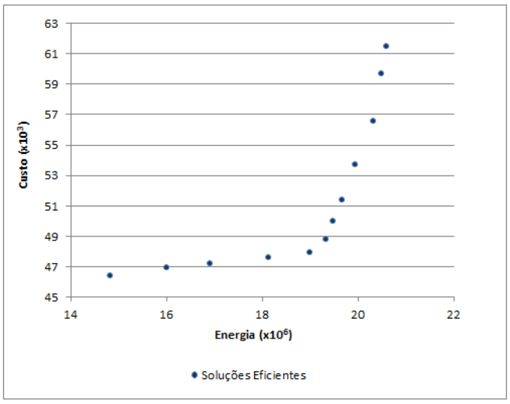


Figura 2: Gráfico que representa as soluções eficientes obtidas pela estratégia do ε-restrito

Tabela 7: Valores de ε que geraram a curva de Pareto ótimo

ε(x10^6)	Custo(CT)	Energia (BET)
20,597	615158,02	20597468,47
20,5	597149,49	20502583,59
20,3	565564,1	20317323,13
19,8	537076,52	19940348,78
19,6	513602,66	19671419,63
19,45	499744,66	19475029,86
19,3	487964,88	19340695,44
19	478936,53	19007485,04
18,1	475951,93	18134310,43
16,9	471937,26	16913348,45
16	469049,92	16009823,53
14,81	464043,93	14819555,52

Os gráficos 1 e 2 representam a relação entre os valores de custo e balanço de energia total, e o fato destes objetivos serem conflitantes, melhorar um objetivo, implica em piorar o outro. As soluções eficientes obtidas são de interesse, pois informam ao produtor a determinação das variedades a serem plantadas, de modo a atender o momento econômico de sua empresa. De modo geral, a aplicação do método PDBB à estratégia do ε-restrito gera um número maior de soluções eficientes do que a estratégia da soma poderada, visto que esta estratégia permite que o usuário escolha o limitante ε, de acordo com suas necessidades. Enquanto que, a estratégia da soma poderada, depende da variação do $\alpha \in [0, 1]$ e da normatização dos coeficientes das funções objetivo.

Todos os resultados obtidos pelo método PDBB foram comparados com aqueles apresentados pelo aplicativo Solver do software Excel. A partir desta comparação, pode-se concluir que ambos determinam as mesmas variedades a serem plantadas nos talhões, e em relação aos valores das funções objetivo, obtêm os mesmos resultados, se diferenciando apenas a partir da quinta ou sexta casa decimal.

5 Conclusões

Neste trabalho, desenvolveu-se um procedimento híbrido que envolve os métodos previsor-corretor primal-dual de pontos interiores e *branch-and-bound* (método PDBB), para serem aplicados em problemas de programação linear inteira 0-1. Primeiramente, obtém-se a solução relaxada do PPL através do método previsor-corretor primal-dual de pontos interiores, e em seguida, as varáveis fracionárias identificadas pelo método PDBB são integralizadas pelo método *branch-and-bound*, determinando-se a solução ótima 0-1 do problema.

O método PDBB foi utilizado para a resolução do modelo multiobjetivo de minimizar o custo da colheita da cana-de-açúcar, e o custo da coleta da biomassa residual, e/ou maximizar o balanço de energia relativo à biomassa residual do processo, baseando-se em [Silva, 2011], [Ramos, 2010], [Homem, 2010], [Florentino et al., 2011] e [Florentino, 2006]. Este foi testado para um caso específico de dezesseis talhões e dez variedades, considerando as áreas dos talhões 3 e 11 semi-mecanizáveis, e as demais, mecanizáveis.

O algoritmo do método PDBB foi implementado através do software Borland C++ Builder 6.0, e aplicado aos problemas destacados, através das estratégias da soma ponderada e ε-restrito, obtendo as soluções eficientes para os casos considerados. Estas foram comparadas com aquelas determinadas pelo aplicativo Solver, do software Excel, e os resultados obtidos, mostraram o bom desempenho do método híbrido proposto.

Portanto, o trabalho atingiu os seus objetivos, relacionados à proposta e ao uso do método PDBB à resolução do modelo de cultivo e de aproveitamento da biomassa residual da cana-de-açúcar, mostrando a viabilidade de se utilizar esta técnica de otimização para auxiliar as usinas na seleção de variedades a serem plantadas, de tal forma a otimizar o processo, respeitando-se as restrições de produção caracterizadas no modelo.

Agradecimentos

Agradecemos a CAPES pela bolsa de mestrado.

Referências

[Bazaraa e Shetty, 1979] Bazaraa, M. S. and Shetty, C. (1979). Nonlinear Programming: Theory and Algorithms *John-Willey & Sons, Inc.*

[Borchers e Mitchell, 1992] Borchers, B. and Mitchell, J. E. (1992). Using an interior point method in a branch-and-bound algorithm for integer programming. *Technical Report 195, Mathematical Sciences, Resselaer Polytechnic Institute, Troy, NY 12180, March 1991, Revised July 7, 1992.*

[Deb, 2004] DEB, K. (2004). Multi-objective optimization using evolutionary algorithms. *John-Willey & Sons Ltda.*

[Florentino, 2006] Florentino, H. O. (2004). Programação linear inteira em problemas de aproveitamento da biomassa residual de colheita da cana-de-açúcar. 64f. *Tese (Livre Docência) Instituto de Biociências de Botucatu, Universidade Estadual Paulista, Botucatu, SP.*

[Florentino et al., 2011] Florentino, H. O.; Lima, A. D; Carvalho L.; Balbo, A. R.; Homem, T. P. D. (2011). Multiobjective 0-1 integer programming for the use of sugarcane residual biomass in energy cogeneration. *International Transactions in Operational Research*, v.18, p.605-615.

[Granville, 1994] Granville, S. (1994). Optimal Reactive Dispatch Through Interior Point Methods. *IEEE Transactions on Power Systems*, v.9, p.136-146.

[Homem, 2010] Homem, T. P. D. (2010). Procedimento híbrido envolvendo os métodos primal-dual de pontos interiores e branch-and-bound em problemas multiobjetivo de aproveitamento de resíduos de cana-de-açúcar. *Dissertação (Mestrado em Engenharia Elétrica), Faculdade de Engenharia, Universidade Estadual Paulista, Bauru, SP.*

[Homem et al., 2011] Homem, T. P. D.; Balbo, A. R.; Silva, H. O. F. (2011). Optimal energy generation with biomass of sugarcane harvest. *Revista IEEE América Latina*, v.1, p.653-658 .

[IBGE, 2012] IBGE - Instituto Brasileiro de Geografia e Estatística. (2012). Levantamento Sistemático da Produção Agrícola - Confronto das safras de 2011 e 2012 - Brasil - Fevereiro de 2012. *Disponível em http://www.ibge.gov.br/home/estatistica/indicadores/agropecuaria/lspa/lspa_201202_5.shtm, acesso em 20 de março de 2012.*

[Lima, 2009] Lima, A. D. (2009). Otimização do aproveitamento do palhiço de cana-de-açúcar. *Tese (Doutorado em Energia na Agricultura), Faculdade de Ciências Agronômicas, Universidade Estadual Paulista, Botucatu, SP.*

[Mehrotra, 1992] Mehrotra, S. (1992). On the implementation of a primal-dual interior point method. *SIAM Journal on Optimization*, v.2, p.572-601.

[Mitchell, 1995] Mitchell, J. E. (1995). Fixing variables and generating classical cutting planes when using an interior point branch-and-cut method to solve integer programming problems.

[Pellegrini, 2002] Pellegrini, M. C. (2002). Inserção de centrais cogeradoras a bagaço de cana no parque energético do Estado de São Paulo: exemplo de aplicação de metodologia para análise dos aspectos locacionais e de integração energética. *Dissertação (Mestrado em Energia), Universidade de São Paulo, São Paulo, SP.*

[Ramos, 2010] Ramos, R. P. (2010). Modelo matemático para custo e energia na produção de açúcar e álcool. *Dissertação (Mestrado em Agronomia/Energia na Agricultura), Faculdade de Ciências Agronômicas, Universidade Estadual Paulista, Botucatu, SP.*

[Ripoli e Ripoli, 2004] Ripoli, T. C. C. Ripoli, M. L. C. (2004). Biomassa de cana-de-açúcar: colheita, energia e ambiente. *Piracicaba, SP.*

[Silva, 2011] Silva, L. M. (2011). Biomassa de cana-de-açúcar: colheita, energia e ambiente. *Piracicaba, SP.*

[Tolentino, 2007] Tolentino, G. (2007). Programação Linear Inteira Aplicada ao Aproveitamento do Palhiço da Cana-de-Açúcar. *Dissertação (Mestrado em Agronomia/Energia na Agricultura), Faculdade de Ciências Agronômicas, Universidade Estadual Paulista, Botucatu, SP.*

[Wu et al., 1994] Wu, Y. C.; Debs, A. S. and Marsten, R. E. (1994). A Direct Nonlinear Predictor-Corrector Primal-Dual Interior Point Algorithm for Optimal Power Flows. *IEEE Transactions on Power Systems*, 9, No.2, p.876-883.

Development of a Multicriteria Decision Aiding Model for monitoring and evaluating the performance of Health Care Units

Diana F. Lopes

CEG-IST, Instituto Superior Técnico, Technical University of Lisbon, Avenida Rovisco Pais, 1049-001 Lisbon, Portugal
E-mail:`diana.lopes@ist.utl.pt`

Abstract

Given the increasing restrictions on the Portuguese health sector, it is critical to have a monitoring of the performance of health care providers. Within this context, this study aimed to assist the South West Primary Health Care Group in building a system to monitor and evaluate the performance of Health Units (HU). For that purpose, it was developed a multicriteria model based on MACBETH and on the Choquet integral, an innovation in the health care literature, being applied to four HU. The development of methods has shown that this model is still in its first steps and needs further research.

Keywords: Primary Health Care, Multicriteria Value Measurement, MACBETH, Choquet integral, Interactions.

1 Introduction

Primary Health Care (PHC) is a key component of the Portuguese's health system, representing the first level of contact of individuals with the National Health Service [1]. Regarding the problems and challenges that the Portuguese PHC sector was facing, in 2005 a set of reforms was initiated. The objective has been to create more and better health care for citizens and to improve accessibility, proximity and quality in health care provision. This set of reforms is still in course and essentially aims to reorganize and modernize Primary Health Care Centres, structuring them in Functional Units (FU) with complementary missions. In addition, a contracting system was recently introduced aiming to bring more accountability and efficiency into the primary care system in line with improving health outcomes given available resources [2].

Given budget cuts and limited resources and the objectives of the Portuguese health care system, it is critical to have a correct identification of major problems of PHC providers. Within this context, the South West Primary Health Care Group (PHCG) wanted to properly evaluate the performance of FU subject to the contracting processes - Health Units (HU): Family Health Units (FHU) and Personalized Health Care Units (PHCU) - in order to identify the need for corrective actions. In fact, the correct identification of major problems in their HU and appropriate management of available resources are imperative for proper decision making. Thus, this study was developed in an attempt to meet the request of the South West PHCG: build a system to evaluate and monitor the performance of HU.

To achieve these objectives, it was made a literature review. The review has shown that literature to address that challenge is scarce and does not provide methods for, in a rational and transparent way, helping and supporting the South West PHCG's objective in a satisfactory way. Therefore this study has developed a multicriteria model based on MACBETH and on an extension of MACBETH to the Choquet integral (CI) operators to evaluate and monitor the HU's performance. With the latter methodology we intended to explore, in an innovative way, the process of modelling interacting points of view (PV).

This paper is structured as follows: Section 2 presents a brief review of related literature. Section 3 contains the proposed methodology. Section 4 describes the implementation of the proposed methodology. Section 5 shows the results of the model application. Finally section 6 presents some discussion and concluding remarks.

2 Literature Review

Performance measurement is fundamental to every management control. Records of performance measurement efforts in health systems can be traced back to at least 250 years ago [3]. However, just over the past thirty years there has been a dramatic growth in health system performance measurement. A confluence of forces has led to this growth. On the supply side, great advances in information technology have made it much cheaper and easier to collect and process data [3]. On the demand side, health systems have come under intense cost containment pressures and there has emerged a growing demand for increased oversight and accountability in health professions and health institutions [3].

Some of the most well known performance measurement frameworks are Six Sigma approach [4], European Foundation for Quality Management model [4], Performance Pyramid [5], Results and Determinants Framework [5] and Balanced Scorecard [6]. In fact, these frameworks emphasize the need of measurement systems to make explicit the trade-offs between the various performance measures, but are vague on how to deal with these trade-offs. Furthermore, there is no information about its aggregation mechanisms and some of the frameworks is related to high costs of implantation [7].

Accordingly, the use of a Multicriteria Decision Analysis (MCDA) approach can be helpful in this context. Up to our knowledge, multicriteria models found in the health literature are scarce and the existing ones do not take into account possible interactions between PV (nor have been used for performance measurement of health care providers). Thus, these models are inaccurate or limited to a few cases where there is, in fact, independence and absence of interactions between PV [8]. In the industrial sector, however, some studies exploited in an integrated manner the interaction phenomena between PV, with the use of the CI operator [9,10,11].

To understand interacting phenomena, it is useful to distinct two preferential conditions [12]: (1) ordinal independence and (2) cardinal independence. The first one is verified when alternatives or options can be ranked with respect to one PV independent of their impact in other aspects; the second one is verified, when in addition the difference in attractiveness between the alternatives can be measured with respect to one PV independent of their impact in other aspects [12]. For instance, consider the following two PV "peak noise level during the night" and "noise level during the whole day" [12]. Whatever the average noise during the day, low peaks during the night are always preferred to high peaks (ordinal independence) [12]. However, the difference of attractiveness between two night peak levels may not remain the same independent of the average noise level during the whole day (cardinal dependence may occur) [12]. In this sense, interactions are related to preferential dependence. They could be expressed in different ways, such as: (a) additivity, the combined effect is the sum of effects; (b) antagonism, the combined effect is less than the sum; and (c) synergism, the combined effect is more than the sum [12].

The literature is not consensual about the interaction phenomena. A common pitfall of analysis is to associate environmental (physical or statistical) relations between PV with independence conditions [13]. Note that these conditions refer to judgments or preferences. Thus, environmental correlations between PV can be ignored unless they are redundant [13,14]. Other pitfall consists of distinguishing *substitutiveness* and *complementarity* from preferential dependence [13]; in fact, the former are types of preferential dependence.

In order to tackle all the gaps mentioned above, this study explored a MCDA model based on the extension of MACBETH to the CI operators, following some steps of the procedure present in the CI application studies [9,10,11]. In fact, this framework allows: (1) to build a transparent model through an iterative and social process; (2) to take into account multiple PV relevant to the Decision-Makers (DM); (3) to consider the interactions between PV; and (4) to determinate the HU's performance in different levels of specification.

3 Methodological Framework

The construction of the multicriteria framework aimed to determine HU's performance and involved several and interconnected steps built using a socio-technical approach (figure 1): (1) model structuring, being constructed a value tree with the relevant areas and PV for DM to evaluate the HU; (2) construction of a value function in each PV; (3) aggregation of the multiple dimensions of performance into a global one, through the use of (a) MACBETH and (b) an extension of MACBETH to a particular CI operator (2-additive CI) in the final branches of the value tree (1^{st} level of the hierarchy) - with (a) and (b) being determinant to measure the attractiveness at the 2^{nd} level of the hierarchy (subareas) -, and (c) the hierarchical additive model to measure the attractiveness of HU in the 3^{rd} level of the hierarchy and in global terms; and (4) construction of performance categories, to more easily distinguish the need for corrective actions.

The social component of building a multicriteria model included various meetings with the Executive Director (ED) of the South West PHCG in the model structuring and construction of performance categories, as well as two decision conferences with the group of DM (ED and the Clinic Council President of the South West PHCG, the coordinator of the PHCU of Torres Vedras and two doctors) for the construction of value functions and determination of area and PV weights.

The methodology and its various activities are specified in figure 1.

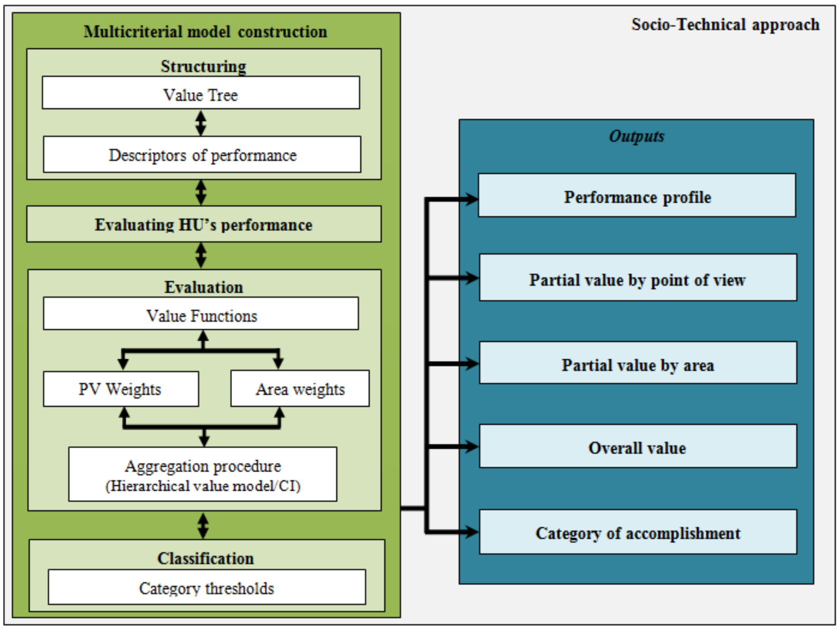


Figure 1: Activities involved in the construction of the methodology. Double arrows mean that this is an iterative model, allowing going back and forth whenever necessary.

3.1 Model structuring

Structuring the multicriteria model consisted of an interactive learning process between the DM, ED and Clinic Council President of the South West PHCG, in order to (1) specify the relevant PV for the DM and (2) operationalize each PV.

To that end, firstly, it was necessary to discuss the South West PHCG's strategic objectives in order to recognize the main areas of concern. Then, it was necessary to reflect about these areas in order to identify specific PV. In the sense of providing a practical visual overview of the structure of concerns in several levels of increasing specification, it is useful to group the PV in a value tree according to the areas of concern [12]. The final model was composed of 3 areas, each one integrating a coherent and an exhaustive set of non-redundant PV (23 in total) (see figure 2). The initial phase of structuring the PV showed the existence of three main areas of concern: *Access*, *Performance of care* and *Efficiency*, which were further explored in a top down approach. Starting by the first branch, *Access*, the DM explained that this concern was mostly related to the use of family planning services, general practice and home visits. A total of 5 PV were identified in this branch. The second branch, *Performance of care*, involved aspects of *Adult Health*, *Maternal Health* and *Child Health*, which correspond to the three subareas in this branch (covering 16 PV). Finally, in the *Efficiency*, it was important to evaluate the policy of medicines and means of diagnosis and therapeutics prescription. The full structure of the value tree is represented in figure 2.

Secondly, it was necessary to operationalize each PV by assigning a descriptor of performance and defining levels of reference [15]. The use of levels of reference (like "target" and "base" used in this study) has three objectives: i) it contributes significantly to the intelligibility of the PV [16]; ii) it objectifies the notion of intrinsic attractiveness of each HU [16]; and iii) it allows us to construct intra-PV and inter-PV information [17]. Note that at the two reference levels, "target" and "base", were assigned 100 and 0 points, respectively (see equation (2)) [18]. Given the nature of the elicited PV, all of the descriptors

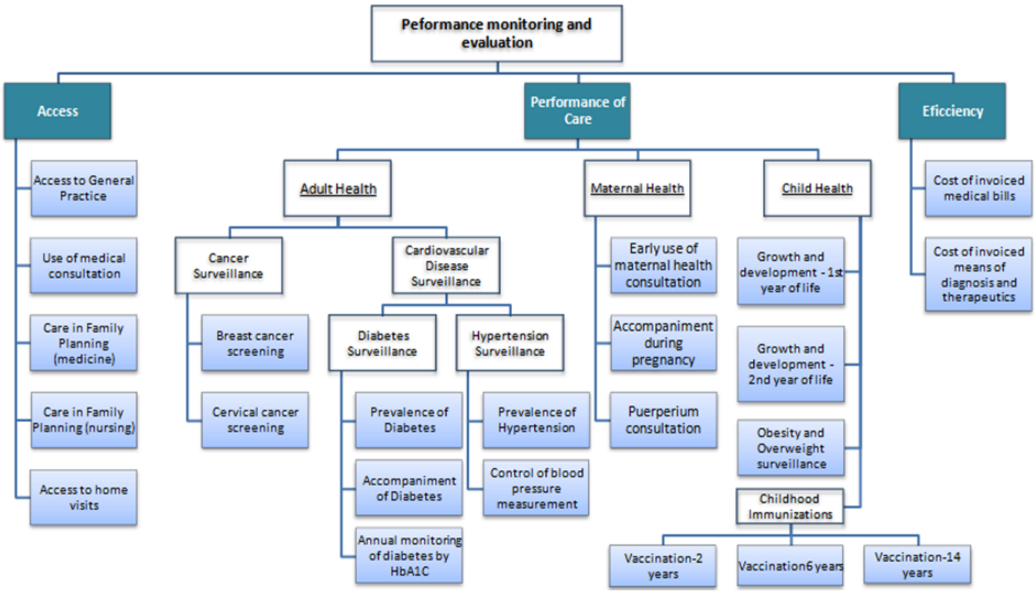


Figure 2: Representation of the value tree. The blue-shadowed nodes, in the final branches of the tree, represent PV.

were quantitative and described by a mathematical formulation. However, it is important to note that, in addition to the two reference levels assigned in each PV ("base" and "target"), plus three performance levels were defined to the majority of the descriptors, with the same meaning ("minimum plausible", "desirable" and "maximum plausible"). The descriptor of performances of the PV *Care in Family Planning (medical)* is on table 1.

Table 1: Descriptor of performances of the PV *Care in Family Planning (medical)*.

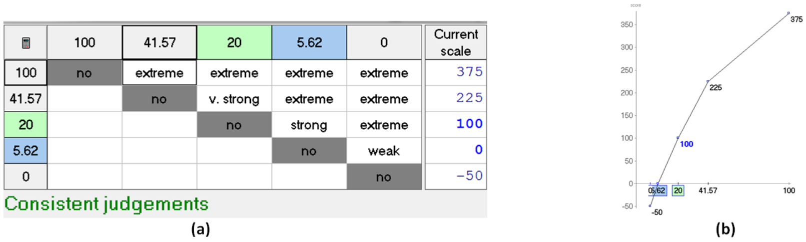
Care in Family Planning (medical)	Levels of performance	
	Minimum plausible	0%
	Base	5.62%
$\frac{Number\,of\,women\,with\,medical\,consultation\,in\,Family\,Planning}{Number\,of\,women\,enrolled\,on\,the\,HU\,([15,49]\,years\,old)} \times 100$	Target	20%
	Desirable	41.57%
	Maximum plausible	100%

3.2 Construction of Value Functions

The process of building a value function for each PV permits the translation of performance into value. The value function on PV i is often constructed by asking questions regarding the preference of the DM on the set X_i (descriptor of performances on PV i) rather than his preferences on $X = X_1 \times X_2 \times ... \times X_n$ [19]. This procedure is completely justified when the additive model is the aggregation function since each PV can be isolated thanks to preferential independence property [19]. When PV interact together, the construction of the value function on X_i must be performed by asking information regarding elements of X [19]. Thus, the construction of each value function was carried out by considering that the HU were on the "base" level of performance on the other PV. This is not considered in the application studies [9, 10], being the value functions constructed regarding the set X_i.

The adoption of the MACBETH approach for this purpose in the South West PHCG context was motivated by the deliberate intent of avoiding the difficulties of some DM to elicit quantitative judgments. In fact, the MACBETH procedure (with the help of M-MACBETH software) allows constructing the value functions by means of qualitative judgments of differences in attractiveness between performance levels, two at a time [20]. To facilitate each pairwise comparison the DM is asked to choose one of the MACBETH qualitative categories of difference in attractiveness: "is the difference very weak, weak, moderate, strong, very strong, or extreme?" [14]. During this questioning process, the facilitator fills in a matrix with the categorical judgments of DM. Note that judgment disagreement or hesitation between two or more categories is also allowed [14]. Each time a qualitative judgment is elicited, the consistency is verified and suggestions are offered to solve eventual inconsistencies [14]. After the consistency verification, the software derives a value function at each PV, which has to be validated by the DM [20]. Note that these functions or scales of performance values could be not limited to the same range of values in every PV and concern negative and positive values being called "unlimited bipolar scales" [21], which was the case of this study.

This activity was carried out in a decision conference with the ED and the Clinic Council President of the South West PHCG, the coordinator of the PHCU of Torres Vedras and two doctors. Figure 3 presents the filled MACBETH matrix (a) that originate the value function of PV *Care in Family Planning (medical)* (b).



	100	41.57	20	5.62	0	Current scale
100	no	extreme	extreme	extreme	extreme	375
41.57		no	v. strong	extreme	extreme	225
20			no	strong	extreme	100
5.62				no	weak	0
0					no	-50

Figure 3: (a) Judgment matrix and (b) value function for the PV *Care in Family Planning (medical).*

In order to evaluate the HU's attractiveness at the level of each subarea, area and in global terms, the value scores on the PV should be aggregated [18]. To that end, firstly, it was necessary to determine the weights of each PV and area.

3.3 Weighting PV

Trade-off, Swing-Weighting and MACBETH are the main methods of eliciting weights. This study, which involved a value tree with an hierarchical structure, proposed the determination of PV weights, through the MACBETH approach (see section A) and through the extension of MACBETH to 2-additive CI operator (see section B), being the latter an innovation in the health care literature. Then, it was possible to measure the attractiveness at the 2^{nd} level of the hierarchy. To measure the attractiveness of HU in terms of areas belonging to the 3^{rd} level of the hierarchy and in global terms it was proposed to use a hierarchical additive model.

A. MACBETH weighting process

The adoption of this procedure implies an additive value model, which implicitly assumes "difference independence", which means: the difference in attractiveness between performance level on one PV does not depend and can be measured independently of the levels on other PV (being called criteria) [18]. So, the first step, in this procedure, consists on verifying this hypothesis.

The mathematical formulation of its aggregation operator is given by the expression:

$$V_{Ag}^{u}(x_{1u},...,x_{nu}) = \sum_{i=1}^{n} w_i v_i(x_{iu}), \text{with } \sum_{i=1}^{n} w_i = 1, w_i > 0. \tag{3.1}$$

$$v_i(target_i) = 100 \text{ and } v_i(neutral_i) = 0 \tag{3.2}$$

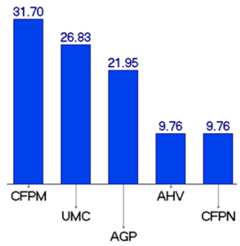
where V_{Ag}^{u} is the performance value of the HU u, which measures its attractiveness in terms of each area; n is the number of criteria involved; x_{iu} is the impact of u in the criterion i; w_i is the weight of criterion i; $v_i(x_{iu})$ is the partial value of u in the ith criterion; $(target)_i$ and $(base)_i$ are, respectively, the target and base impact levels in the descriptor of the ith criterion.

Let $(HU)^F$ (with $F = 0,...,n$) be a reference set of fictitious HU, where $(HU)^0$ is a reference fictitious HU with the base impacts in all criteria, and $(HU)^j$ $(j = 1,..,n)$ is a reference fictitious HU with the target impact in the jth criterion and the base impacts in all the other criteria. Note that n is the number of criteria present in each area. After this mathematical overview, the MACBETH procedure for weighting the criteria consists in asking the DM: (1) to order the $(HU)^j$ in terms of attractiveness; (2) to judge qualitatively the difference of attractiveness between any two reference HU [15]. These judgments are introduced in a matrix present in the M-MACBETH software, being, each at a time, consistently verified. After that, the software derives a weighting scale, which needs to be discussed, possibly adjusted, and then validated by the DM.

In particular, it was adopted the MACBETH procedure to determine the PV weights present in the *Access* area and in the *Child Health* subarea, in a decision conference with the same people mentioned in the previous activity. This adoption was due to the number of PV in each of these areas (5 and 6, respectively), in which the 2-additive CI procedure would be impractical due the time and complexity. Note that, before this adoption, it was verified the “difference independence” in these PV. Figure 4 presents the MACBETH matrix filled with the DM’s judgements, used to determine the criteria weights in the *Access* area.

	[CFPM]	[UMC]	[AGP]	[AHV]	[CFPN]	[tudo inf.]
[CFPM]	no	weak	weak	moderate	moderate	extreme
[UMC]		no	weak	positive	positive	v. strong
[AGP]			no	moderate	positive	moderate
[AHV]				no	no	weak
[CFPN]				no	no	weak
[tudo inf.]						no

(a)



(b)

Figure 4: (a) Judgment matrix for (b) weighting criteria on the *Access* area. (CFPM: Care in Family Planning (Medicine); UMC: Use of Medical Consultation; AGP: Access to General Practice; AHV: Access to Home Visits; and CFPN: Care in Family Planning (Nursing)).

This procedure provides a simple and transparent approach in solving complex multicriteria problems, hence its wide applicability to most methods for MCDA [15,16,18,22,23,24,25]. However, this mechanism of aggregation does not consider possible interactions between PV. Thus, in the next subsection is described the 2-additive CI mechanism which was applied to the rest of PV present in the first level of the hierarchy.

B. The 2-additive Choquet integral

The operators of the CI family, introduced by Choquet [26], belong to the non-additive measures family and their basic idea consists on assigning a “weight” to every possible subset of PV (i.e. coalitions of PV) [27]. Due to its exponential complexity, many particular families of this operator have been proposed to solve this issue, such as the k-additive ones [19]. K-additivity property fixes the degree of interaction between PV: 1-additivity does not permit interaction (additive model); 2-additivity allows interaction up to 2 PV, etc. Note that a k-additive operator needs $\sum_{a=1}^{k}\binom{n}{a}$ coefficients to be defined, which makes, in practice, the 2-additivity the best compromise between complexity and richness of the model [19]. Thus, in this study we proposed the use of the 2-additive CI operator, which dictates the explanatory nature of this study. This operator is based on two types of parameters [21]:

- Importance of each PV in relation to all the other contributions to the overall performance evaluation by the Shapley parameters (s_i) that satisfy the condition $\sum_{i=1}^{n} s_i = 1$;
- The interaction parameters I_{ij} of any pair of PV, that range in [-1,1]: (a) $I_{ij} > 0$ implies a synergistic behaviour between PV i and j, (b) $I_{ij} < 0$ implies a antagonistic behaviour between PV i and j, and (c) $I_{ij} = 0$ implies that no interaction exists and s_i acts as the weights in the additive model.

The associated aggregation function is given by:

$$V_{Ag}^{u} = \sum_{i=1}^{n} s_i v_i(x_{iu}) - \frac{1}{2}\sum_{i,j,i\neq j} I_{ij}|v_i(x_{iu}) - v_j(x_{ju})|. \tag{3.3}$$

with the property $\left(s_i - \frac{1}{2}\sum_{i\neq j}|I_{ij}|\right) \geq 0$, $\forall i \in [1,n]$ and $j \neq i$.

However, it is important to note that this mathematical formulation is coherent with the performance values on an unipolar interval scale (on the universe [0,1]) [21]. As mentioned above, the performance values are on an unlimited bipolar scale. So, to use this operator, it was necessary to convert the scale proposed by MACBETH (value function) in the unipolar scale required.

To determine the CI parameters, this paper followed the procedure present in the CI application studies [9,10,11]. According to them, firstly, the DM have to compare fictitious HU characterized by having all the performances in the best and/or worst levels. However, since the objective is to make such comparisons more realistic, simpler and understandable, we considered more appropriate the use of fictitious HU based on target and/or base levels. The number of fictitious HU is intended to be equal to the sum of the CI parameters plus one (α, which is a coefficient necessary to guarantee that performance values are in the range [0,1]) [10]. In the case of two PV, it involves the existence of 3 CI parameters (v_1,v_2 and I_{12}) and, thus, 4 fictitious HU. These fictitious HU should be (B,B), (B,T), (T,B) and (T,T), where (B,T) represents a fictitious HU with performances in the base and target levels in the first and second PV, respectively. The next step consists in their ordination in terms of attractiveness and to judge semantically the difference of attractiveness between two consecutive ordered fictitious HU. Note that these judgments are introduced in a matrix.

This procedure was applied to the rest of PV. Table 2 illustrates this process on *Cancer Surveillance* subarea. From the preference ordering of the fictitious HU and the strength of that preference order given by the DM, and with the help of equations (3), it was possible to construct an equation system. As we can see in table 2, the DM said that: "$V_{Ag}^{(B,T)}$ is strongly preferred to $V_{Ag}^{(B,B)}$". This piece of information leads to $V_{Ag}^{(B,T)} - V_{Ag}^{(B,B)} = 4\alpha \iff [v_1(B_1)s_1 + v_2(T_2)s_2 - 0,5 \times |v_1(B_1) - v_2(T_2)|] - [v_1(B_1)s_1 + v_2(B_2)s_2 - 0,5 \times |v_1(B_1) - v_2(B_2)|] = 4\alpha$. Other equations were written in the same logical, originating an equation system. The resolution of the system gave the values of the Shapley and interactions parameters and α (see table 3). Then the aggregated performance could be determined by using equation (3). From table 3, it was concluded that there were interactions, some of which cannot be ignored, between PV. The obtained positive interaction values in the PV present in *Cancer Surveillance* subarea, show that for the DM, a HU must be successful from both the PV in order to be considered successful (synergistic behaviour). The same was observed for the PV present in *Hypertension Surveillance* and *Efficiency* areas.

Table 2: Matrix with preference relations between consecutive ordered fictitious FU on *Cancer Surveillance* subarea ($UF(T,T) \succeq UF(T,B) \succeq UF(B,T) \succeq UF(B,B)$).

	UF(T,T)	UF(T,B)	UF(B,T)	UF(B,B)
UF(T,T)	Null	V.strong		
UF(T,B)		Null	Null	
UF(B,T)			Null	Strong
UF(B,B)				Null

Table 3: Obtained CI parameters.

Subarea or Area	Points of view	Shapley parameters	Interaction parameters
Cancer Surveillance	1: *Breast cancer surveillance*	0.512	$I_{12} = 0.14$
	2: *Cervical cancer surveillance*	0.488	
Diabetes Surveillance	1: *Prevalence of diabetes*	0.370	$I_{12} = -0.1138$
	2: *Accompaniment of diabetes*	0.170	$I_{13} = -0.20279$
	3: *Annual monitoring of diabetes by HbA1C*	0.460	$I_{23} = -0.08854$
Hypertension Surveillance	1: *Prevalence of hypertension*	0.560	$I_{12} = 0.42$
	2: *Control of blood pressure measurement*	0.440	
Maternal Health	1: *Early use of maternal health consultation*	0.689	$I_{12} = 6.8 \times 10^{-15}$
	2: *Accompaniment during pregnancy*	0.235	$I_{13} = -2.5 \times 10^{-16}$
	3: *Puerperium consultation*	0.076	$I_{23} = -1.9 \times 10^{-16}$
Efficiency	1: *Cost of invoiced medical bills*	0.587	$I_{12} = 0.092$
	2: *Cost of invoiced means of diagnosis and therapeutics*	0.413	

C. Weighting areas

By the previous step, it was possible to measure the attractiveness at the 2^{nd} level of the hierarchy. To measure the attractiveness of HU in terms of areas belonging to the 3^{rd} level of the hierarchy, it was proposed to use a hierarchical additive model. Thus, the scores obtained in the subareas of the 2^{nd} level are aggregated through the additive model, yielding a single score in each area (the third level). To obtain a value of overall performance for each state the procedure is identical: the scores obtained in each area are aggregated through the additive model, originating an overall performance value [18]. This model is simple, has wide acceptance in the literature [28], and was shown to be appropriate for the specific evaluation context.

In order to determine the weights from the subareas present in the 2^{nd} level of the hierarchy, it was chosen the swing weighting procedure. In a bottom-up approach, this procedure, firstly, considers the most elementary areas (2^{nd} level) and then hierarchically areas superior to them, and so successively until reaching the top of the hierarchy. The description of this procedure can be consulted in [28].

It is important to note some difficulties felt by the DM to quantitatively judge the difference of attractiveness between two stimuli.

3.4 Construction of categorical performances

In order to easily identify the needs for corrective actions, the status of the HU in each subarea, area and in overall terms can be classified within a category scale. In this study it was used a five category scale: “Very Weak”, “Weak”, “Acceptable”, “High” and “Very High”. These designations are common to all areas, but their description varies from area (subarea) to area (subarea) and also in overall terms. These categorical performances were constructed using the bottom-up procedure described in the study [18]. Due to time constraints, this activity was only realized with the ED of the South West PHCG.

4 Results

After having constructed the evaluation model, it was necessary to perform its implementation. In this sense, this model was applied on three FHU - Arandis, Dom Jordão and Gama - and one PHCU - Alenquer. Due to time constraints, the facilitator met only with the ED to build the performance profiles for each HU. These profiles are based on values observed in December 2010, since there was no later available information. The value scores obtained by the HU in each area and in overall terms are shown in figure 5, with each score assigned to a category of accomplishment. As can be seen, the FHU are in an overall very high state and the PHCU is in an overall acceptable state. It is worthwhile noting that the analysis made at the level of the areas resulted in the same conclusion. In fact the FHU’s performances values are more attractive than the PHCU’s performances values in the most levels of the hierarchy. This is in line with expectations, in the sense that the FHU have higher dynamic in their professional teams and a higher degree of organization autonomy than the PHCU. Other aspect that could be related is the fact that the FHU was the first HU to be on contracting processes. In addition, PHCU are not rewarded as a result of good practices, which may arise the demotivation for good results.

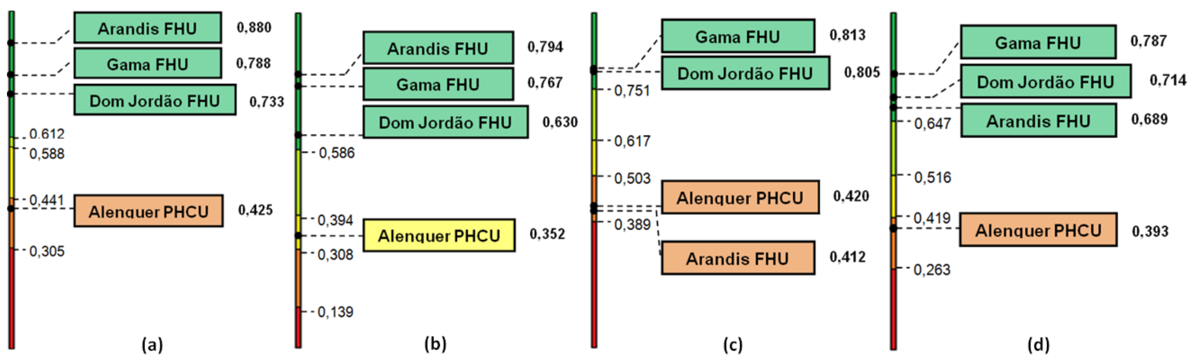


Figure 5: States of the four HU under analysis, in each area (Access (a), Performance Care (b) and Efficiency (c)) and in overall terms (d). Note that the red colour represents a very weak state; orange, a weak state; yellow, a acceptable state; light green, a high state; and dark green, a very high state.

5 Discussion

The methodology constructed intended to support South West PHCG in building a system to monitor and evaluate the performance of HU, by adopting a multicriteria methodology throughout a interactive consulting path. The methodology applied was based on the extension of MACBETH to the 2-additive CI operators that is more general than the most common additive MACBETH model formulation. In fact, these operators permit the modelling of the interactions between PV, which dictate the exploratory nature of this work. The constructed model was applied to a set of four HU: three FHU (Arandis, Dom Jordão and Gama) and one PHCU (Alenquer). The HU's performances values are in line with expectations, since FHU's performances values are more attractive than the PHCU's performances values in the most levels of specification.

Despite the consistent results, it is important to note some of the limitations associated with the constructed model. Two of them are related to the adoption of the 2-additive CI procedure: (1) despite its richness on the interaction modelling, this procedure becomes impractical and complex when applied with more than four PV (thus, it was not used in *Access* area and *Child Health* subarea); (2) since it intends to model interactions between PV, the construction of value functions within the Choquet procedure might not be appropriate in the sense that for a specific PV we can obtain different value functions if we consider that the HU are in different levels in the other PV; and (3) the value of the semantic categories used to construct value functions in M-MACBETH and to fill the matrix that allows to determine the Shapley and interaction parameters may not be the same. Other limitation is related to time constraints since two activities were constructed with a sole person (ED, being those activities dependent of his views, compromising in some way the richness of the model. Finally, it is important to state that the DM showed some difficulties to elicit quantitative judgments in swing weights procedure to the area weights determination, which might have introduced some deviations into the model. Once the results of application of the model were consistent with expectations, it is considered that the limitations of the model did not preclude the development of the methodology.

Future research might explore the use of MACBETH to build cardinal scales when points of view are preferentially dependent, namely by overall comparing a set of profiles within a MACBETH matrix, overcoming the problems present in areas with more than four PV. Also, methods should attempt to use unlimited bipolar scales, in order to make the process easier and less time consuming, avoiding the changes of scales. Although there is already a mathematical formulation adapted for limited bipolar scales ([-1,1]), this leads to a very complex process, having no examples of application in real cases [21].

Concluding, the results support the South West Primary Health Care Group in monitoring and evaluating the performance of the Health Care Units. It was also concluded that the extension of MACBETH to the CI operators is still in the first steps to make its application easier and more integrated. Given the relevance of this methodology (combination of the MACBETH with the CI) on accounting for possible judgmental interactions between PV that could be present in many contexts such as in the health sector, further research is needed to improve the formulation avoiding the referred problems.

References

[1] OMS and UNICEF (1978). Declaration of Alma-Ata. In *International Conference on primary Health Care*, Alma-Ata.

[2] Missão para os Cuidados de Saúde Primários (2010). Cuidados de Saúde Primários - Metodologia de Contratualização. In *http://mcsp.min-saude.pt/*, accessed on December of 2010.

[3] Smith, P., Mossialos, E., et al. (2008). Performance measurement for health system improvement: experiences, challenges and prospects. *World Health Organization.*

[4] Rego, M. (2010). O Balanced Scorecard como ferramenta de suporte para o alinhamento de objectivos estratégicos organizacionais - um estudo de caso no CENPES. In *VI Congresso Nacional de Excelência em Gestão: Energia, Inovação, Tecnologia e Complexidade para a Gestão Sustentável*, Niterói, Brasil.

[5] Silva, E., Lima, E., et al. (2009). Análise da produção académica brasileira em medição de desempenho e modelos de gestão. *Revista Gestão Industrial*, 5(3):198-218.

[6] Grugoroudis, E., Orfanoudaki, E., et al. (2011). Strategic performance measurement in a healthcare organisation: A multiple criteria approach based on balanced scorecard. *The Internation Journal of Management Science*, 40(1):104–119.

[7] Bourne, M., Mills, J., et al. (2000). Designing, implementing and updating performance measurement systems. *International Journal of Operations & Production Management*, 20(7):754–771.

[8] Santos, S., Belton, V., et al. (2001). Adding Value to Performance Measurement by Using Systems Dynamics and Multicriteria Analysis. *Management Science: Theory, Method & Practice*, 22(11):1246–1272.

[9] Berrah, L. and Cliville, V.(2007). Towards an aggregation performance measurement system model in a supply chain context. *Computers in Industry*, 58(7):709–719.

[10] Berrah, L., Cliville, V., et al.(2007). Quantitative expression and aggregation of performance measurements based on the MACBETH multi-criteria method. *International Journal of Production Economics*, 105(1):171–189.

[11] Cliville, V., Mauris, G., et al. (2006). A Quantified Industrial Performance Measurement System Based on a Choquet Fuzzy Integral. In *IEEE International Conference on Fuzzy Systems*, Vancouver.

[12] Bana e Costa, C. A. and Beinat, E. (2005). Model-structuring in public decision-aiding. *Working Paper LSE OR 05.79*, London School of Economics.

[13] Marichal, J. -L. (2000). An axiomatic approach of the discrete Choquet integral as a tool to aggregate interacting criteria. In *IEEE Transactions on Fuzzy Systems*, 8(6):800–807.

[14] Bana e Costa, C. A., de Corte, J. -M., et al.(2012). MACBETH. *International Journal of Information Technology & Decision Making*, 11(2):359–387.

[15] Bana e Costa, C. A., Ensslin, L., et al. (1999). Decision Support Systems in action: Integrated application in a multicriteria decision aid process. *European Journal of Operational Research*, 113(2):315–335.

[16] Bana e Costa, C. A., Corrêa, E. C., et al. (2002). Facilitating bid evaluation in public call for tenders: a socio-technical approach. *The International Journal of Management Science*, 30(3):227–242.

[17] Labreuche, C. and Grabisch, M.(2003). The Choquet integral for the aggregation of interval scales in multicriteria decision making. *Fuzzy Sets and Systems*, 137:11–26.

[18] Bana e Costa, C. A., Carnero, M. C., et al.(2012). A Multi-Criteria Model for Auditing a Predictive Maintenance Programme. *European Journal of Operational Research*, 217(2):381–393.

[19] Grabisch, M. and Labreuche, C.(2010). A decade of application of the Choquet and Sugeno integrals in multi-criteria decision aid. *Annals of Operations Research*, 175(1):247–290.

[20] Bana e Costa, C. A., Lourenço, J. C., et al. (2008). Development of reusable bid evaluation models for the Portuguese Electric Transmission Company. *Decision Analysis*, 5(1):22–42.

[21] Berrah, L., Mauris, G., et al. (2006). Industrial performance measurement: an approach based on the aggregation of unipolar or bipolar expression. *International Journal of Production Research*, 44(18-19):4145–4158.

[22] Bana e Costa, C. A., Costa-Lobo, M. L., et al.(2002). Multicriteria approach for strategic town planning: the case of Barcelos, in Aiding decisions with Multiple Criteria: Essays in Honour of Bernard Roy, D. Bouyssou, E. Jacquet-Lagrèze, P. Perny, R. Slowinsky, D. Vanderpooten, P. Vincke. *International Series in Operations Research & Management Science*, 44:429–456.

[23] Bana e Costa, C. A., Ferreira, J. A., et al. (1995). Avaliação multicritério de propostas: o caso de uma nova linha do metropolitano de Lisboa. *Revista de Transportes e Tecnologia*, VII, 14:31–65.

[24] Carnero, M. C. (2006). Predictive maintenance programme audit with MACBETH. Safety and Reliability for Managing Risk, Soares C. G. and Zio E. eds.. *Taylor & Francis Ltd*, 1:109–115.

[25] Barin, A., Martins, L. N., et al.(2012). Selection of renewable energy technologies in landfills: The M-MACBETH decision support. Fuelling the Future: Advances in Science and Technologies for Energy Generation, Transmission and Storage. Mendez-Vilas A. eds.. *Universal-Publishers*, 98–102.

[26] Choquet, G. (1953). Theory of capacities. *Annales de l'Institut Fourier*, 5:131–295.

[27] Grabisch, M. (2004). The Choquet integral as a linear interpolator. In *10th Int. Conf. on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, Perugia, Italy:373–378.

[28] Goodwin, P. and Wright, G.(1991). Decision Analysis for Management Judgement. *Jonh Wiley & Sons*.

On numerical testing of the regularity of Semidefinite problems

Eloísa Macedo*

* *Center for Research & Development in Mathematics and Applications (CIDMA), Department of Mathematics, University of Aveiro, Portugal*
*E-mail:*macedo@ua.pt

Abstract

This paper is devoted to study regularity of Semidefinite Programming (SDP) problems. Current methods for SDP rely on assumptions of regularity such as constraint qualifications and well-posedness. Absence of regularity may compromise characterization of optimality and algorithms may present numerical difficulties. Prior that solving problems, one should evaluate the expected efficiency of algorithms. Therefore, it is important to have simple procedures that verify regularity. Here we use an algorithm to test regularity of linear SDP problems in terms of Slater's condition. We present numerical tests using problems from SDPLIB and compare our results with those from others available in literature.

Key words: Constraint qualifications, optimality conditions, regularity, semi-infinite programming, semidefinite programming, well-posedness.

1 Introduction

Semidefinite Programming (SDP) deals with problems of minimization of a linear objective function subject to constraints in the form of linear matrix inequalities and can be considered as a generalization of Linear Programming (LP), where real matrices are used instead of real variables. SDP is an active area of Optimization and has various applications in control and approximation theory, sensor network localization, principal component analysis, etc. (see [1, 18] for further details).

The methods developed to solve SDP problems, the duality theory, and most of optimality conditions are based on certain assumptions of regularity [6, 12, 18]. In the absence of regularity, characterization of optimality of feasible solutions may fail. With respect to algorithms, the regularity of a problem is a condition that guarantees their stability and efficiency [6].

In the literature, different concepts are being associated to the notion of regularity. Usually, an optimization problem is considered to be regular if certain constraint qualification (CQ) is satisfied ([9]). The most efficient optimality conditions are formulated under the fulfilment of some CQ. The regularity conditions play also an important role in deriving duality relations, sensitivity/stability analysis and convergence of computational methods [11]. One of the most stronger regularity conditions is the Slater CQ (see [4, 12]) that consists in the nonemptiness of the interior of the feasible set.

Another kind of regularity in Optimization and Numerical Methods is known as well-posedness of a problem. There exist different definitions of well-posedness. According to [5, 13], an optimization problem is well-posed in the sense of Hadamard if it has a unique solution that depends continuously on data. According to [5, 16], a problem is well-posed in the sense of Tikhonov if it has a unique solution that is stable, meaning that small perturbations of data cause small variations on solution.

In [13], it is shown that under the Slater CQ, Hadamard's well-posedness is equivalent to that of Tikhonov. In [13], other definitions of well-posedness are presented as well (Levitin-Polyak well-posedness and strong well-posedness). It is noticed, in particular, that "in finite dimensions uniqueness of the solution to a convex minimizing problem ... is enough to guarantee its Tikhonov well-posedness ... This is no longer valid in infinite dimensions ...". Theoretical study of well-posedness of certain optimization problems is a rather difficult issue. In practice, well-posedness is often verified in terms of convergence of certain minimizing sequences of some optimization problems, that is not always easy to verify.

A problem that is not well-posed is called ill-posed. Ill-posed problems are quite common in applications and, according to [10], the ill-posedness may occur due to the lack of precise mathematical formulations.

In [7, 10], a practical characterization of well-posedness of SDP problems is proposed and it is based on a so called condition number defined by Renegar in [15]. The Renegar's condition number is defined as a scale-invariant reciprocal of a problem instance to be infeasible. A SDP problem is considered to be well-posed if its Renegar condition number is finite, and ill-posed, otherwise. In [10], the calculus of this condition number is connected with upper bounds for the optimal values of SDP problems. The approach proposed in [10] for characterization of regularity of a SDP problem is constructive and is based on obtaining rigorous bounds and also error bounds for the optimal values, by properly postprocessing the output of a SDP solver. The main feature of this approach is that computation of the rigorous bounds takes into account all rounding errors and the possible small errors presented in the input data.

There exist some studies dedicated to interrelation between regularity and well-posedness of optimization problems. In [13] different notions of well-posedness of general convex problems are studied and compared; in [8, 10, 17] the relationship between well-posedness and regularity in the sense of Slater is considered for SDP problems.

In this paper, we apply a simple algorithmic procedure suggested in [14] to verify the Slater CQ for problems from SDPLIB, a linear SDP database [3]. We will compare the obtained results with tests of well-posedness described in [7, 10].

The paper is structured as follows. The basic notions are presented in Section 2. The study of different notions of regularity of linear SDP problems and the description of the algorithm to test the Slater CQ for these problems are carried out in Section 3. The numerical results are presented in Section 4. The final section is devoted to the conclusions of our work and future research.

2 Linear Semidefinite Programming Problem and its Semi-Infinite presentation

In what follows, we denote by $\mathcal{S}(s) \subset \mathbb{R}^{s\times s}$, $s \in \mathbb{N}$ the subspace of $s \times s$ real symmetric matrices. The set $\mathcal{S}(s)$ can be considered as a vector space with the trace inner product defined by

$$tr(AB) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}b_{ji}$$

for $A, B \in \mathcal{S}(s)$. A matrix $A \in \mathcal{S}(s)$ is positive semidefinite if $x^T Ax \geq 0, \forall x \in \mathbb{R}^s$. Given $A \in \mathcal{S}(s)$, to denote that A is positive (negative) semidefinite, we write: $A \succeq 0$ ($A \preceq 0$). Let $\mathcal{P}(s) \subset \mathcal{S}(s)$ be the cone of positive semidefinite symmetric $s \times s$ matrices.

Consider the following linear SDP problem:

$$\min c^T x, \quad \text{s. t. } \mathcal{A}(x) \preceq 0, \tag{2.1}$$

where $x \in \mathbb{R}^n$, $\mathcal{A}(x) := \sum_{i=1}^{n} A_i x_i + A_0$, $A_i \in \mathcal{S}(s)$, $i = 0, 1, ..., n$. Problem (2.1) is a convex problem and its (convex) feasible set is $\mathcal{X} = \{x \in \mathbb{R}^n : \mathcal{A}(x) \preceq 0\}$. We will refer to this problem as primal SDP problem.

The Lagrangian dual problem to problem (2.1) is given by

$$\max\ tr\,(A_0 Z)\,, \quad \text{s. t. } -tr\,(A_i Z) = c_i, \forall i = 1, \dots, n, \quad Z \succeq 0, \tag{2.2}$$

where $Z \in \mathcal{P}(s)$. The feasible set of problem (2.2) is $\mathcal{Z} = \{Z \in \mathcal{P}(s) : -tr\,(A_i Z) = c_i, i = 1, ..., n\}$.

Notice here that some authors consider that the primal SDP problem has the form (2.2), and the dual problem has the form (2.1). This is not an issue, since it is possible to transform the problem in the form (2.1) to the form (2.2) and vice-versa (see [18]).

The duality results in SDP are more subtle than in Linear Programming (LP). The following property of LP problems still holds for SDP, inducing a lower bound on the value of the primal problem:

Theorem 1. *[Weak Duality] Given a pair of primal and dual feasible solutions $x \in \mathcal{X}, Z \in \mathcal{Z}$ of SDP problems* (2.1) *and* (2.2)*, the inequality $c^T x \geq tr(A_0 Z)$ always holds.*

Definition 1. *Denote by p^* the optimal value of the objective function of the primal SDP problem* (2.1) *and by d^* the optimal value of the objective function of the dual problem* (2.2)*. The difference $p^* - d^*$ is called duality gap.*

Unlike LP, a nonzero duality gap can occur in SDP and to guarantee strong duality some additional assumptions have to be made [1, 17].

Semi-Infinite Programming (SIP) studies optimization problems in the form

$$\min c^T x, \quad \text{s. t. } f(x,t) \le 0, t \in T,$$

where the index set T is some subset of $\mathbb{R}^s$ containing an infinite number of elements.

Consider the linear SDP problem (2.1). It is easy to see that this problem is equivalent to the following convex semi-infinite problem ([14]):

$$\min c^T x, \quad \text{s. t. } l^T \mathcal{A}(x) l \le 0, \forall l \in L := \{l \in \mathbb{R}^s : \|l\| = 1\}. \tag{2.3}$$

The feasible set of problem (2.3) is $\left\{x \in \mathbb{R}^n : l^T \mathcal{A}(x) l \le 0, \forall l \in L\right\}$ and it coincides with the feasible set of problem (2.1).

In [14], a new approach to optimality conditions for convex SIP and linear SDP was suggested. This approach is based on the notions of immobile indices for SIP problems and subspace of immobile indices for SDP.

Definition 2. *Given a convex SIP problem* (2.3)*, an index $l^* \in L$ is called immobile if $l^{*T} \mathcal{A}(x) l^* = 0, \forall x \in \mathcal{X}$.*

It is proved in [14] that, given a pair of equivalent problems (2.1) and (2.3), the set of immobile indices $L^* = \left\{l \in L : l^T \mathcal{A}(x) l = 0, \forall x \in \mathcal{X}\right\}$ of problem (2.3) can be presented in the form $L^* = L \cap \mathcal{M}$, where $\mathcal{M}$ is a subspace of $\mathbb{R}^s$ defined by

$$\mathcal{M} := \left\{l \in \mathbb{R}^s : l^T \mathcal{A}(x) l = 0, \forall x \in \mathcal{X}\right\} = \{l \in \mathbb{R}^s : \mathcal{A}(x) l = 0, \forall x \in \mathcal{X}\}, \tag{2.4}$$

and is called subspace of immobile indices of the SDP problem (2.1).

3 Regularity of SDP Problems

3.1 Constraint Qualifications

Constraint qualifications are essential for deriving primal-dual characterization of solutions of optimization problems, and play an important role in duality theory, sensitivity and stability analysis, and in the convergence properties of computational algorithms ([11]).

The Slater CQ is one of the most widely known CQ in finite and infinite optimization that guarantees the vanishing of the dualiaty gap, therefore many authors assume in their studies that this condition holds (see [4, 8, 12, 17]).

Definition 3. *The constraints of problem* (2.1) *satisfy the Slater CQ if the feasible set $\mathcal{X}$ has a nonempty interior, i.e., $\exists\ \bar{x} \in \mathbb{R}^n : \mathcal{A}(\bar{x}) \prec 0$.*

Here, $A \prec 0$ ($A \succ 0$) denotes that matrix $A \in S(s)$ is negative (positive) definite. The Slater CQ is sometimes called strict feasibility [17] or Slater regularity condition [12].

The analogous definition can be introduced for the dual SDP problem.

Definition 4. *The constraints of the dual SDP problem* (2.2) *satisfy the Slater CQ if there exists a matrix $Z \in \mathcal{P}(s)$, such that $-tr\,(A_i Z) = c_i, \forall i = 1, \ldots, n$ and $Z \succ 0$.*

In this paper, we will consider a SDP problem (2.1) to be regular if its constraints satisfy the Slater CQ and nonregular otherwise.

If it is assumed that problem (2.1) is regular, then the optimal values of problems (2.1) and (2.2) coincide and the following property is satisfied ([4, 17]):

Theorem 2. *[Strong Duality] Under the Slater CQ for the linear SDP problem* (2.1)*, if the primal optimal value is finite, then the duality gap vanishes and the (dual) optimal value of* (2.2) *is attained.*

The first order necessary and sufficient optimality conditions for regular linear SDP can be formulated in the form of the following theorem from ([2]):

Theorem 3. *If problem* (2.1) *satisfies the Slater CQ, then $x^* \in \mathcal{X}$ is an optimal solution if and only if there exists a matrix $Z^* \in \mathcal{P}(s)$ such that*

$$tr\,(Z^* A_i) + c_i = 0, i = 1, ..., n, \ \ tr\,(Z^* \mathcal{A}(x^*)) = 0. \tag{3.1}$$

In the absence of the Slater CQ, a duality gap in SDP can exist and conditions (3.1) (also called complementary conditions) may fail ([4, 14, 17]). Hence, the complete characterization of optimality of feasible solutions (either primal or dual) may fail (see the examples provided in [1, 14, 17]). Since many algorithms are based on solving systems of type (3.1), the failure of the strong duality can result in numerical difficulties, and thus, it is important to know in advance if the problem is regular. Notice also that unlike LP, the primal and dual SDP problems do not necessarily satisfy or not the Slater CQ simultaneously ([17]).

For the SIP problem in form (2.3), the Slater CQ is defined as follows.

Definition 5. *The SIP problem* (2.3) *satisfies the Slater CQ if there exists a feasible point* $\bar{x} \in \mathbb{R}^n$ *such that the inequalities* $l^T \mathcal{A}(\bar{x}) l < 0$ *hold, for all indices* $l \in L$.

It is easy to verify that the equivalent SDP and SIP problems (2.1) and (2.3) satisfy or not the Slater CQ simultaneously [14]. The following propositions are proved in [14].

Proposition 1. *The convex SIP problem* (2.3) *satisfies the Slater CQ if and only if the set* L^* *is empty.*

Proposition 2. *The SDP problem* (2.1) *satisfies the Slater CQ if and only if the set of immobile indices in the corresponding SIP problem* (2.3) *is empty.*

From Proposition 1, it follows that problem (2.1) is regular if and only if the subspace of immobile indices $\mathcal{M}$ defined in (2.4) is null, i.e., $\mathcal{M} = \{0\}$. Notice that the dimension s^* of the subspace $\mathcal{M}$ can be considered as the irregularity degree of problem (2.1) and

- if $s^* = 0$, then the problem is regular (satisfies the Slater CQ);
- if $s^* = 1$, then the problem is nonregular, with minimal irregularity degree;
- if $s^* = s$, then the problem is nonregular, with maximal irregularity degree.

In [14], it is shown that the subspace of immobile indices plays an important role in characterization of optimality of SDP problems and a new CQ-free optimality criterion is formulated, based on the explicit determination of the subspace $\mathcal{M}$ of immobile indices. A constructive algorithm (the DIIS algorithm) that finds a basis of the subspace $\mathcal{M}$ is described and justified in [14].

In this paper, we will use the DIIS algorithm just to check if a given SDP problem in the form (2.1) is regular.

3.1.1 Algorithm DIIS

Consider a linear SDP problem (2.1) and suppose that its feasible set is nonempty. The DIIS Algorithm proposed in [14] constructs a basis $M = (m_i, i = 1, ..., s^*)$ of the subspace of immobile indices $\mathcal{M}$. At the k-th iteration, I^k denotes some auxiliary set of indices and M^k denotes an auxiliary set of vectors. Suppose that $s > 1$, with $s \in \mathbb{N}$.

The brief description of the algorithm is as follows:

DIIS Algorithm

`input:` $s \times s$ matrices $A_j, j = 0, 1, ..., n$. Set $k := 1$, $I^1 := \emptyset$, $M^1 := \emptyset$

`repeat`

Given k, I^k , M^k:

compute $p_k := s - \left|I^k\right|$

solve the quadratic system:

$$\begin{cases} \sum\limits_{i=1}^{p_k} l_i^T A_j l_i + \sum_{i \in I^k} \gamma_i^T A_j m_i = 0, \; j = 0, 1, \ldots, n, \\ \sum\limits_{i=1}^{p_k} \|l_i\|^2 = 1, \\ l_i^T m_j = 0 \;, \; j \in I^k, \; i = 1, \ldots, \; p_k, \end{cases} \tag{3.2}$$

where $l_i \in \mathbb{R}^s, i = 1, ..., p_k$ and $\gamma_i \in \mathbb{R}^s, i \in I^k$

`if` system (3.2) does not have a solution, `then` stop and return the current M^k.

`else` given the solution $\left\{l_i \in \mathbb{R}^s, i = 1, \ldots, p_k, \gamma_i \in \mathbb{R}^s, i \in I^k\right\}$ of (3.2):

construct the maximal subset of linearly independent vectors

$\{m_1, \ldots, m_{s_k}\} \subset \{l_1, \ldots, l_{p_k}\}$

update: $\triangle I^k := \left\{\left|I^k\right| + 1, \ldots, \left|I^k\right| + s_k\right\}$

$M^{k+1} := M^k \cup \left\{m_j, \; j \in \triangle I^k\right\}$

$I^{k+1} := I^k \cup \triangle I^k$.

`do` $k := k + 1$

The outputs of the algorithm are:

- if the Slater CQ is satisfied: the algorithm stops at the first iteration, $k = 1$, $\mathcal{M} = \{0\}$ and $dim(\mathcal{M}) = 0$;
- if the Slater CQ is violated: the algorithm returns a basis $M = M^k$, such that $rank(M) = dim(\mathcal{M}) = s^* > 0$.

In [14], it is proved that the DIIS algorithm constructs a basis of the vector subspace $\mathcal{M}$ in a finite number of iterations.

It is easy to see that the algorithm's implementation is simple and the main numerical procedure on each iteration is solving the quadratic system (3.2).

3.2 Well-posedness in SDP

In [7] and [10], two constructive approaches to the classification of well-posed SDP problems are proposed, both based on the concept of well-posedness in the sense defined by Renegar in [15]. The Renegar condition number C of a problem's instance is defined as a scale-invariant reciprocal of the distance to infeasibility (the smallest data perturbation that results in either primal or dual infeasibility). If the distance to infeasibility is close to zero, then $C = \infty$ and the problem is said to be ill-posed; otherwise, it is well-posed.

The approach described in [7] is based on the estimation of lower and upper bounds of the Renegar's condition number C of a SDP problem. The distance to primal infeasibility is obtained by solving several auxiliary SDP problems of compatible size to the original SDP problem and the distance to dual infeasibility is found by solving one single SDP auxiliary problem. According to Proposition 3 in [7], the estimation of the norm of data can be done with the help of its upper and lower bounds using straightforward matrix norms and maximum eigenvalue computations. Notice that it is necessary to choose adequate norms for computing the distances to primal and dual infeasibility. The SDP problem is ill-posed if C approaches to infinity.

In [10], the characterization of well-posedness of SDP problems is based on the calculus of rigorous lower and upper bounds of their optimal values. It is shown that for the ill-posed problems (in the sense of the infinite condition number), the rigorous upper bound $\bar{p}^*$ of the primal objective function is infinite. In [10], an algorithm for computing this upper bound is described. On its iterations, some auxiliary perturbed "midpoint" SDP problems are solved using a SDP solver and special interval matrices are constructed on the basis of their solutions. The constructed interval matrices must contain a primal feasible solution of the perturbed "midpoint" problem and satisfy the conditions (4.1) and (4.2) of Theorem 4.1 from [10]. If such interval matrix can be computed, then the optimal value of the primal objective function of the SDP problem is bounded from above by $\bar{p}^*$, which is the primal objective function value considering the obtained interval matrix. Besides requiring a SDP solver for computing the approximate solutions of the perturbed problems, the approach proposed in [10] also needs verified solvers for interval linear systems and eigenvalue problems.

Therefore, we can conclude that technically, testing of regularity of SDP problems with the help of the DIIS algorithm from [14] is much easier than testing their well-posedness using the methods from [7] and [10].

3.3 Relations between regularity and well-posedness in SDP

Nevertheless the definitions of regularity and well-posedness of SDP problems introduced above are different, there exist a deep connection between them. According to [17], the lack of regularity implies ill-posedness of the problem.

The following lemma can be easily proved.

Lemma 1. *If the linear SDP problem in the form* (2.1) *is not regular, then it is ill-posed.*

Indeed, since the Slater CQ is not satisfied, all the feasible solutions of the problem lie on the boundary of the feasible set. Therefore, any small perturbations may lead to the loss of feasibility, meaning that the problem is ill-posed.

The reciprocal is not true. The following example shows that there exist regular problems that are ill-posed.

Example 1. *Consider the primal SDP problem*

$$\begin{array}{ll} \min & x_1 - x_2 - x_3 \\ \text{s.t.} & \begin{bmatrix} 1 & -1 & 0 \\ -1 & x_2 & 0 \\ 0 & 0 & x_3 \end{bmatrix} \succeq 0. \end{array} \tag{3.3}$$

This problem can be easily written in form (2.1)*:*

$$\begin{array}{ll} \min & x_1 - x_2 - x_3 \\ \text{s.t.} & \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} x_1 + \begin{bmatrix} 0 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{bmatrix} x_2 + \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -1 \end{bmatrix} x_3 + \begin{bmatrix} -1 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \preceq 0. \end{array}$$

The dual problem to (3.3) *has the form*

$$\begin{array}{ll} \max & y_1 + y_2 \\ \text{s.t.} & \begin{bmatrix} -y_2 & \frac{1+y_1}{2} & -y_3 \\ \frac{1+y_1}{2} & -1 & -y_4 \\ -y_3 & -y_4 & -1 \end{bmatrix} \succeq 0. \end{array} \tag{3.4}$$

The constraints of the primal problem satisfy the Slater's CQ, since there exists a strictly feasible solution: e.g., $X = \begin{bmatrix} -1 & 1 & 0 \\ 1 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ *and* $X \prec 0$.

However, problem (3.3) *is not well-posed, since its dual problem is infeasible, the distance to dual infeasibility is zero and the Renegar condition number is infinite.*

4 Regularity Tests for SDP Problems

4.1 Test Problems for SDP

For the numerical tests we have used problems from the literature and from the SDPLIB suite, a collection of 92 Linear SDP test problems, provided by Brian Borchers ([3]). This SDP data base contains problems ranging in size from 6 variables and 13 constraints up to 7000 variables and 7000 constraints. The problems are drawn from a variety of applications, such as truss topology design, control systems engineering and relaxations of combinatorial optimization problems.

4.2 Regularity Tests

To test the regularity of linear SDP problems we used the DIIS Algorithm that was implemented in Matlab 7.10.0.499 (R2010a) on a computer with an Intel Core i7-2630QM, 2.0GHz, with Windows 7 64 bits and 6 Gb RAM.

We have chosen 26 test problems from SDPLIB that were also tested in [7] and [10], in terms of their well-posedness. For these problems we checked regularity using the DIIS algorithm.

The results are displayed in Table 1. The first column of the table contains the instance's name used in SDPLIB data base; the next two columns refer to the number of variables, n, and the dimension of the constraints matrices, s. The next column represents the results of the regularity tests that find the dimension of the immobile index subspace, s^*. If $s^* = 0$, then the problem is regular. Column 5 contains the lower and upper bounds of the condition number C reported in [7] and the last column presents the upper bound for the primal objective function from [10].

From Table 1 we see that 20 from 26 tested problems are regular, i.e., their constraints satisfy the Slater CQ. Moreover, the tests provide valuable information about the irregularity degree for a nonregular SDP problem. Notice that for the tested problems, all the nonregular problems in Table 1 have maximal irregularity degree.

Tables 2 and 3 compare the results of testing SDP problems in terms of their regularity and well-posedness.

In Table 2, the lines correspond to well-posed and ill-posed problems according to the test from [7] and the columns correspond to regular and nonregular problems. On the intersection we have the number of problems that satisfy both corresponding conditions. Table 3 is constructed in a similar way, but

Table 1: Numerical results using the DIIS Algorithm to test regularity (the SDP problem satisfies the Slater CQ if $s^* = 0$), the lower and upper bounds of the Renegar condition number from [7] and the upper bound of the optimal value from [10] (if C or $\bar{p}^*$ is finite, then the problem is well-posed).

Problem	n	s	s^*	C		$\bar{p}^*$
				lower bound	upper bound	
control1	21	15	0	8.3×10^5	1.8×10^6	-1.7782×10^1
control2	66	30	0	3.9×10^6	1.3×10^7	-8.2909×10^0
control3	136	45	0	2.0×10^6	1.2×10^7	-1.3615×10^1
hinf1	13	14	0	∞	∞	∞
hinf2	13	16	0	3.5×10^5	5.6×10^5	-7.1598×10^0
hinf3	13	16	0	∞	∞	∞
hinf4	13	16	0	∞	∞	∞
hinf5	13	16	16	∞	∞	∞
hinf6	13	16	16	∞	∞	∞
hinf7	13	16	16	∞	∞	∞
hinf8	13	16	0	∞	∞	∞
hinf9	13	16	16	2.0×10^7	3.6×10^7	∞
hinf10	21	18	0	∞	∞	∞
hinf11	31	22	0	∞	∞	∞
hinf12	43	24	0	∞	∞	∞
hinf13	57	30	30	∞	∞	∞
hinf14	73	34	0	∞	∞	∞
hinf15	91	37	37	∞	∞	∞
qap5	136	26	0	∞	∞	∞
qap6	229	37	0	∞	∞	∞
qap7	358	50	0	∞	∞	∞
qap8	529	65	0	∞	∞	∞
theta1	104	50	0	2.0×10^2	2.1×10^2	-2.3000×10^1
truss1	6	13	0	2.2×10^2	3.0×10^2	9.0000×10^0
truss3	27	31	0	7.4×10^2	1.9×10^3	9.1100×10^0
truss4	12	19	0	3.6×10^2	7.7×10^2	9.0100×10^0

Table 2: Regularity and well-posedness according to [7].

		Regularity (Slater CQ)	
		Regular	*Nonregular*
Classification	*well-posed*	8	1
	ill-posed	12	5

Table 3: Regularity and well-posedness according to [10].

		Regularity (Slater CQ)	
		Regular	*Nonregular*
Classification	*well-posed*	8	0
	ill-posed	12	6

the lines correspond to the number of the well-posed and ill-posed problems classified on the basis of the experiments in [10]. From Table 2 we can see that 13 from the tested problems are regular and well-posed or nonregular and ill-posed, simultaneously, and 12 of the ill-posed problems are regular. The only exception is problem *hinf9* that is nonregular and well-posed according to [7]. This contradiction to Lemma 1 can be explained by the fact that the numerical procedures are based on approximated calculus and may be not precise. Comparing our regularity results with those from [10] (w.r.t. ill-posedness), regarding Table 3 we conclude that for 14 problems these results coincide, i.e., the problems are regular and well-posed, or nonregular and ill-posed, simultaneously.

Notice that the numerical results of well-posedness obtained in [7] and in [10] do not coincide: problem *hinf9* is well-posed according to [7] and ill-posed according to [10]. This can be connected with the fact that nevertheless the condition number C is finite, it is rather big and the problem is close to be ill-posed, and it may also be due to the tests were performed in nonexact arithmetic and/or with different numerical procedures.

Finally, notice that in [10], it is reported that problem *hinf8* is well-posed, although the results presented in [10] (and also in [7]) have shown that this problem is ill-posed. Our numerical tests show that this problem is regular.

Therefore, the numerical experiences have showed that the DIIS algorithm can be efficiently used to study the regularity of SDP problems. The comparison of these tests with those from [7] and [10] confirm the conclusions about relation between regularity and well-posedness in SDP.

4.3 Conclusions and Future Work

The DIIS Algorithm permits to verify easily if a given SDP problem in the form (2.1) satisfies the Slater CQ. The numerical tests show that the DIIS algorithm is an efficient procedure to check the regularity of small to medium-scale SDP problems. For problems of bigger dimension the program ran out of memory. The DIIS algorithm can be used to check if a given SDP problem in form (2.1) is regular or not in a single iteration. The main advantage of using this numerical procedure is that one does not need to solve any SDP problem: the DIIS algorithm only deals with a quadratic system (3.2). To solve this system, we used in our tests the Levenberg-Marquardt method that is implemented in the routine `fsolve` from Matlab.

In our procedure, we have not checked the feasibility of the SDP problems: we worked under the assumption that their feasible sets were not empty.

To permit a more universal and precise use of the DIIS algorithm, some improvements should be made in the future. Thus, the algorithm can be modified to handle large-scale SDP problems. The procedure of solving the quadratic system (3.2) should be as much exact as possible, so it is reasonable to develop here specific algorithms that precisely verify the situations when the system is inconsistent. It is also important to implement the procedure that verifies feasibility of the SDP problems. Finally, it is reasonable to implement the DIIS algorithm in other programming language than Matlab (e.g., C++), since the implementation in Matlab can cause some loss of efficiency.

Notice that the tests of well-posedness of SDP problems using the procedures described in [7] and [10] present much more difficulties compared with the numerical test presented in this paper. Thus, using the method described in [7], one has to estimate the Renegar condition number that depends on the computation of three quantities that can be more or less numerically hard, depending on the choice of norms. To compute estimations of distance to primal and dual infeasibility, one has to solve several auxiliary SDP problems whose structure and size are compatible with the original SDP primal and dual problems. This may be computationally hard and not precise, since in nonregular cases, the auxiliary problems are ill-posed and the solutions found by the numerical methods may be not correct. The procedure in [10] contains parameters and to find them an auxiliary perturbed SDP problem must be solved, that turns into a difficult problem in the case of nonregular problems. To compute the rigorous upper bound, it is required to find certain interval matrices satisfying conditions of Theorem 4.1. Specially adapted solvers for interval linear systems and eigenvalue problems as well as a SDP solver for computing approximate solutions for the perturbed "midpoint" problem are needed.

On the basis of the results of the numerical experiences we can make the following conclusions: it is important to introduce a unified treatment of regularity for SDP problems and to have numerical tools to verify regularity of problems and establish clear relationship between different notions of regularity.

In the future, we intend to provide more precise and extensive regularity tests with SDP and SIP problems, and compare them with the available results of well-posedness tests.

Acknowledgements

This work was supported by *FEDER* funds through *COMPETE*–Operational Programme Factors of Competitiveness (“Programa Operacional Factores de Competitividade”) and by Portuguese funds through the *Center for Research and Development in Mathematics and Applications* (University of Aveiro) and the Portuguese Foundation for Science and Technology (“FCT–Fundação para a Ciência e a Tecnologia”), within project PEst-C/MAT/UI4106/2011 with COMPETE number FCOMP-01-0124-FEDER-022690.

References

[1] Anjos, M.F. and Lasserre, J.B. (Eds.) (2012). Handbook of Semidefinite, Conic and Polynomial Optimization: Theory, Algorithms, Software and Applications. *International Series in Operational Research and Management Science*, 166, Springer.

[2] Bonnans, J. F. and Shapiro, A. (2000). Perturbation Analysis of Optimization Problems. *Springer-Verlag*, New-York.

[3] Borchers, B. (1999). SDPLIB 1.2, A library of Semidefinite Programming Test Problems. *Optimization Methods and Software*, 11(1):683–690.

[4] Cheung, Y., Schurr, S. and Wolkowicz, H. (2013). Preprocessing and Reduction for Degenerate Semidefinite Programs. Research Report CORR 2011-02, http://www.optimization-online.org/DB_FILE/2011/02/2929.pdf, Revised in January 2013.

[5] Dontchev, A. L., Zolezzi, T. (1993). Well-Posed Optimization Problems. *Lecture Notes in Mathematics, Springer-Verlag*, Berlin, 1543.

[6] Freund, R. M. (1995). Complexity of an Algorithm for Finding an Approximate Solution of a Semi-Definite Program, with no Regularity Condition. Technical Report OR 302-94, Op. Research Center, MIT, Revised in December 1995.

[7] Freund, R. M., Ordóñez, F. and Toh, K. C. (2007). Behavioral Measures and their Correlation with IPM Iteration Counts on Semi-Definite Programming Problems. *Mathematical Programming*, Springer-Verlag, New York, 109(2):445–475.

[8] Helmberg, C. (2000). Semidefinite Programming for Combinatorial Optimization. ZIB Report, Berlin.

[9] Hernández-Jiménez, B., Rojas-Medar, M.A., Osuna-Gómez, R., Beato-Moreno, A. (2009). Generalized convexity in non-regular programming problems with inequality-type constraints. Journal of Mathematical Analysis and Applications, 352(2):604–613.

[10] Jansson, C., Chaykin, D. and Keil, C. (2007). Rigorous Error Bounds for the Optimal Value in Semidefinite Programming. *SIAM Journal on Numerical Analysis archive*, 46(1):180–200.

[11] Klatte, D. (2009). First Order Constraint Qualifications. *Encyclopedia of Optimization*, Floudas, C. A. and Pardalos, P. M. (Eds.), Second Edition, Springer, 1055–1060.

[12] Klerk, E. de (2004). Aspects of Semidefinite Programming - Interior Point Algorithms and Selected Applications. *Applied Optimization*, 65, Kluwer.

[13] Konsulova, A. S. and Revalski,J. P. (1994). Constrained convex optimization problems - well-posedness and stability. *Numerical Functional Analysis and Optimization*, 15(7-8):889–907.

[14] Kostyukova, O. I. and Tchemisova, T.V. (2012). Optimality Criterion without Constraint Qualification for Linear Semidefinite Problems. Journal of Mathematical Sciences, 182(2):126–143.

[15] Renegar, J. (1994). Some Perturbation-Theory for Linear-Programming. *Mathematical Programming*, 65(1):73–91.

[16] Tikhonov, A. N. and Arsenin, V. Y. (1977). Solutions of ill-posed problems. *John Wiley and Sons.*

[17] Wolkowicz, H. (2009). Duality for semidefinite programming. *Encyclopedia of Optimization*, Floudas, C. A. and Pardalos, P. M. (Eds.), Second Edition, Springer, 811–814.

[18] Wolkowicz, H., Saigal, R. and Vandenberghe, L. (2000). Handbook of semidefinite programming: theory, algorithms, and applications. *Kluwer*.

Dynamic location problem with uncertainty: a branch&bound approach

Maria do Céu Marques*, Joana Matos Dias†

* *Instituto Superior de Engenharia de Coimbra, Instituto Politécnico de Coimbra, Portugal*
E-mail:`cmarques@isec.pt`

† *Faculdade de Economia e INESC–Coimbra, Universidade de Coimbra, Portugal*
E-mail:`joana@fe.uc.pt`

Abstract

We consider the dynamic uncapacitated facility location problem where uncertainty, regarding future potential facility locations, customers and costs is considered using scenarios. The objective is to minimize the expected total cost. We assume that once a facility is opened it stays open until the end of the planning horizon. Whilst assignment decisions can be scenario dependent, location decisions cannot. This problem contains the deterministic static and dynamic uncapacitated facility location problems as particular cases. We propose a branch&bound algorithm incorporating an efficient primal–dual heuristic to solve the problem. Computational results are discussed and compared with the results using CPLEX.

Keywords: dynamic location, uncertainty, scenarios, heuristics, branch&bound.

1 Introduction

The strategic nature of most location decision problems associated with the limited knowledge about problem parameters at the time of decision making, makes facility location problems under uncertainty an active area of research within the location research field. Models and solution methods that deal explicitly with uncertainties are even more complex than deterministic versions and with higher computational difficulties to achieve optimal or near–optimal solutions. As emphasized in the literature, the subject flourished during the last years and is far from being depleted [Snyder, 2006].

In this work we consider the dynamic uncapacitated facility location problem (DUFLP) where uncertainty is explicitly incorporated, represented by a finite and discrete set of future scenarios. In our model, fixed and assignment costs are scenario dependent, as well as the set of customers and the set of potential locations for facilities. We formulate the problem as an integer linear programming problem, that contains the static uncapacitated facility location problem (UFLP) and the DUFLP as particular problems (NP–hard problems [Cornuejols, 1990]). We also propose a branch&bound algorithm incorporating an efficient primal-dual heuristic to solve the problem to optimality.

The remainder of this paper is organized as follows. In the following section, the problem is described. In section 3 the solution method is summarized. In section 4 computational experiments with results are provided. Section 5 concludes this paper with some notes on future work.

2 Problem description

Consider a planning horizon represented by a discrete set of time periods $\mathcal{T} = \{1, ..., t, ..., T\}$. The *future* will be one of a finite set of possibilities, represented by a discrete set of *scenarios* $\mathcal{S} = \{1, ..., s, ..., S\}$, where each scenario characterizes the value of all uncertain elements. Suppose that each $s \in \mathcal{S}$ will occur with probability p^s such that $\sum_{s\in\mathcal{S}} p^s = 1$.

Let the set of potential facility sites be denoted by $J = \{1, ..., j, ..., M\}$ and the set of possible customer locations (or demand points) by $I = \{1, ..., i, ..., N\}$. In reality, these sets include all the potential facility locations and all the potential customers for all possible scenarios, despite the fact that for each scenario in particular possibly only a subset of potential locations and a subset of customers is considered. The reason for this is that we consider uncertainty associated not only with the fixed and variable costs, but also associated with the existence of customers and the future existence of potential locations. Let us define δ_{it}^s as equal to 1 if customer i has a demand that has to be fulfilled during period t for scenario s, and 0 otherwise. Then we have to guarantee that all customers such that $\delta_{it}^s = 1$ are assigned to an open facility, for all $(t, s) \in \mathcal{T} \times \mathcal{S}$.

In terms of costs, the model considers not only fixed costs (opening and operating), but also variable costs associated with the assignment of customers to the facilities. For $(j, t, s) \in J \times \mathcal{T} \times \mathcal{S}$, let f_{jt}^s be the fixed cost of establishing (opening) facility j at the beginning of period t plus the operating and subsequent costs in that period, under scenario s; for $(i, j, t, s) \in I \times J \times \mathcal{T} \times \mathcal{S}$, c_{ijt}^s represents the assignment cost of customer i to facility j in period t and under scenario s. If it is not possible to open facility j at the beginning of time period t, under scenario s, then the corresponding fixed cost will be considered equal to $+\infty$. We assume that once a facility is opened, it stays open until the end of the planning horizon.

The decisions to be made are where and when to locate new facilities, and how to assign the existing customers to open facilities over the whole planning horizon and under each scenario. Thus, we define the following binary decision variables: x_{jt} equals 1 if facility j is opened at the beginning of period t, and 0 otherwise; y_{ijt}^s equals 1 if customer i is assigned to facility j in period t and under scenario s, and 0 otherwise. As a matter of fact, assignment decisions are considered to be taken a period at a time, so they can be changed according to the scenario that occurred. Location decisions are hard to revert, so we have to live with the decision taken whatever the scenario that came to occur. Our aim is to make the best location decisions, considering the uncertainty associated with the future. Considering the minimization of expected total costs, we can formulate the problem as follows:

$$\min \quad \sum_{t\in\mathcal{T}}\sum_{j\in J}\sum_{s\in\mathcal{S}} p^s f_{jt}^s\, x_{jt} + \sum_{s\in\mathcal{S}}\sum_{t\in\mathcal{T}}\sum_{i\in I}\sum_{j\in J} p^s c_{ijt}^s\, y_{ijt}^s \tag{2.1}$$

subject to

$$\sum_{j\in J} y_{ijt}^s = \delta_{it}^s \qquad \forall i \in I, t \in \mathcal{T}, s \in \mathcal{S} \tag{2.2}$$

$$\sum_{\tau=1}^{t} x_{j\tau} - y_{ijt}^s \geq 0 \qquad \forall i \in I, j \in J, t \in \mathcal{T}, s \in \mathcal{S} \tag{2.3}$$

$$\sum_{t\in\mathcal{T}} (-x_{jt}) \geq -1 \qquad \forall j \in J \tag{2.4}$$

$$x_{jt} \in \{0, 1\} \qquad \forall j \in J, t \in \mathcal{T} \tag{2.5}$$

$$y_{ijt}^s \in \{0, 1\} \qquad \forall i \in I, j \in J, t \in \mathcal{T}, s \in \mathcal{S} \tag{2.6}$$

The objective function (2.1) minimizes the expected total costs (fixed plus variable costs). Constraints (2.2) require that, under each scenario and in every time period, an existing customer is assigned to exactly one facility. Constraints (2.3) impose that an existing customer can only be assigned to open facilities. A customer can be assigned to different facilities at different time periods and different scenarios. Constraints (2.4) ensure that each facility is opened at most once during the time horizon (located at the same site in all scenarios). Finally, (2.5)–(2.6) restrict the decision variables to be binary.

The above formulation contains the UFLP ($|\mathcal{T}| = |\mathcal{S}| = 1$) and the DUFLP ($|\mathcal{T}| > 1, |\mathcal{S}| = 1$) as particular problems, and has $|J|\,|\mathcal{T}| + |J|\,|I|\,|\mathcal{T}|\,|\mathcal{S}|$ binary variables and $|I|\,|\mathcal{T}|\,|\mathcal{S}| + |J|\,|I|\,|\mathcal{T}|\,|\mathcal{S}| + |J|$ restrictions (not counting the zero-one constraints). Even for moderate dimensions of these sets, (2.1)–(2.6) becomes a quite large integer linear program.

3 Solution method

In this paper we propose a branch&bound algorithm capable of calculating the optimal solution of the problem. Instead of solving to optimality relaxed versions of the original problems in each node of the branch&bound tree, we decided to use a very efficient dual-based heuristic to solve each problem. This dual-based heuristic is inspired on the approaches developed in [Erlenkotter, 1978] and [Van Roy, 1982], designed for the static and dynamic versions of the UFLP, respectively. A formal description of this heuristic, along with dual problem formulations and complementary slackness conditions, is given in [Marques, 2011]. The heuristic's procedures (dual ascent, primal and adjustment procedures) are designed to reduce progressively the duality gap between dual and primal objective function values. We can summarize its behavior as follows. The ascent procedure starts with a dual feasible solution and tries to increase iteratively the values of dual variables associated with restrictions (2.2). This will lead to an increase of the dual objective function value and, simultaneously, to the decrease of some slacks' values. It stops when all dual variables are blocked from increasing by at least one slack, and thus no further improvements of the dual objective function value are possible. The output of this procedure is a dual feasible solution and an associated set of candidate facility locations defined by the slacks that are equal to zero. A corresponding candidate primal feasible solution will be constructed within the primal procedure. The facilities that are opened first, belonging to the set of candidate facilities that can be opened without violating complementary slackness conditions, are the ones that at a given time t should be assigned to a given customer i under some scenario s, the so–called *essential* facilities. Other facilities are only opened if strictly necessary, that is if exists a customer that cannot be assigned to an *essential* facility. Finally, as we are considering uncapacitated facilities, for each scenario s and period t, each customer i will be assigned to the facility operating in t with the lowest assignment cost. If the dual and primal solutions satisfy all complementary slackness conditions, then the solutions are optimal and the heuristic stops. If not, the heuristic continues with adjustments in order to improve these solutions. Basically, it will try to enforce some complementary conditions that can still be violated. If this is the case, then at least one dual variable is decreased which causes the increase of at least two slacks, associated with distinct facilities, so some dual variables may be increased improving the dual objective value.

If the heuristic is unable to find the optimal solution, it is still able to provide a primal admissible solution and a lower bound to the optimal objective function value. In the branch&bound algorithm, the original problem is first solved in the root node using the dual-based heuristic. If the solution calculated is not the optimal solution (or in cases where it is, but we cannot prove it because of a duality gap), the searching proceeds with a branch&bound scheme that guarantees that the optimal solution is found (if enough time and computational resources are available). The branching is based on those location decision variables that contribute to the complementary slackness violations of the current solution. After some tests, we decided to follow a simple rule and choose the first location variable found that contributes to these violations. Other rules were tested (taking into account the fixed facility costs, expected gains/losses in terms of assignment costs in choosing a second–best source instead of selecting the best source for a given customer), but no significant improvements were observed, especially in large sized problems. Inspired on previous works [Dias, 2007, Erlenkotter, 1978, Van Roy, 1982], location variables are fixed first to zero and then to one. The tree is searched using a depth search procedure. Setting a variable to one is achieved by changing the corresponding fixed cost to zero. To use the current dual solution in the next branch&bound tree node, some changes may have to be made to guarantee dual admissibility (some dual variables must be reduced, with a corresponding increase in some of the slacks). When fixing a variable to zero, its fixed cost is set equal to $+\infty$, guaranteeing the admissibility of the current dual solution that will be used in the next tree node. A node is fathomed only if the current problem is infeasible, the optimal solution of the current problem has been found or the current dual objective function value is worse than the best primal objective function value found so far.

4 Computational experiments and results

We have randomly generated different problem instances by varying the number S of scenarios, number T of time periods, number M of possible facility locations and number N of possible customers according to Table 1. For each combination of (S, T, M, N), with $N > M$, five instances were randomly generated (details about the generation procedure are available from the authors). We have, in total, 780 instances, that were solved by our algorithm and by CPLEX MIP optimizer, v12.4, that was used with its default settings. We have established a maximum computational time for the execution of our algorithm equal to one hour[1] (no time limit was imposed to CPLEX). We note that the smallest instance considered has 1025 variables with 1205 constraints but the largest has 3000750 variables with 3060050 constraints.

Table 1: Numerical data.

S	2	5	10	20
T	5	10	15	
M	5	10	20	50
N	20	50	100	200

Our algorithm was coded in C–language and the computational experiments were carried out on a

[1]This criterion is tested only at the beginning of each node, thus the final computational time may in fact be higher than the time limit established à priori.

AMD Turion(tm) X2 Dual–Core Mobile RM–70 processor at 2.00 GHz with 3.00 GB of RAM.

Tables 2–3 summarize the computational results obtained in terms of primal solution quality achieved in the root node and by the branch&bound algorithm. We report the minimum, average and maximum gap on the five instances solved for each combination of (S, T, M, N). Gap is given by $(f_P - f_{LB})/f_{LB}$, in percentage, where f_P represents the primal objective function value and f_{LB} is the best known lower bound on the optimal primal objective function value. The average results for all S–scenario problems are reported in the last row of the corresponding tables. Tables 4–7 show the solution times (minimum, average and maximum times, in seconds, on the five instances) of the branch&bound, CPLEX, and also the time needed to calculate the admissible solution of the root node. We note that time results do not include the time required to read the problems, only the time to solve them. Due to the time limit restriction, the branch&bound was not able to calculate the optimal solution of some instances. As far as CPLEX results are concerned, the solver could not also solve to optimality some of the problems out of the five instances, due to lack of memory to proceed the calculation. We report these cases and solution gaps are provided. However, if these solution gaps exceeded 10% (*gaps excessively high* when compared with solution gaps provided by our procedure), we have decided to exclude them from the time statistics. We report these cases and CPLEX statistics refer only to those instances that were solved to optimality or presented a *reasonable* gap. Whenever CPLEX was not able to solve any of the five instances, the solver time is given as '*' (in such cases, due to lack of memory to read the problems).

The computational results show that the admissible primal solution calculated in the root node is of very good quality, and is obtained in reasonable computational times. The maximum time needed to compute the root node solution is, for most problems (around 60%), lower than the minimum time required by CPLEX for the same problems. The worst results in terms of gap are observed in instances with $M \in \{20, 50\}$, but still with a maximum gap of 4.01% ($(S, T, M, N) = (20, 15, 50, 100)$). Within each S-scenario problems, in average, the larger gaps are observed in instances with largest M and N. Nevertheless, the branch&bound algorithm is able to improve significantly the quality of the primal solution calculated in the root node. It should be noted that CPLEX has better computational times than branch&bound for $M \in \{20, 50\}$ and $N \in \{100, 200\}$, in general, but as the number of scenarios increases (especially for problems with 20 scenarios), CPLEX shows difficulties in providing a better solution or even to be able to generate a feasible solution. From our computational tests we have observed that different problem instances of the same size can make the optimization algorithms behave very differently, both in terms of the computational times and solution quality. To give an example, considering the 5 instances with size $(S, T, M, N) = (20, 10, 50, 100)$, we have observed the following: the branch&bound algorithm was able to calculate the optimal solution of 2 out of the 5 problems using 1 (after only 215.5 sec) and 3 nodes of the tree, respectively. For the other problems, the algorithm was unable to calculate the optimal solutions due to the time limit restriction, but still improved the solution of two problems (using 6 and 7 nodes of the tree). CPLEX was able to calculate the optimum of one problem only (715 sec), and could not provide feasible solutions for any of the other problems due to memory restrictions. These different behaviors make us think that time should be spent looking at the problem's characteristics to try and delineate more efficient branching rules.

5 Conclusions and future work

In this paper we present a branch&bound algorithm that uses an efficient dual-based heuristic to solve the problems in each of the tree nodes. The computational results show that this approach is competitive for problems up to 50 potential facility locations. CPLEX shows a better behavior for a limited set of problems, but then presents difficulties in solving problems of higher dimension, especially due to memory constraints. For these larger problems, especially considering 10 or more scenarios, and 15 or more time periods, branch&bound appears again as a preferred alternative. It should also be noted that the admissible primal solution calculated in the root node is, in general, of very good quality, presenting small duality gaps, and is obtained reasonably fast. Looking at the structure of the branch&bound tree, other ways of choosing the branching variable can be tested, to diminish the number of nodes that the algorithm needs to explore. Instead of fixing only location variables, it could also be interesting to consider fixing assignment variables (noticing that when they are fixed to one, they implicitly imply fixing a location variable to one). In the model presented, we are minimizing expected costs, and are thus assuming that the decision-maker is risk-neutral. Other objective functions should be considered, as well as additional features to the model (like capacity restrictions), that will raise other interesting questions (like robustness concerns).

Table 2: Solution quality (in %) for problems with 2 and 5 scenarios.

			S=2						S=5					
T	M	N	Root			B&B			Root			B&B		
			min	aver	max	min	aver	max	min	aver	max	min	aver	max
5	5	20	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5	5	50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5	5	100	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5	5	200	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5	10	20	0.00	0.11	0.44	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5	10	50	0.00	0.13	0.36	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5	10	100	0.00	0.03	0.14	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5	10	200	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5	20	50	0.00	0.41	1.52	0.00	0.00	0.00	0.00	0.39	1.41	0.00	0.00	0.00
5	20	100	0.00	0.02	0.12	0.00	0.00	0.00	0.00	0.19	0.56	0.00	0.00	0.00
5	20	200	0.00	0.01	0.04	0.00	0.00	0.00	0.00	0.08	0.26	0.00	0.00	0.00
5	50	100	0.03	0.62	1.85	0.00	0.00	0.00	0.00	0.15	0.49	0.00	0.00	0.00
5	50	200	0.00	0.30	0.58	0.00	0.00	0.00	0.16	0.23	0.34	0.00	0.00	0.00
10	5	20	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.29	0.00	0.00	0.00
10	5	50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.31	0.00	0.00	0.00
10	5	100	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
10	5	200	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
10	10	20	0.00	0.03	0.12	0.00	0.00	0.00	0.00	0.29	1.46	0.00	0.00	0.00
10	10	50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.33	0.00	0.00	0.00
10	10	100	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
10	10	200	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
10	20	50	0.00	0.06	0.32	0.00	0.00	0.00	0.00	0.25	0.57	0.00	0.00	0.00
10	20	100	0.00	0.04	0.20	0.00	0.00	0.00	0.00	0.02	0.12	0.00	0.00	0.00
10	20	200	0.00	0.01	0.06	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.00
10	50	100	0.37	0.68	1.17	0.00	0.00	0.00	0.08	0.46	1.25	0.00	0.02	0.10
10	50	200	0.02	0.25	0.45	0.00	0.02	0.08	0.02	0.15	0.40	0.00	0.10	0.23
15	5	20	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
15	5	50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
15	5	100	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
15	5	200	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
15	10	20	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
15	10	50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
15	10	100	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.04	0.00	0.00	0.00
15	10	200	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
15	20	50	0.00	0.03	0.13	0.00	0.00	0.00	0.00	0.11	0.39	0.00	0.00	0.00
15	20	100	0.00	0.05	0.21	0.00	0.00	0.00	0.00	0.05	0.13	0.00	0.00	0.00
15	20	200	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.03	0.00	0.00	0.00
15	50	100	0.26	0.52	0.90	0.00	0.00	0.00	0.29	0.61	1.20	0.00	0.48	1.20
15	50	200	0.00	0.34	1.47	0.00	0.11	0.57	0.00	0.37	1.06	0.00	0.26	0.75
			0.02	**0.09**	**0.26**	**0.00**	**0.00**	**0.02**	**0.01**	**0.09**	**0.27**	**0.00**	**0.02**	**0.06**

Table 3: Solution quality (in %) for problems with 10 and 20 scenarios.

			S=10						S=20					
T	M	N	Root			B&B			Root			B&B		
			min	aver	max	min	aver	max	min	aver	max	min	aver	max
5	5	20	0.00	0.07	0.36	0.00	0.00	0.00	0.00	0.01	0.07	0.00	0.00	0.00
5	5	50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5	5	100	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5	5	200	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5	10	20	0.00	0.03	0.11	0.00	0.00	0.00	0.00	0.22	1.08	0.00	0.00	0.00
5	10	50	0.00	0.03	0.14	0.00	0.00	0.00	0.00	0.08	0.39	0.00	0.00	0.00
5	10	100	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5	10	200	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5	20	50	0.00	0.14	0.46	0.00	0.00	0.00	0.00	0.13	0.63	0.00	0.00	0.00
5	20	100	0.00	0.25	1.27	0.00	0.00	0.00	0.00	0.19	0.82	0.00	0.00	0.00
5	20	200	0.00	0.03	0.10	0.00	0.00	0.00	0.00	0.19	0.61	0.00	0.03	0.15
5	50	100	0.00	0.17	0.57	0.00	0.00	0.00	0.08	0.70	1.42	0.00	0.46	1.42
5	50	200	0.31	0.84	1.63	0.24	0.60	1.02	1.85	2.47	3.33	1.62	2.10	2.87
10	5	20	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
10	5	50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
10	5	100	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
10	5	200	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
10	10	20	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.23	1.16	0.00	0.00	0.00
10	10	50	0.00	0.06	0.30	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
10	10	100	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
10	10	200	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
10	20	50	0.00	0.48	1.43	0.00	0.00	0.00	0.00	0.12	0.56	0.00	0.00	0.00
10	20	100	0.00	0.06	0.20	0.00	0.00	0.00	0.00	0.14	0.34	0.00	0.00	0.00
10	20	200	0.00	0.01	0.03	0.00	0.00	0.00	0.00	0.07	0.22	0.00	0.04	0.13
10	50	100	0.00	0.64	0.92	0.00	0.41	0.77	0.00	1.43	3.55	0.00	1.10	2.57
10	50	200	0.13	0.39	1.05	0.00	0.31	1.05	0.56	1.08	2.25	0.56	1.08	2.25
15	5	20	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
15	5	50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
15	5	100	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
15	5	200	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
15	10	20	0.00	0.07	0.37	0.00	0.00	0.00	0.00	0.14	0.72	0.00	0.00	0.00
15	10	50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
15	10	100	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.08	0.00	0.00	0.00
15	10	200	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
15	20	50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.28	0.00	0.00	0.00
15	20	100	0.00	0.06	0.24	0.00	0.00	0.00	0.00	0.02	0.07	0.00	0.01	0.05
15	20	200	0.00	0.01	0.03	0.00	0.00	0.00	0.00	0.01	0.04	0.00	0.01	0.04
15	50	100	0.05	1.05	2.25	0.00	0.42	1.14	0.95	2.30	4.01	0.65	2.17	3.81
15	50	200	0.26	1.12	1.79	0.26	1.01	1.79	1.27	1.84	3.01	1.27	1.84	3.01
			0.02	**0.14**	**0.34**	**0.01**	**0.07**	**0.15**	**0.12**	**0.29**	**0.63**	**0.11**	**0.23**	**0.42**

Table 4: Computational time (in sec.) for 2–scenario problems.

T	M	N	Root			B&B			CPLEX[1]		
			min	aver	max	min	aver	max	min	aver	max
5	5	20	0.00	0.00	0.02	0.00	0.01	0.02	0.06	0.07	0.11
5	5	50	0.00	0.00	0.00	0.00	0.00	0.00	0.16	0.17	0.17
5	5	100	0.00	0.00	0.00	0.00	0.00	0.00	0.34	0.51	0.73
5	5	200	0.00	0.03	0.08	0.00	0.03	0.08	0.92	1.03	1.25
5	10	20	0.00	0.01	0.03	0.00	0.02	0.06	0.11	0.17	0.27
5	10	50	0.00	0.06	0.17	0.00	0.11	0.30	0.34	0.36	0.41
5	10	100	0.00	0.08	0.23	0.00	0.12	0.31	0.76	0.88	0.95
5	10	200	0.00	0.26	1.28	0.00	0.26	1.28	2.04	2.19	2.37
5	20	50	0.03	0.13	0.30	0.03	1.60	5.57	0.72	2.38	6.93
5	20	100	0.03	0.83	1.51	0.03	1.82	5.87	1.89	3.18	5.46
5	20	200	0.02	3.24	12.29	0.02	9.29	26.86	4.77	6.32	11.67
5	50	100	0.48	3.23	5.13	0.89	76.01	184.24	5.16	27.34	67.10
5	50	200	6.29	13.41	19.44	45.24	259.14	600.12	17.53	27.47	58.70
10	5	20	0.00	0.00	0.02	0.00	0.00	0.02	0.11	0.13	0.16
10	5	50	0.00	0.00	0.00	0.00	0.00	0.00	0.36	0.38	0.41
10	5	100	0.00	0.03	0.08	0.00	0.03	0.08	0.80	0.90	1.00
10	5	200	0.00	0.02	0.08	0.00	0.05	0.22	2.12	2.24	2.31
10	10	20	0.00	0.01	0.02	0.00	0.03	0.09	0.23	0.27	0.30
10	10	50	0.00	0.01	0.03	0.00	0.02	0.08	0.73	0.86	1.09
10	10	100	0.00	0.01	0.02	0.00	0.01	0.02	2.09	2.16	2.25
10	10	200	0.00	0.02	0.08	0.00	0.02	0.08	5.05	5.23	5.66
10	20	50	0.08	0.58	1.25	0.08	1.09	3.24	1.56	2.50	5.54
10	20	100	0.09	0.89	2.26	0.09	2.47	10.16	4.32	4.57	4.79
10	20	200	0.09	1.68	6.57	0.09	1.79	6.57	12.04	12.39	12.59
10	50	100	1.95	6.33	11.25	33.68	247.76	432.53	17.83	64.68	106.77
10	50	200	40.17	52.61	90.46	434.76	1238.78	3624.30	44.73	69.14	156.31
15	5	20	0.00	0.01	0.06	0.00	0.01	0.06	0.17	0.20	0.23
15	5	50	0.00	0.00	0.00	0.00	0.00	0.00	0.58	0.66	0.70
15	5	100	0.00	0.03	0.09	0.00	0.03	0.09	1.19	1.32	1.45
15	5	200	0.00	0.32	1.56	0.00	0.32	1.56	3.20	3.53	3.90
15	10	20	0.00	0.00	0.02	0.00	0.00	0.02	0.39	0.41	0.44
15	10	50	0.00	0.01	0.02	0.00	0.01	0.02	1.06	1.27	1.44
15	10	100	0.00	0.03	0.11	0.00	0.03	0.11	2.67	3.26	3.88
15	10	200	0.02	0.02	0.02	0.02	0.02	0.02	9.11	9.50	9.67
15	20	50	0.31	0.99	1.97	0.31	1.16	2.62	2.54	3.32	5.16
15	20	100	0.02	1.55	7.27	0.02	1.56	7.27	6.44	6.98	7.74
15	20	200	0.20	0.95	2.26	0.20	1.18	2.73	23.31	23.91	24.62
15	50	100	2.39	5.40	9.50	417.53	1751.79	3604.30	72.45	185.26	314.83
15	50	200	58.62	106.15	210.16	81.59	1469.96	3617.16	59.87	90.70	183.60
			2.84	**5.10**	**9.89**	**26.01**	**129.91**	**311.23**	**7.94**	**14.56**	**25.97**

[1] $(T, M, N) = (15, 50, 200)$: solution gap of 2.54% in one instance.

Table 5: Computational time (in sec.) for 5–scenario problems.

T	M	N	Root			B&B			CPLEX[(1)]		
			min	aver	max	min	aver	max	min	aver	max
5	5	20	0.00	0.00	0.00	0.00	0.00	0.00	0.14	0.16	0.17
5	5	50	0.00	0.01	0.02	0.00	0.01	0.02	0.50	0.55	0.64
5	5	100	0.00	0.01	0.02	0.00	0.01	0.02	1.23	1.27	1.30
5	5	200	0.00	0.01	0.03	0.00	0.01	0.03	2.98	3.15	3.37
5	10	20	0.00	0.01	0.05	0.00	0.01	0.05	0.33	0.34	0.36
5	10	50	0.00	0.18	0.55	0.00	0.37	1.47	1.15	1.26	1.40
5	10	100	0.00	2.63	10.19	0.00	3.16	10.19	2.43	3.36	5.94
5	10	200	0.00	0.23	0.66	0.00	0.23	0.66	7.21	7.37	7.52
5	20	50	0.08	0.70	1.72	0.08	3.08	6.93	2.29	3.37	5.87
5	20	100	0.02	5.38	10.78	0.02	13.00	22.48	5.77	11.65	29.22
5	20	200	2.14	34.26	52.57	5.71	131.14	384.53	22.67	42.81	99.15
5	50	100	4.57	14.99	23.76	25.16	71.50	124.04	22.25	36.06	79.39
5	50	200	49.97	94.49	188.82	537.22	1708.13	3121.40	77.69	131.91	254.75
10	5	20	0.00	0.04	0.20	0.00	0.05	0.25	0.38	0.55	1.19
10	5	50	0.00	0.13	0.50	0.00	0.39	1.78	1.20	1.27	1.31
10	5	100	0.00	0.02	0.03	0.00	0.02	0.03	2.75	3.00	3.37
10	5	200	0.02	0.03	0.05	0.02	0.03	0.05	7.29	7.65	7.89
10	10	20	0.00	0.31	1.11	0.00	0.80	2.62	0.83	0.95	1.11
10	10	50	0.00	0.86	3.48	0.00	2.05	6.91	2.56	2.83	3.25
10	10	100	0.02	0.16	0.53	0.02	0.16	0.53	6.46	6.68	6.88
10	10	200	0.03	0.04	0.05	0.03	0.04	0.05	22.40	23.27	24.09
10	20	50	1.45	4.93	8.81	7.00	22.13	39.56	6.41	12.19	23.18
10	20	100	0.25	9.70	27.44	0.25	15.21	50.59	18.21	18.96	20.64
10	20	200	2.15	19.59	68.11	2.15	127.16	566.25	53.68	57.85	69.75
10	50	100	11.22	50.01	82.74	440.05	1672.13	3601.46	69.59	321.40	556.18
10	50	200	210.62	344.60	432.31	3672.54	3723.90	3843.65	200.18	244.33	297.48
15	5	20	0.00	0.00	0.00	0.00	0.00	0.00	0.51	0.59	0.66
15	5	50	0.00	0.02	0.03	0.00	0.02	0.03	1.86	1.98	2.03
15	5	100	0.03	0.06	0.19	0.03	0.06	0.19	4.43	4.68	4.90
15	5	200	0.05	0.06	0.09	0.05	0.06	0.09	15.29	15.60	15.91
15	10	20	0.22	0.63	1.95	0.33	1.06	3.76	1.23	1.81	3.42
15	10	50	0.00	0.03	0.06	0.00	0.03	0.06	4.51	4.60	4.65
15	10	100	0.03	1.97	8.94	0.03	2.58	12.00	12.22	14.67	23.31
15	10	200	0.06	0.32	1.36	0.06	0.32	1.36	38.41	39.64	40.73
15	20	50	2.81	10.82	25.55	2.81	14.30	32.21	9.95	11.94	14.03
15	20	100	4.99	23.75	48.55	4.99	74.52	161.87	33.23	41.48	63.80
15	20	200	19.19	64.08	156.41	19.19	95.58	282.41	91.23	93.13	95.00
15	50	100	23.43	60.40	120.53	738.02	2553.79	3768.07	111.53	594.58	1426.71
15	50	200	20.58	338.01	639.04	2986.28	3663.47	4178.27	264.97	264.97	264.97
			9.07	**27.78**	**49.16**	**216.46**	**356.42**	**518.61**	**28.92**	**52.15**	**88.86**

[(1)] $(T, M, N) = (10, 50, 200)$: solution gap of 5.04% in one instance; $(T, M, N) = (15, 50, 200)$: statistics refer only to one instance, as gaps on the other four were *excessively high*–ranged from 61% to 70%.

Table 6: Computational time (in sec.) for 10–scenario problems.

T	M	N	Root			B&B			CPLEX[(1)]		
			min	aver	max	min	aver	max	min	aver	max
5	5	20	0.00	0.20	0.53	0.00	0.24	0.73	0.38	0.66	1.65
5	5	50	0.00	0.11	0.50	0.00	0.11	0.50	1.12	1.23	1.34
5	5	100	0.00	0.01	0.02	0.00	0.01	0.02	2.92	3.17	3.48
5	5	200	0.02	0.03	0.03	0.02	0.03	0.03	7.64	7.95	8.38
5	10	20	0.00	0.33	0.45	0.00	0.54	1.28	0.92	1.25	2.14
5	10	50	0.00	1.91	6.94	0.00	6.77	31.25	2.81	4.90	12.78
5	10	100	0.05	5.30	23.34	0.05	6.27	23.34	7.13	7.85	8.69
5	10	200	0.05	3.47	9.61	0.05	3.47	9.61	21.72	22.82	26.07
5	20	50	1.73	5.06	8.74	1.73	7.05	12.56	6.90	7.94	10.58
5	20	100	2.40	8.13	17.85	4.49	31.27	85.94	21.17	32.71	67.05
5	20	200	26.57	164.40	341.20	26.57	221.01	437.28	58.87	64.29	72.42
5	50	100	34.94	79.29	121.93	52.39	588.56	1301.41	69.14	139.48	259.37
5	50	200	418.86	634.12	946.89	3723.94	3935.26	4273.85	524.82	959.66	1527.31
10	5	20	0.00	0.01	0.02	0.00	0.01	0.02	0.83	0.90	0.98
10	5	50	0.02	0.03	0.08	0.02	0.03	0.08	2.67	2.88	3.17
10	5	100	0.03	0.04	0.05	0.03	0.04	0.05	7.44	7.88	8.14
10	5	200	0.06	0.08	0.08	0.06	0.08	0.08	21.68	23.37	24.98
10	10	20	0.00	0.29	1.20	0.00	0.29	1.20	1.93	2.10	2.43
10	10	50	0.03	9.70	34.91	0.03	15.07	59.16	6.66	9.36	19.19
10	10	100	0.06	1.23	3.78	0.06	1.23	3.78	21.11	22.02	22.50
10	10	200	0.11	0.11	0.13	0.11	0.11	0.13	55.16	57.02	59.45
10	20	50	6.16	27.61	49.41	6.16	91.57	207.54	18.80	55.25	132.16
10	20	100	7.22	73.58	205.44	7.22	357.21	803.21	56.55	118.88	226.20
10	20	200	1.64	241.84	460.86	1.64	302.18	648.24	127.48	136.49	152.48
10	50	100	73.26	225.40	334.34	73.26	2847.19	3658.22	161.01	612.17	841.99
10	50	200	1091.86	1871.35	2703.61	3344.55	4004.48	4996.68	401.34	610.96	820.58
15	5	20	0.02	1.33	6.54	0.02	2.06	10.22	1.53	1.78	1.95
15	5	50	0.05	0.07	0.13	0.05	0.07	0.13	4.98	5.10	5.34
15	5	100	0.08	0.20	0.41	0.08	0.20	0.41	14.98	15.69	16.65
15	5	200	0.14	0.43	1.47	0.14	0.43	1.47	43.07	45.16	47.69
15	10	20	0.28	1.62	4.68	0.28	10.41	34.41	3.21	3.77	4.88
15	10	50	0.05	1.24	3.57	0.05	1.24	3.57	12.81	13.24	13.73
15	10	100	0.11	10.94	41.96	0.11	10.94	41.96	37.46	40.84	46.11
15	10	200	0.17	0.21	0.23	0.17	0.21	0.23	99.33	103.01	105.32
15	20	50	0.78	17.82	40.72	0.78	46.89	158.57	32.93	43.70	72.59
15	20	100	8.19	58.64	105.66	8.19	270.82	796.27	95.40	164.64	398.57
15	20	200	0.28	508.17	1231.29	0.28	625.82	1231.29	217.14	248.22	310.01
15	50	100	187.43	427.05	785.32	2662.76	3490.24	3750.85	313.22	361.24	409.27
15	50	200	526.03	1771.76	3170.56	3674.49	4156.30	4983.22	*	*	*
			61.25	**157.77**	**273.45**	**348.46**	**539.38**	**706.89**	**65.37**	**104.20**	**151.25**

[(1)] $(T, M, N) = (5, 50, 200)$: solution gap of 2.28% in one instance, and solution gaps *excessively high* in two instances; $(T, M, N) = (10, 50, 100)$: two *excessively high* gaps (15% and 19%); $(T, M, N) = (10, 50, 200)$: three *excessively high* gaps (around 60%). $(T, M, N) = (15, 20, 200)$: one solution gap of 0.29%; $(T, M, N) = (15, 50, 100)$: two *excessively high* gaps (62% and 69%) and no feasible solution was provided for one instance.

Table 7: Computational time (in sec.) for 20–scenario problems.

T	M	N	Root			B&B			CPLEX[1]		
			min	aver	max	min	aver	max	min	aver	max
5	5	20	0.00	0.13	0.61	0.00	0.38	1.87	0.80	1.48	3.00
5	5	50	0.02	0.65	3.15	0.02	0.65	3.15	2.85	3.16	3.53
5	5	100	0.03	30.86	154.19	0.03	30.86	154.19	7.86	9.43	14.35
5	5	200	0.06	0.08	0.09	0.06	0.08	0.09	25.26	26.22	27.11
5	10	20	0.02	2.52	4.90	0.02	6.26	20.87	2.11	2.51	2.75
5	10	50	1.09	13.61	23.99	1.09	21.88	42.84	7.89	16.82	47.86
5	10	100	3.09	70.80	245.59	3.09	108.01	431.64	24.32	28.15	37.46
5	10	200	0.09	145.04	638.04	0.09	145.04	638.04	64.19	73.17	101.15
5	20	50	0.34	21.11	44.29	0.34	113.64	435.97	20.61	74.18	221.68
5	20	100	19.64	101.85	222.91	19.64	403.43	1015.67	56.89	117.68	288.10
5	20	200	47.05	621.55	1342.97	58.31	2219.18	4636.38	164.66	537.78	1190.23
5	50	100	202.60	310.82	359.07	988.31	2449.90	3983.55	264.14	468.16	883.07
5	50	200	383.79	1812.58	2803.42	3981.65	4692.20	5493.17	*	*	*
10	5	20	0.03	0.55	1.78	0.03	1.15	4.79	2.11	2.22	2.29
10	5	50	0.06	0.07	0.09	0.06	0.07	0.09	8.25	8.55	8.71
10	5	100	0.13	0.15	0.19	0.13	0.15	0.19	23.53	24.93	26.16
10	5	200	0.23	0.27	0.33	0.23	0.27	0.33	58.75	61.76	65.38
10	10	20	0.03	2.56	6.22	0.03	3.01	6.22	5.46	5.95	7.49
10	10	50	0.08	23.42	45.74	0.08	23.42	45.74	25.91	27.31	29.84
10	10	100	0.16	7.62	35.22	0.16	7.62	35.22	55.93	58.88	61.87
10	10	200	0.30	1.62	3.76	0.30	2.45	7.57	131.67	137.53	140.46
10	20	50	34.41	140.33	225.34	34.41	352.87	1008.43	62.28	75.39	97.60
10	20	100	53.57	193.76	409.70	107.17	1642.17	3957.66	141.54	228.71	470.41
10	20	200	689.88	1786.88	3325.52	689.88	2822.17	4521.18	313.94	390.26	512.89
10	50	100	215.51	748.85	1289.93	215.51	3259.06	4482.38	715.61	715.61	715.61
10	50	200	1860.63	3639.29	4646.76	4283.56	4696.01	5646.12	*	*	*
15	5	20	0.06	9.97	49.55	0.06	9.97	49.55	4.23	4.46	4.57
15	5	50	0.14	0.17	0.19	0.14	0.17	0.19	15.34	16.33	17.22
15	5	100	0.27	0.42	0.86	0.27	0.42	0.86	40.19	45.06	47.83
15	5	200	0.50	0.57	0.70	0.50	0.57	0.70	97.42	105.32	109.15
15	10	20	0.06	12.31	43.73	0.06	32.01	142.21	10.97	20.22	54.18
15	10	50	0.17	79.79	297.60	0.17	157.06	683.95	42.01	46.75	61.62
15	10	100	0.34	8.53	37.82	0.34	35.66	139.39	104.63	108.50	111.84
15	10	200	0.61	0.65	0.73	0.61	0.65	0.73	221.55	233.70	241.26
15	20	50	49.64	198.52	353.08	49.64	1109.26	3475.71	104.99	244.58	616.14
15	20	100	109.61	435.24	641.11	109.61	2080.56	3728.63	228.11	283.64	332.78
15	20	200	364.42	2457.07	5352.72	364.42	2525.55	5352.72	*	*	*
15	50	100	414.34	1973.06	3138.74	3914.32	4599.44	5882.31	*	*	*
15	50	200	6064.95	12619.44	15228.66	6064.95	12619.44	15228.66	*	*	*
			269.69	**704.43**	**1050.75**	**535.62**	**1183.91**	**1827.15**	**89.88**	**123.66**	**192.81**

[1] $(T, M, N) = (5, 50, 100)$: one *excessively high* gap (39%); $(T, M, N) = (10, 50, 100)$: CPLEX was only able to solve one of the instances (no feasible solutions were provided for the other four).

References

[Cornuejols, 1990] Cornuejols, G., Nemhauser, G. L., and Wolsey, L. (1990). The Uncapacitated Facility Location Problem. In *Mirchandani, P. B., and Francis, R. L. (editors), Discrete Location Theory, 119–171. John Wiley&Sons.*

[Dias, 2007] Dias, J., Captivo, M. E., and Clímaco, J. (2007). Efficient primal-dual heuristic for a dynamic location problem. *Computers & Operations Research*, 34(6):1800–1823.

[Erlenkotter, 1978] Erlenkotter, D. (1978), A Dual-Based Procedure for Uncapacitated Facility Location. *Operations Research*, 26(6):992–1009.

[Marques, 2011] Marques, M. C. and Dias, J. M. (2011). Simple dynamic location problem with uncertainty: a primal-dual heuristic approach. *Relatórios de Investigação do INESC Coimbra, nº 14/2011, INESC Coimbra.*

[Van Roy, 1982] Van Roy, T. J. and Erlenkotter, D. (1982). A Dual-Based Procedure for Dynamic Facility Location. *Management Science*, 28(10):1091–1105.

[Snyder, 2006] Snyder, L. V. (2006). Facility location under uncertainty: a review. *IIE Transactions*, 28:537–554.

Dantzig-Wolfe reformulations for the forest harvest scheduling subject to maximum area restrictions

Isabel Martins†, Filipe Alvelos‡, Miguel Constantino§, Ricardo Magalhães♯

† *Centro de Investigação Operacional/DCEB, Instituto Superior de Agronomia, Portugal*
E-mail:`isabelinha@isa.utl.pt`

‡ *Centro Algoritmi/DPS, Universidade do Minho, Portugal*
E-mail:`falvelos@dps.uminho.pt`

§ *Centro de Investigação Operacional/DEIO, Faculdade de Ciências de Lisboa, Portugal*
E-mail:`miguel.constantino@fc.ul.pt`

♯ *Centro de Investigação Operacional/DEIO, Faculdade de Ciências de Lisboa, Portugal*
E-mail:`rjmagalhaes@fc.ul.pt`

Abstract

We describe two Dantzig-Wolfe decompositions of the so-called bucket formulation for the forest harvest scheduling problem with maximum area restrictions. A heuristic solution to the problem is obtained by solving the final restricted master problem provided by column generation, enforcing the integrality constraints. We compare the approaches and present preliminary computational results.

Keywords: Forest harvest scheduling, Dantzig-Wolfe decomposition.

1 Introduction

Forest harvesting for timber production causes negative environmental impacts, primarily habitat alteration and loss of biodiversity, soil, water quality and scenic beauty. A common practice in many countries to reduce these impacts has been to restrict the areas of clearcuts. Addressing these constraints has led to an evolution of model approaches that support forest management. The most recent approach, the so-called area restriction model (ARM), lets the formulation itself suggest stand aggregation when the sum of the areas does not violate the maximum clearcut area. Three main basic integer programming models for the ARM have been described in the literature: the path formulation, with an exponential number of constraints, the cluster formulation, with an exponential number of variables, and the bucket formulation, with a polynomial number of variables and constraints (for a survey of integer programming approaches to solving the ARM, we refer those interested to ([4]).

The harvest scheduling problem that we shall consider consists of selecting, for each period in the planning horizon, a set of stands to be harvested, in order to maximize the timber's net present value. The stand selection is subject to several restrictions. Maximum area restrictions (constraints A1) impose that the area of each clearcut does not exceed the maximum allowed size. Each stand is harvested at the most once in the planning horizon, *i.e.* the minimum rotation in the stand is longer than the latter (constraints A2). Other requirement is a steady flow of harvested timber (constraints A3). This restriction is mainly to ensure that the industry is able to continue operating with similar levels of machine and labor utilizations.

In this work, we propose two Dantzig-Wolfe decompositions of the bucket formulation for this problem. We describe the bucket formulation in Section 2, and the decompositions in Sections 3 and 4. In Section 5, we briefly describe how the pricing subproblems are solved. We compare the approaches presented and report on the computational results in Section 6. Finally, we present the main conclusions in Section 7.

2 A compact model

In this section, the compact *bucket formulation* ([1]) is described. As each stand is harvested once at the most, the harvested area in a forest is a set of clearcuts (maximal harvested connected regions) which do not overlap in the course of the planning horizon. Thus, each clearcut may be represented by one of its stands, for example the one with the smallest index. In this paper two stands are considered to be adjacent when both share a boundary with positive length, *i.e* that is not a discrete set of points (the so-called strong adjacency). According to this definition, clearcuts in Figure 1 are, in period 1, regions $\{2,3\}$ and $\{6,8\}$ and, in period 2, $\{1,5\}$ and $\{7\}$. These clearcuts may be represented by stands 2, 6, 1 and 7, respectively.

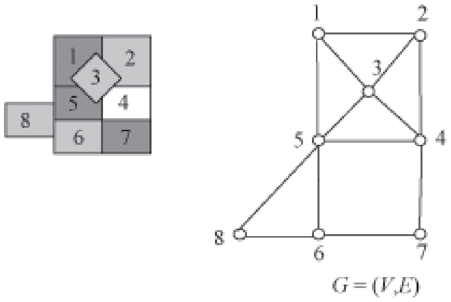


Figure 1: Forest with eight stands and its graph representation. Stands 2, 3, 6 and 8 are harvested in period 1, stands 1, 5 and 7 in period 2, and there is no intervention in stand 4. The area of each stand is 1 ha.

Let us consider the empty set C_k (*bucket*) for each stand k. Assigning stands to C_k (stand k and stands $i > k$) corresponds to selecting these stands to be harvested. The model assigns stands to buckets in such a way that constraints A1, A2 and A3 are not violated. A bucket remains empty if no stands are assigned to it. Each non empty bucket represents a feasible clearcut or a set of feasible clearcuts. The stand with the smallest index in a non empty bucket C_k is k.

We introduce next the notation used.

• T - set of time periods indexed by $t = 1, ..., |T|$
• I - set of stands indexed by $i = 1, ..., |I|$
• a_i - area of stand i; $i \in I$
• A_{max} - maximum clearcut area
• p_i^t - timber net present value from stand i if it is harvested in period t; $i \in I$; $t \in T$
• v_i^t - volume of timber of stand i in period t; $i \in I$; $t \in T$
• Δ - maximum allowed variation on volume of timber harvested between two consecutive periods.

Let $G = (I, E)$ be a graph, where each stand in I is represented by a vertex and the endpoints of each edge in E correspond to two adjacent stands. Using the definition of strong adjacency, the graph is planar, *i.e* it can be drawn in a plane surface without crossing edges.

Let $\mathcal{Q}$ be the set of maximal cliques of G indexed by $P \in \mathcal{Q}$. A clique is the set of nodes of a complete subgraph of the graph, which has an edge between each pair of vertices, and it is maximal if it is not contained in any other clique. Since the graph is planar there are no cliques with more than four vertices ([2]). For the graph in Figure 1, $\mathcal{Q} = \{\{1,2,3\},\{1,3,5\},\{2,3,4\},\{3,4,5\},\{4,7\},\{5,6,8\},\{6,7\}\}$.

We can represent the set of buckets in the forest as $\mathrm{C} = \{C_1, ..., C_{|I|}\}$.

The decision variables in the bucket model are therefore as follows:

$$x_i^{kt} = \begin{cases} 1 & \text{if stand } i \text{ is selected to belong to bucket } C_k \text{ in period } t \\ 0 & \text{otherwise; } \quad k \in I; t \in T; i = k, ..., |I| \end{cases}$$

$$w_P^{kt} = \begin{cases} 1 & \text{if at least one stand from clique } P \text{ is selected to belong to bucket } C_k \\ & \text{in period } t \\ 0 & \text{otherwise; } \quad k \in I; t \in T; P \in \mathcal{Q} : \max\limits_{i \in P}\{i\} \geq k. \end{cases}$$

The model is the following:

$$\max \sum_{t \in T} \sum_{k \in I} \sum_{i=k}^{|I|} p_i^t x_i^{kt} \tag{2.1}$$

subject to

$$x_i^{kt} \leq w_P^{kt}; k \in I; t \in T; i \geq k; P \in \mathcal{Q} : i \in P \tag{2.2}$$

$$\sum_{k \leq \max\limits_{i \in P}\{i\}} w_P^{kt} \leq 1; t \in T; P \in \mathcal{Q} \tag{2.3}$$

$$\sum_{i=k+1}^{|I|} a_i x_i^{kt} \leq (A_{max} - a_k) x_k^{kt}; k \in I; t \in T \tag{2.4}$$

$$\sum_{t \in T} \sum_{k=1}^{i} x_i^{kt} \leq 1; i \in I \tag{2.5}$$

$$\sum_{i \in I} v_i^t \sum_{k=1}^{i} x_i^{kt} \geq (1-\Delta) \sum_{i \in I} v_i^{t-1} \sum_{k=1}^{i} x_i^{k,t-1}; t = 2, ..., |T| \tag{2.6}$$

$$\sum_{i \in I} v_i^t \sum_{k=1}^{i} x_i^{kt} \leq (1+\Delta) \sum_{i \in I} v_i^{t-1} \sum_{k=1}^{i} x_i^{k,t-1}; t = 2, ..., |T| \tag{2.7}$$

$$x_i^{kt} \in \{0,1\}; k \in I; t \in T; i = k, ..., |I| \tag{2.8}$$

$$w_P^{kt} \geq 0; k \in I; t \in T; P \in \mathcal{Q} : \max_{i \in P}\{i\} \geq k. \tag{2.9}$$

The expression (2.1) states the management objective of maximizing the net present value of timber harvested. Constraints (2.2) define the relationship between variables x and w. Constraints (2.2) and (2.3) ensure that in each period every two adjacent stands are in one bucket at the most. Constraints (2.4) guarantee that each bucket does not exceed the maximum allowed size (constraints A1). Constraints (2.4) also state that if C_k is non-empty then C_k contains stand k, and thus the stand in C_k with the smallest index is k. Constraints (2.5) state that each stand is harvested at the most once in the planning horizon (constraints A2). Constraints (2.6) and (2.7) allow harvested volumes in each period to range from $1-\Delta$ to $1+\Delta$ times the harvested volume in the previous period (constraints A3). The other constraints state binary and non-negativity requirements on variables. The integrality of variables x, together with constraints (2.2), implies the integrality of variables w in at least one optimal solution. Note that a non empty bucket is a region that might be disconnected since there are no constraints to enforce its connectivity. However, any solution with a disconnected harvested set C_k is equivalent to the solution where C_k is replaced by its clearcuts, each with an area not exceeding A_{max}. Hence there is no need to add explicit constraints in the model to enforce connectivity of the buckets.

The number of variables and constraints of this model can be reduced ([1]). Observe that if a stand is too "far" from stand k then it is not worth assigning it to bucket C_k, because the area of any connected region with both stands would exceed the maximum. Back to Figure 1, considering $A_{max} = 2$, for $k = 1$, variables x_4^{1t}, x_6^{1t}, x_7^{1t}, x_8^{1t}, $w_{4,7}^{1t}$, $w_{6,7}^{1t}$ are null in any solution, and thus may not be considered.

We consider two main Dantzig-Wolfe decompositions for the model presented above. The *knapsak decomposition* is obtained by reformulating the set $\mathcal{K}$ defined by constraints (2.4) and (2.8), while the *knapsack-and-clique decomposition* corresponds to the reformulation of the set $\mathcal{C}$ defined by (2.2), (2.4), (2.8) and (2.9). In each case, the representation of the convex hull of the corresponding set by extreme points is used to strengthen the model. As those convex hulls have an exponential number of extreme points in general, the linear programming relaxations of the resulting models are solved by column generation. The pricing subproblems consider reduced cost objective functions over the sets $\mathcal{K}$ and $\mathcal{C}$ respectively.

3 Knapsack decomposition

In this section, we propose a Dantzig-Wolfe reformulation of the compact model, by reformulating the set defined by constraints (2.4) and (2.8).

3.1 The master problem

Let $d = |T| \times |I| \times (|I|+1)/2$ be the number of variables x_i^{kt} in the model and $\mathcal{K} = \{x \in \mathbb{R}^d$ satisfying (2.4) and 2.8$\}$. Observe that $\mathcal{K}$ can be decomposed into $|I| \times |T|$ sets $\mathcal{K}^{kt} = \{x \in \{0,1\}^{|I|-k+1} : \sum_{i=k+1}^{|I|} a_i x_i^{kt} \leq (A_{max} - a_k) x_k^{kt}\}$. Observe further that for a given k, the sets are identical, and that an element of a set $\mathcal{K}^{kt}$ is either the null vector or the incidence vector of a region with stands $i \geq k$, containing k and with area not greater than A_{max}. That is, $\mathcal{K}^{kt} = \{0\} \cup \{\bar{x}_S^k : S \in \mathcal{S}^k\}$ where $\mathcal{S}^k$ is the set of all regions S with stands $i \geq k$ such that $k \in S$ and the area of S is not greater than A_{max}, and $\bar{x}_S^k$ is the incidence vector of S:

$$\bar{x}_{iS}^k = \begin{cases} 1 & \text{if stand } i \text{ belongs to region } S \in \mathcal{S}^k \\ 0 & \text{otherwise.} \end{cases}$$

Now, for $t \in T$, $k \in I$ and $S \in \mathcal{S}^k$ define the variables

$$y_S^{kt} = \begin{cases} 1 & \text{if region } S \text{ is selected to be harvested in period } t \\ 0 & \text{otherwise} \end{cases}$$

and for each $k \in I$ and $t \in T$ define the variable y_0^{kt} that assumes the unitary value if none of sets of $\mathcal{S}^k$ is selected to be harvested in period t and the null value otherwise. We have $\mathcal{K}^{kt} = \{x \in \mathbb{R}^{|I|-k+1} : x = \sum_{S\in\mathcal{S}^k} \bar{x}_S^k y_S^{kt}, y_0^{kt} + \sum_{S\in\mathcal{S}^k} y_S^{kt} = 1, y_S^{kt} \in \{0,1\}, S \in \mathcal{S}^k\}$, for each $k \in I$ and $t \in T$. By incorporating this reformulation of the sets $\mathcal{K}^{kt}$ into the bucket formulation, we obtain the master problem

$$\max \sum_{t\in T}\sum_{k\in I}\sum_{S\in\mathcal{S}^k}\sum_{i=k}^{|I|} p_i^t \bar{x}_{iS}^k y_S^{kt} \tag{3.1}$$

$subject\, to$

$$y_0^{kt} + \sum_{S\in\mathcal{S}^k} y_S^{kt} = 1; k \in I; t \in T \tag{3.2}$$

$$\sum_{S\in\mathcal{S}^k} \bar{x}_{iS}^k y_S^{kt} \le w_P^{kt}; k \in I; t \in T; i \ge k; P \in \mathcal{Q} : i \in P \tag{3.3}$$

$$\sum_{k\le\max\{i:i\in P\}} w_P^{kt} \le 1; t \in T; P \in \mathcal{Q} \tag{3.4}$$

$$\sum_{t\in T}\sum_{k=1}^{i}\sum_{S\in\mathcal{S}^k} \bar{x}_{iS}^k y_S^{kt} \le 1; i \in I \tag{3.5}$$

$$\sum_{i\in I} v_i^t \sum_{k=1}^{i}\sum_{S\in\mathcal{S}^k} \bar{x}_{iS}^k y_S^{kt} \ge (1-\Delta)\sum_{i\in I} v_i^{t-1}\sum_{k=1}^{i}\sum_{S\in\mathcal{S}^k} \bar{x}_{iS}^k y_S^{k,t-1}; t = 2, ..., |T| \tag{3.6}$$

$$\sum_{i\in I} v_i^t \sum_{k=1}^{i}\sum_{S\in\mathcal{S}^k} \bar{x}_{iS}^k y_S^{kt} \le (1+\Delta)\sum_{i\in I} v_i^{t-1}\sum_{k=1}^{i}\sum_{S\in\mathcal{S}^k} \bar{x}_{iS}^k y_S^{k,t-1}; t = 2, ..., |T| \tag{3.7}$$

$$y_S^{kt} \in \{0,1\}; k \in I; t \in T; S \in \mathcal{S}^k \tag{3.8}$$

$$y_0^{kt} \in \{0,1\}; k \in I; t \in T \tag{3.9}$$

$$w_P^{kt} \ge 0; k \in I; t \in T; P \in \mathcal{Q} : \max_{i\in P}\{i\} \ge k. \tag{3.10}$$

We shall refer to this formulation as $\mathcal{S}$-knapsack decomposition. Note that constraints (3.5) imply constraints (3.2), so these can be removed.

Observe also that variables x in the bucket formulation and variables y in the $\mathcal{S}$-knapsack decomposition are related through the equations $x_i^{kt} = \sum_{S\in\mathcal{S}^k} \bar{x}_{iS}^k y_S^{kt}$. The linear programming relaxation of the $\mathcal{S}$-knapsack decomposition corresponds to replacing, in the bucket model, the set $\mathcal{K}$ defined by constraints (2.4) and (2.8) by its convex hull. Given that the extreme points of this convex hull are not necessarily integer, we may state the following:

Proposition 3.1. *The LP bound of the $\mathcal{S}$-knapsack decomposition is better than or equal to that of the bucket formulation.*

3.2 The pricing subproblem

Relaxation of the binary requirement on the variables y_S^{kt} leads to the linear relaxation of the master problem. Constraints (3.8) and (3.9) are simply replaced respectively by $y_S^{kt} \ge 0$ and $y_0^{kt} \ge 0$ because constraints (3.2) guarantee $y_S^{kt} \le 1$ and $y_0^{kt} \le 1$. The pricing subproblem kt for each node $k \in I$ and for each period $t \in T$ consists of finding a variable y_S^{kt}, with $S \in \mathcal{S}^k$, such that the corresponding reduced cost is maximum. The variables y_0^{kt} and w_P^{kt} are inserted into the first restricted master problem.

Let Ω^{kt}, α_{iP}^{kt}, β_P^t, θ_i, μ^t and ν^t denote the dual variables associated with constraints (3.2), (3.3), (3.4), (3.5), (3.6) and (3.7) of the linear relaxation of the master problem and Ω^{kt*}, α_{iP}^{kt*}, β_P^{t*}, θ_i^*, μ^{t*} and ν^{t*} assume an optimal dual solution of the linear relaxation of a restricted master problem. By definition of reduced cost, the objective function of the pricing subproblem kt is given by

$$\max_{S\in\mathcal{S}^k} \{\sum_{i\in S} \epsilon_i^{t*}\} + \Omega^{kt*}$$

where

$$\epsilon_i^{1*} = p_i^1 + \theta_i^* - \mu^{2*}(1-\Delta)v_i^1 - \nu^{2*}(1+\Delta)v_i^1 + \sum_{\substack{P\in\mathcal{Q}:\\ i\in P}} \alpha_{iP}^{k1*}$$

$$\epsilon_i^{t*} = p_i^t + \theta_i^* + \mu^{t*}v_i^t + \nu^{t*}v_i^t - \mu^{t+1*}(1-\Delta)v_i^t - \nu^{t+1*}(1+\Delta)v_i^t + \sum_{\substack{P\in\mathcal{Q}:\\ i\in P}} \alpha_{iP}^{kt*},\ t=2,...,|T|-1$$

$$\epsilon_i^{|T|*} = p_i^{|T|} + \theta_i^* + \mu^{|T|*}v_i^{|T|} + \nu^{|T|*}v_i^{|T|} + \sum_{\substack{P\in\mathcal{Q}:\\ i\in P}} \alpha_{iP}^{k|T|*} \text{ (with } \theta_i^* \le 0,\ \mu^{t*} \ge 0,\ \nu^{t*} \le 0 \text{ and } \alpha_{iP}^{kt*} \le 0).$$

The subproblem kt can be formulated by the following integer program. The variables will be

$$x_i = \begin{cases} 1 & \text{if stand } i \text{ is selected to belong to region } S \\ 0 & \text{otherwise; } i = k, \ldots, |I|. \end{cases}$$

And the formulation is as follows:

$$\max \sum_{i=k}^{|I|} \epsilon_i^{t*} x_i + \Omega^{kt*} \tag{3.11}$$

$$\textit{subject to}$$

$$\sum_{i=k+1}^{|I|} a_i x_i \le A_{max} - a_k \tag{3.12}$$

$$x_k = 1 \tag{3.13}$$

$$x_i \in \{0,1\}\,; i = 1, \ldots, k. \tag{3.14}$$

The objective function (3.11) is to maximize the sum of the node weights ϵ_i^{t*} over the selected nodes and to add the value Ω^{kt*} to the optimal sum. Constraint (3.13) guarantees that region S contains stand k. Constraints (3.12) ensure that the area of S does not exceed A_{max}. Constraints (3.14) state the variable types.

3.3 An alternative Knapsack decomposition

In the first section, we pointed out there is no need to enforce connectivity in the bucket formulation. Indeed, if a bucket with area not exceeding A_{max} is disconnected, then it can be broken up into smaller connected buckets with area not exceeding A_{max}. Nevertheless, we can consider an alternative formulation with connectivity constraints. Consider a decomposition to this alternative model, similar to the knapsack decomposition, with the connectivity constraints to be included in the knapsack sets to be reformulated. Since now this sets are more constrained, the linear programming bounds obtained by the knapsack decomposition with connectivity are not worse than those obtained by the knapsack decomposition without connectivity.

Let $\mathcal{R}^k$ denote the set of all regions from $\mathcal{S}^k$ that are connected, for $k \in I$. We shall refer to the master problem where sets $\mathcal{S}^k$ are replaced by $\mathcal{R}^k$ as $\mathcal{R}$-knapsack decomposition. The solution set of the master problem does not change if sets $\mathcal{S}^k$ are replaced by $\mathcal{R}^k$. However, we have the following:

Proposition 3.2. *The LP bound of the $\mathcal{R}$-knapsack decomposition is better than or equal to that of the $\mathcal{S}^k$-knapsack decomposition.*

That the LP bound of the $\mathcal{R}$-knapsack decomposition is never worse than that of the $\mathcal{S}$-knapsack decomposition follows from the fact that there are more constraints in the bucket formulation defining the set which is reformulated, and the remaining constraints in the master problem are the same. Next, we will give an example that shows there are instances for which the LP bound of the $\mathcal{R}$-knapsack decomposition is strictly better.

Consider a forest with seven stands (Table 1) and let $A_{max} = 3.0$ and $\Delta = 1.0$. Optimal solutions of the linear relaxations of the $\mathcal{S}$-knapsack and $\mathcal{R}$-knapsack decompositions are the following, respectively (only the non-null values are displayed):

$y_{\{3\}}^{3,1} = 0.(3)$, $y_{\{7\}}^{7,1} = 0.(3)$, $y_{\{5,6\}}^{5,1} = 1$, $y_{\{1,2,3\}}^{1,1} = 0.(3)$, $y_{\{1,2,7\}}^{1,1} = 0.(3)$, $y_{\{1,3,7\}}^{1,1} = 0.(3)$, $y_{\{2\}}^{2,2} = 0.(3)$, $y_{\{4\}}^{4,2} = 1$ (note that set $\{1,3,7\}$ is disconnected),

$w^{1,1}_{\{1,2\}} = 1$, $w^{1,1}_{\{2,3\}} = 0.(6)$, $w^{3,1}_{\{2,3\}} = 0.(3)$, $w^{1,1}_{\{2,7\}} = 0.(6)$, $w^{7,1}_{\{2,7\}} = 0.(3)$, $w^{1,1}_{\{3,4\}} = 0.(6)$, $w^{3,1}_{\{3,4\}} = 0.(3)$, $w^{5,1}_{\{5,6\}} = 1$, $w^{2,2}_{\{1,2\}} = 0.(3)$, $w^{2,2}_{\{2,3\}} = 0.(3)$, $w^{2,2}_{\{2,7\}} = 0.(3)$, $w^{4,2}_{\{3,4\}} = 1.0$, $w^{4,2}_{\{4,5\}} = 1.0$,

$y^{1,2}_0 = y^{3,2}_0 = y^{4,1}_0 = y^{5,2}_0 = y^{6,2}_0 = y^{7,2}_0 = 1.0$, $y^{2,1}_0 = 0.(3)$, $y^{2,2}_0 = 0.(6)$, with the value of the objective function 165104.1;

$y^{1,1}_{\{1,2,7\}} = 1.0$, $y^{4,1}_{\{4,5,6\}} = 1.0$, $y^{3,2}_{\{3\}} = 1.0$,

$w^{1,1}_{\{1,2\}} = w^{1,1}_{\{2,3\}} = w^{1,1}_{\{2,7\}} = w^{4,1}_{\{3,4\}} = w^{4,1}_{\{4,5\}} = w^{3,2}_{\{2,3\}} = w^{3,2}_{\{3,4\}} = 1.0$,

$y^{1,2}_0 = y^{2,2}_0 = y^{3,1}_0 = y^{4,2}_0 = y^{5,2}_0 = y^{6,2}_0 = y^{7,2}_0 = 1.0$, with the value of the objective function 161346.1. These solutions are obtained using CPLEX 12.2 ([5]) as an integer programming solver.

Table 1: Example.

Stand i	a_i	p^1_i	p^2_i	v^1_i	v^2_i	Nodes adjacent to i
1	1.00	27135.0	18387.2	524.2	558.5	2
2	1.00	26524.1	18030.2	512.9	548.7	1, 3, 7
3	1.00	26524.1	18030.2	512.9	548.7	2, 4
4	1.00	7693.6	5789.0	263.2	341.5	3, 5
5	1.00	28094.4	18950.6	542.0	573.9	4, 6
6	1.00	26934.4	18269.8	520.5	555.3	5
7	1.00	26934.4	18269.8	520.5	555.3	2

Concerning the pricing subproblem for the $\mathcal{R}$-knapsack decomposition, it can be formulated as (3.11) - (3.14), with additional constraints "the set $\{i : x_i = 1\}$ induces a connected subgraph of $G = (I, E)$".

4 Knapsack-and-clique decomposition

In this section, we define the knapsack-and-clique Dantzig-Wolfe decomposition of the compact model, by reformulating the set defined by constraints (2.2), (2.4), (2.8) and (2.9).

4.1 The master problem

Let d and e be the number of variables x^{kt}_i and w^{kt}_P in the model respectively, $e = \sum_{k,t} e(k,t)$ where $e(k,t) = |\{P \in \mathcal{Q} : P \cap \{k,\ldots,|I|\} \neq \emptyset\}|$, and $\mathcal{C} = \{(x,w) \in \mathbb{R}^{d+e}$ satisfying (2.2), (2.4), (2.8) and (2.9)$\}$.

As before, $\mathcal{C}$ can be decomposed into $|I|\times|T|$ sets $\mathcal{C}^{kt} = \{(x,w) \in \{0,1\}^{(|I|-k+1)+e(k,t)} : \sum_{i=k+1}^{|I|} a_i x^{kt}_i \leq (A_{max} - a_k)x^{kt}_k$ and $x^{kt}_i \leq w^{kt}_P; i \geq k; P \in \mathcal{Q} : i \in P\}$.

Again, for a given k, the sets are identical, and an element (x,w) of a set $\mathcal{C}^{kt}$ is either the null vector or a vector $(\bar{x}^k_S, \bar{w}^k_S)$ where S is a region with stands $i \geq k$, containing k and with area not greater than A_{max}, and $\bar{x}^k_S$, $\bar{w}^k_S$ are defined as

$$\bar{x}^k_{iS} = \begin{cases} 1 & \text{if stand } i \text{ belongs to region } S \\ 0 & \text{otherwise} \end{cases}$$

$$\bar{w}^k_{PS} = \begin{cases} 1 & \text{if clique } P \text{ is such that } P \cap S \neq \emptyset \\ 0 & \text{otherwise.} \end{cases}$$

For $t \in T$, $k \in I$ and $S \in \mathcal{S}^k$, we consider again variables y^{kt}_S and y^{kt}_0. The master problem is as follows:

$$\max \sum_{t\in T}\sum_{k\in I}\sum_{S\in\mathcal{S}^k}\sum_{i=k}^{|I|} p^t_i \bar{x}^k_{iS} y^{kt}_S \tag{4.1}$$

subject to

$$y^{kt}_0 + \sum_{S\in\mathcal{S}^k} y^{kt}_S = 1; k \in I; t \in T \tag{4.2}$$

$$\sum_{k\leq\max\{i:i\in P\}}\sum_{S\in\mathcal{S}^k}\bar{w}^k_{PS}y^{kt}_S\leq 1; P\in\mathcal{Q}; t\in T \tag{4.3}$$

$$\sum_{t\in T}\sum_{k=1}^{i}\sum_{S\in\mathcal{S}^k}\bar{x}^k_{iS}y^{kt}_S\leq 1; i\in I \tag{4.4}$$

$$\sum_{i\in I}v^t_i\sum_{k=1}^{i}\sum_{S\in\mathcal{S}^k}\bar{x}^k_{iS}y^{kt}_S\geq(1-\Delta)\sum_{i\in I}v^{t-1}_i\sum_{k=1}^{i}\sum_{S\in\mathcal{S}^k}\bar{x}^k_{iS}y^{k,t-1}_S; t=2,...,|T| \tag{4.5}$$

$$\sum_{i\in I}v^t_i\sum_{k=1}^{i}\sum_{S\in\mathcal{S}^k}\bar{x}^k_{iS}y^{kt}_S\leq(1+\Delta)\sum_{i\in I}v^{t-1}_i\sum_{k=1}^{i}\sum_{S\in\mathcal{S}^k}\bar{x}^k_{iS}y^{k,t-1}_S; t=2,...,|T| \tag{4.6}$$

$$y^{kt}_S\in\{0,1\}; k\in I; t\in T; S\in\mathcal{S}^k \tag{4.7}$$

$$y^{kt}_0\in\{0,1\}; k\in I; t\in T. \tag{4.8}$$

We shall refer to this formulation as $\mathcal{S}$-knapsack-and-clique decomposition.

As for the knapsack formulation, one can consider the reformulation of the compact bucket model with enforcement of the connectivity of clearcuts, as well as the corresponding $\mathcal{R}$-knapsack-and-clique decomposition. It turns out that the corresponding master problem coincides with the so-called *cluster formulation* which has been considered in ([3],[8],[10],[9]).

The authors in [9] showed that the linear relaxation bound of the cluster formulation remains the same if besides connected clusters, disconnected clusters (buckets) are allowed in the model. This means that in this case the $\mathcal{S}$ and $\mathcal{R}$ knapsack-and-clique decompositions yield the same LP bounds.

The reformulated sets in the $\mathcal{S}$ or $\mathcal{R}$ knapsack-and-clique decompositions are contained in the reformulated sets in the corresponding knapsack decompositions, *i.e.* they are defined by a larger set of constraints. This means the LP bounds obtained by the knapsack-and-clique decompositions are never worse than those obtained by the corresponding knapsack decompositions, and it turns out that they are better in many instances (see Section 6).

From the above discussion we may state the following results:

Proposition 4.1. *The LP bound of the $\mathcal{S}$-knapsack-and-clique decomposition is equal to that of the $\mathcal{R}$-knapsack-and-clique decomposition.*

Proposition 4.2. *The LP bounds of the $\mathcal{S}$ or $\mathcal{R}$ knapsack-and-clique decompositions are better than or equal to those of the $\mathcal{S}$ and $\mathcal{R}$ knapsack decompositions.*

4.2 The pricing subproblem

For the linear relaxation of the master problem, constraints (4.7) and (4.8) are simply replaced by $y^{kt}_S\geq 0$ and $y^{kt}_0\geq 0$ respectively.

Let Ω^{kt}, β^t_P, θ_i, μ^t and ν^t denote the dual variables associated with constraints (4.2), (4.3), (4.4), (4.5) and (4.6) of the linear relaxation of the master problem and Ω^{kt*}, β^{t*}_P, θ^*_i, μ^{t*} and ν^{t*} assume an optimal dual solution of the linear relaxation of a restricted master problem. By definition of reduced cost, the objective function of the pricing subproblem kt is given by

$$\max_{S\in\mathcal{S}^k}\{\sum_{i\in S}\epsilon^{t*}_i+\sum_{P:S\cap P\neq\emptyset}\beta^{t*}_P\}+\Omega^{kt*}$$

where

$\epsilon^{1*}_i=p^1_i+\theta^*_i-\mu^{2*}(1-\Delta)v^1_i-\nu^{2*}(1+\Delta)v^1_i$

$\epsilon^{t*}_i=p^t_i+\theta^*_i+\mu^{t*}v^t_i+\nu^{t*}v^t_i-\mu^{t+1*}(1-\Delta)v^t_i-\nu^{t+1*}(1+\Delta)v^t_i$, $t=2,...,|T|-1$

$\epsilon^{|T|*}_i=p^{|T|}_i+\theta^*_i+\mu^{|T|*}v^{|T|}_i+\nu^{|T|*}v^{|T|}_i$ (with $\theta^*_i\leq 0$, $\mu^{t*}\geq 0$ and $\nu^{t*}\leq 0$).

The subproblem kt can be formulated by the following integer program. The variables will be

$$x_i=\begin{cases}1 & \text{if stand } i \text{ is selected to belong to region } S\\ 0 & \text{otherwise}; i=k,\ldots,|I|\end{cases}$$

$$w_P=\begin{cases}1 & \text{if at least one stand from clique } P \text{ is selected to belong to region } S\\ 0 & \text{otherwise}; P\in\mathcal{Q}:\max\limits_{i\in P}\{i\}\geq k.\end{cases}$$

The formulation is as follows:

$$\max \sum_{i=k}^{|I|} \epsilon_i^{t*} x_i + \sum_{P:\max_{i \in P}\{i\} \geq k} \beta_P^{t*} w_P + \Omega^{kt*} \tag{4.9}$$

subject to

$$x_i \leq w_P; i \geq k; P \in \mathcal{Q} : i \in P \tag{4.10}$$

$$\sum_{i=k+1}^{|I|} a_i x_i \leq A_{max} - a_k \tag{4.11}$$

$$x_k = 1 \tag{4.12}$$

$$x_i \in \{0,1\}; i = k, ..., |I| \tag{4.13}$$

$$w_P \geq 0; P \in \mathcal{Q} : \max_{i \in P}\{i\} \geq k. \tag{4.14}$$

The objective function (4.9) is to maximize the sum of the node weights ϵ_i^{t*} and the clique weights β_P^{t*} over the selected nodes and to add the value Ω^{kt*} to the optimal sum. Constraints (4.10) ensure that if a node is selected, then any maximal clique with this node is also selected. Constraint (4.12) guarantees that region S contains stand k. Constraints (4.11) ensure that the area of region S does not exceed A_{max}. Constraints (4.13) and (4.14) state the variable types.

5 Solving the subproblems

The subproblems for the $\mathcal{S}$-Knapsack decomposition were solved alternatively in two ways: using CPLEX 12.2 as an integer programming solver or with the Horowitz-Sahni exact solution algorithm [6], specific for the knapsack problem. For the subproblems of the $\mathcal{R}$-Knapsack decomposition, an algorithm for enumerating sets ([4]) was used. This algorithm generates, for each $k \in I$, all regions from set $\mathcal{R}^k$. The subproblems for the Knapsack-and-clique decompositions were solved using the integer programming solver ($\mathcal{S}$-Knapsack-and-clique decomposition) and the algorithm for enumerating sets ($\mathcal{R}$-Knapsack-and-clique decomposition).

6 Computational experience

The computational tests we performed were intended (i) to test the quality of the lower bounds provided by the proposed reformulations, and (ii) to test the behavior in terms of solution quality and speed of a straightforward heuristic (MIPHeur), that consists in solving the final restricted master problem of column generation by branch-and-bound.

Relatively to column generation, we began by generating all feasible clusters with two stands at the most. Then the linear relaxation of the reformulation was solved by adding meaningful variables to the linear relaxation of the restricted master problem. Concerning MIPHeur, the final restricted master problem, with all added variables, is exactly solved, enforcing the integrality constraints. This heuristic was allowed to run for two hours at the most, and the limit on the relative gap tolerance was set to 1%. Column generation and MIPHeur were performed with CPLEX 12.2 as linear and integer programming solver. CPLEX default parameters were used throughout. Computations were performed on a desktop computer with an Intel Core i7 - 2.8 GHz processor and with 8 GB RAM.

We report results for both real and hypothetical test forests (Table 2). Real test forest is Leiria National Forest (LNF) in Portugal. LNF and the hypothetical test forests F and G are referred to in [1]. We also report results for the hypothetical instance WLC which is referred to in [1] and partly available at the website www.unbf.ca/fmos/.

The quality of the integer solutions obtained by MIPHeur was measured by using the deviation (in percentage) of its value (vis) from the optimal value of the linear relaxation (bup): gap= $\frac{bup-vis}{vis}100$. For the set of all instances except LNF, the mean, (sample) standard deviation and coefficient of variation (ratio of the standard deviation to the mean) of the consumption times of column generation and MIPHeur are displayed in Table 3 (Knapsack decompositions) and Table 4 (Knapsack-and-clique decompositions). For the same set, Table 5 displays these measures for the gap of the solution obtained by MIPHeur in each decomposition. Note that, although the standard deviation measures the dispersion of the values relative

Table 2: Size of the instances ($\overline{a}_k$ is the average of the stand's area).

Instance	No. nodes	No. edges	No. cliques	$\overline{a}_k$ (ha)	$A_{\max}/\overline{a}_k$	$\|T\|$
F10x10: -3, -4	100	180	180	1	3, 4	7
F15x15: -3, -4	225	420	420	1	3, 4	7
F20x20: -3, -4	400	760	760	1	3, 4	7
F25x25: -3, -4, -5, -6	625	1200	1200	1	3, 4, 5, 6	7
G15x7: -3, -4	105	247	145	1	3, 4	7
G40x10: -3, -4	400	988	596	1	3, 4	7
G40x14: -3, -4	560	1396	844	1	3, 4	7
G60x10: -3, -4, -5, -6	600	1492	904	1	3, 4, 5, 6	7
LNF	574	1152	740	14.96	3.46	6
WLC	73	98	63	10.12	3.95	7

to the mean, the coefficient of variation is useful in comparing decompositions which differ considerably in magnitude for the values. The real instance is referred to separately in all tables.

Table 3: Consumption time of column generation and MIPHeur, for the Knapsack decompositions.

Instance	Time (sec.)	$\mathcal{S}$-knap. CPLEX CG	MIPHeur	Horowitz-Sahni CG	MIPHeur	$\mathcal{R}$-knap. Enumeration CG	MIPHeur
All except LNF	Mean (sec.)	189.2	5476.7	30.8	5668.7	9.8	4850.8
	Standard deviation (sec.)	236.39	1513.93	68.00	1567.64	14.84	1401.3
	Coeficient of variation	1.250	0.276	2.210	0.277	1.518	0.289
LNF		747.8	212.8	28.0	183.4	8.7	27.4

Table 4: Consumption time of column generation and MIPHeur, for the Knapsack-and-clique decompositions.

Instance	Time (sec.)	$\mathcal{S}$-knapcliq. CPLEX CG	MIPHeur	$\mathcal{R}$-knapcliq. Enumeration CG	MIPHeur
All except LNF	Mean (sec.)	42.8	6897.9	1.7	6896.7
	Standard deviation (sec.)	41.80	1116.44	1.92	1379.73
	Coeficient of variation	0.977	0.162	1.155	0.200
LNF		136.8	20.9	5.3	25.2

Table 5: Quality of the integer solutions obtained by MIPHeur, for all decompositions.

Instance	Gap (%)	$\mathcal{S}$-knap. CPLEX	Horowitz-Sahni	$\mathcal{R}$-knap. Enumeration	$\mathcal{S}$-knapcliq. CPLEX	$\mathcal{R}$-knapcliq. Enumeration
All except LNF	Mean (%)	0.98	0.82	0.90	0.47	0.46
	Standard deviation (%)	0.674	0.575	0.584	0.428	0.484
	Coeficient of variation	0.690	0.700	0.652	0.906	1.051
LNF		0.03	0.03	0.01	0.01	0.01

Relatively to the consumption time of column generation, the decompositions with connectivity show smaller mean values than the others. Considering MIPHeur, the smallest mean value is observed with $\mathcal{R}$-knapsack decomposition. Note that the respective coefficients of variation along the decompositions are not quite different. For LNF, the discussion related to column generation is the same. Relatively to MIPHeur, the best values are obtained with $\mathcal{S}$-knapsack-and-clique decomposition and the decompositions with connectivity.

Regarding the gap of the solution obtained by MIPHeur, the knapsack-and-clique decompositions, and also the $\mathcal{R}$-knapsack decomposition for the LNF, provide the best (mean) values. This result is primarily due to the increasing of the net present value of the feasible solutions. The decreasing of the LP bound only occurs with five instances (including LNF). The solution is, in average, within 1% of the optimum with all decompositions, and for LNF, the best gap is 0.01%. Hence all decompositions described in this work seem to provide tight LP bounds on the optimal values and first high-quality feasible solutions.

7 Conclusions

In this paper we considered the so-called bucket formulation for the harvest scheduling problem with maximum area restrictions, and presented two Dantzig-Wolfe decompositions (the *knapsack* and the *knapsack-and-clique* decompositions), and two variants of each one ($\mathcal{S}$ and $\mathcal{R}$). The $\mathcal{S}$-knapsack decomposition yields to a new formulation for the problem, while the knapsack-and-clique reformulations correspond to the so-called cluster formulation for the problem. We showed theoretically that the bounds provided by the knapsack-and-clique reformulations are better than or equal to those obtained by the knapsack reformulations. Computationally, the linear programming relaxations of both reformulations provide very tight bounds, and a heuristic proposed - MIPHeur - is able to obtain feasible solutions within 1% of the optimum in average, with a slight advantage for the knapsack-and-clique decomposition.

While these results show the Dantzig-Wolfe decompositions proposed are competitive with other approaches, including the compact formulation, for medium size instances ([1], [4]), further tests are needed to assess its ability to solve very large scale instances (more than 5000 stands).

Acknowledgement This work have been partially funded by FCT (Fundação para a Ciência e a Tecnologia - Portugal) through project "SearchCol: Metaheuristic search by column generation" (PTDC/EIA-EIA/100645/2008).

References

[1] M. Constantino, I. Martins, and J. G. Borges: A New Mixed-Integer Programming Model for Harvest Scheduling Subject to Maximum Area Restrictions, Operations Research, Vol. 56, No. 3, 2008, pp. 542-551.

[2] R. Diestel: Graph Theory, Graduate texts in Mathematics, 2nd Edition, Springer, 2000.

[3] M. Goycoolea, A.T. Murray, F. Barahona, R. Epstein and A. Weintraub: Harvest scheduling subject to maximum area restrictions: exploring exact approaches, Operations Research, Vol. 53, No. 3, 2005, pp. 90-500.

[4] M. Goycoolea, A.T. Murray, J.P. Vielma and A. Weintraub: Evaluating Approaches for Solving the Area Restricted Model in Harvest Scheduling, Forest Science, Vol. 55, No. 2, 2009, pp. 149-165.

[5] ILOG, ILOG CPLEX 12.2 - User's Manual, 2010.

[6] S. Martello, P. Toth, Knapsack Problems, Wiley, 1990.

[7] I. Martins, M. Constantino and J.G. Borges: Forest management models with spatial structure constraints, Working Paper No. 2/1999, C.I.O., Faculty of Sciences, University of Lisbon, 1999.

[8] I. Martins, M. Constantino and J.G. Borges: A column generation approach for solving a non-temporal forest harvest model with spatial structure constraints, European Journal of Operational Research, Vol. 161, No. 2, 2005, pp. 478-498.

[9] I. Martins, F. Alvelos and M. Constantino: A branch-and-price approach for harvest scheduling subject to maximum area restrictions, Computational Optimization and Applications, Vol. 51, 2012, pp. 363-385.

[10] J.P. Vielma, A.T. Murray, D.M. Ryan and A. Weintraub: Improving computational capabilities for adressing volume constraints in forest harvest scheduling problems, European Journal of Operational Research, Vol. 176, No. 2, 2007, pp. 1246-1264.

Numerical Experiments with a Modified Regularization Scheme for Mathematical Programs with Complementarity Constraints

Teófilo Miguel M. Melo*, João Luís H. Matias†, M. Teresa T. Monteiro‡

*CIICESI, School of Technology and Management of Felgueiras, Polytechnic of Porto, Portugal,
E-mail:tmm@estgf.ipp.pt

† Centre of Mathematics, CM-UTAD, University of Trás-os-Montes e Alto Douro, Portugal,
E-mail:j_matias@utad.pt

‡ R&D Centre Algoritmi, Department of Production and Systems, University of Minho, Portugal,
E-mail:tm@dps.uminho.pt

Abstract

On this paper we present a modified regularization scheme for Mathematical Programs with Complementarity Constraints. In the regularized formulations the complementarity condition is replaced by a constraint involving a positive parameter that can be decreased to zero. In our approach both the complementarity condition and the nonnegativity constraints are relaxed. An iterative algorithm is implemented in MATLAB language and a set of AMPL problems from MacMPEC database were tested.

Keywords: complementarity constraints, regularization scheme, SQP

1 Introduction

Mathematical Programs with Complementarity Constraints (MPCC) is an exciting application of nonlinear programming techniques. These kind of constraints may come in the form of a game, a variational inequality or as stationary conditions of an optimization problem. The main applications areas are Engineering and Economics [Ferris and Pang, 1997], [Outrata et al., 1998]. They are so pervasive in these areas because the concept of complementarity is tantamount with the notion of system equilibrium. They are very difficult to solve as the usual constraint qualifications necessary to guarantee the algorithms convergence fail in all feasible points [Chen and Florian, 1995]. This complexity is caused by the disjunctive nature of the complementarity constraints, from a geometric point of view, its feasible region is not convex nor generally connected.

There have been proposed some nonlinear approaches to solve MPCC, starting with the smoothing scheme [Facchinei et al., 1996] and [Fukushima and Pang, 1999], the regularization scheme [Scholtes, 2001], the modified relaxation scheme [Lin and Fukushima, 2005] and [Demiguel et al., 2005] where a two-sided relaxation scheme is presented. While on this work we focus on a regularization scheme to solve MPCC, there are other nonlinear approaches. For example the penalty approaches [Hu and Ralph, 2004], [Ralph and Wright, 2004], [Monteiro and Meira, 2011], [Monteiro and Rodrigues, 2011] and the "elastic mode" for nonlinear programming in conjunction with a sequential quadratic programming (SQP) algorithm [Anitescu, 2005]. We also emphasize the work [Raghunathan and Biegler, 2003] that uses interior point methods. In [Fletcher et al., 2006], SQP is guaranteed, to under relatively mild conditions, quadratically converge near a stationary point.

This paper is organized as follows. Next section defines the MPCC problem and presents our modified regularization scheme. Some concepts related to the optimality conditions are presented in Section 3. The implemented algorithm in MATLAB environment is detailed in Section 4. Numerical experiments to test the algorithm, are presented in Section 5. Some conclusions and future work ideas are carried out in Section 6.

2 Problem definition

We consider Mathematical Program with Complementarity Constraints (MPCC) as:

$$\begin{array}{ll} \min\limits_{x} & f(x) \\ \text{subject to} & g(x) \geq 0,\ h(x) = 0, \\ & 0 \leq G(x) \perp H(x) \geq 0, \end{array} \tag{2.1}$$

where $f : \mathbb{R}^n \to \mathbb{R}, g : \mathbb{R}^n \to \mathbb{R}^m, h : \mathbb{R}^n \to \mathbb{R}^p, G : \mathbb{R}^n \to \mathbb{R}^q, H : \mathbb{R}^n \to \mathbb{R}^q$ are all twice continuously differentiable functions. The notation $G(x) \perp H(x)$, means that $G(x)^T H(x) = 0$, due the complementarity nature of the complementarity constraints or $G_i(x) = 0$ or $H_i(x) = 0$, $i = 1, \dots, q$. One attractive way of solving (2.1) is to replace the complementarity constraints by a set of nonlinear inequalities, such as $G_i(x)H_i(x) \leq 0$, and then solve the equivalent nonlinear program (NLP):

$$\begin{array}{ll} \min\limits_{x} & f(x) \\ \text{subject to} & g(x) \geq 0,\ h(x) = 0, \\ & G(x) \geq 0,\ H(x) \geq 0, \\ & G_i(x)H_i(x) \leq 0, \\ & i = 1, 2, \dots, q. \end{array} \tag{2.2}$$

The major difficulty in solving (2.2) is that its constraints fail to satisfy the Mangasarian Fromovitz constraint qualification (MFQC) at any feasible point [Scholtes, 2001].
In this paper, we propose the following modified regularization scheme to solve the problem (2.1):

$$\begin{array}{ll} \min\limits_{x} & f(x) + \rho t \\ \text{subject to} & g(x) \geq 0,\ h(x) = 0, \\ & (G_i(x) + t)(H_i(x) + t) \geq t^2 \\ & (G_i(x) - t)(H_i(x) - t) \leq t^2 \\ & i = 1, 2, \dots, q, \end{array} \tag{2.3}$$

where ρ is a parameter to penalize the relaxation parameter t. This reformulation has less constraints than problem (2.2). Figure 1 represents the feasible region of the complementarity constraints for the relaxation scheme proposed.

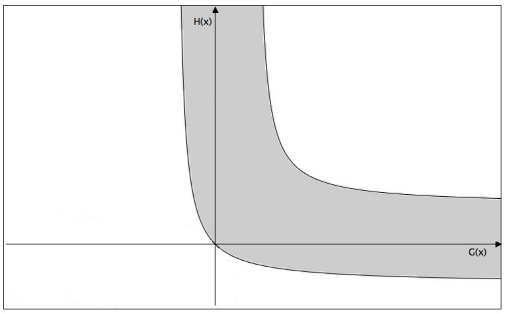


Figure 1: Feasible region of the complementarity constraints.

The novelty of our approach is the combination of [Lin and Fukushima, 2005] and [Kadrani et al., 2009] strategies also penalizing the regularization parameter t. This is similar to the so-called elastic mode, but in our work the complementarity constraints are maintained as constraints of problem. In the elastic mode strategy, the complementarity constraints are removed from the set of constraints and are included in the penalty function as a penalty term.

3 Optimal issues

Recent developments [Anitescu et al., 2006] show that a SQP method with an elastic mode is used to solve MPCC and there is a relationship between strong stationarity defined by [Scheel and Scholtes, 2000] and the Karush-Kuhn-Tucker (KKT) points. This relationship establishes convergence of SQP methods for MPCC formulated as NLP. Concepts like, stationarity, constraints qualification (LICQ - linear independence constraint qualification) and second order conditions (SOSC - second order sufficient condition) of the MPCC problem, will be defined in terms of the relaxed nonlinear program (RNLP) for (2.1) as follows:

$$\begin{array}{ll} \min\limits_{x} & f(x) \\ \text{subject to} & g(x) \geq 0,\ h(x) = 0, \\ & G_i(x) \geq 0,\ i \in I_G \setminus I_H, \\ & H_i(x) \geq 0,\ i \in I_H \setminus I_G, \\ & G_i(x) \geq 0,\ H_i(x) \geq 0,\ i \in I_G \cap I_H, \end{array} \tag{3.1}$$

where I_g, I_G and I_H are the following active sets at the point x^* feasible for (2.1):

$$\begin{array}{l} I_g = \{i \in \{1,2,\ldots,m\} | g_i(x^*) = 0\}, \\ I_G = \{i \in \{1,2,\ldots,q\} | G_i(x^*) = 0\}, \\ I_H = \{i \in \{1,2,\ldots,q\} | H_i(x^*) = 0\}. \end{array}$$

For a general constraints vector $v \in \mathbb{R}^n$ we consider a matrix ∇v containing the constraints gradients along the columns. The optimality concepts follow [Ralph and Wright, 2004] and the corresponding proves can be consulted in this work. There are several kinds of stationarity defined for MPCC problem, among them, the strong stationarity is the following one.

Definition 1. A point x^* that is feasible for (2.1) is strongly stationary if $d = 0$ ($d \in \mathbb{R}^n$) solves the following linear program:

$$\begin{array}{ll} \min\limits_{d} & \nabla f(x^*)^T d \\ \text{subject to} & g(x^*) + \nabla g(x^*)^T d \geq 0,\ h(x^*) + \nabla h(x^*)^T d = 0, \\ & \nabla G_i(x^*)^T d = 0,\ i \in I_G \setminus I_H, \\ & \nabla H_i(x^*)^T d = 0,\ i \in I_H \setminus I_G, \\ & \nabla G_i(x^*)^T d \geq 0,\ \nabla H_i(x^*)^T d \geq 0,\ i \in I_G \cap I_H. \end{array} \tag{3.2}$$

Combining the optimality conditions for (3.2) with the feasibility conditions for x^*, we obtain:

$$0 = \nabla f(x^*) - \sum_{i \in I_g} \lambda_i^* \nabla g_i(x^*) - \sum_{i=1}^{p} \mu_i^* \nabla h_i(x^*) - \sum_{i \in I_G} \tau_i^* \nabla G_i(x^*) - \sum_{i \in I_H} \nu_i^* \nabla H_i(x^*) \tag{3.3}$$

$$\begin{array}{l} 0 = h_i(x^*),\ i = 1,2,\ldots,p, \\ 0 = g_i(x^*),\ i \in I_g, \\ 0 < g_i(x^*),\ i \in \{1,2,\ldots,m\} \setminus I_g, \\ 0 \leq \lambda_i^*,\ i \in I_g, \\ 0 = G_i(x^*),\ i \in I_G, \\ 0 < G_i(x^*),\ i \in \{1,2,\ldots,q\} \setminus I_G, \\ 0 = H_i(x^*),\ i \in I_H, \\ 0 < H_i(x^*),\ i \in \{1,2,\ldots,q\} \setminus I_H, \\ 0 \leq \tau_i^*,\ i \in I_G \cap I_H, \\ 0 \leq \nu_i^*,\ i \in I_G \cap I_H. \end{array} \tag{3.4}$$

Definition 2. The MPCC-LICQ is satisfied at the point x^* if the following set of vectors is linear independent:

$$\{\nabla g_i(x^*) | i \in I_g\} \cup \{\nabla h_i(x^*) | i = 1,2,\ldots,p\} \cup \{\nabla G_i(x^*) | i \in I_G\} \cup \{\nabla H_i(x^*) | i \in I_H\}.$$

The linear independence constraint qualification (LICQ) is satisfied for RNLP (3.1).

For a strongly stationary point x^* and for some $(\lambda_i^*, \mu_i^*, \tau_i^*, \nu_i^*)$ satisfying (3.3-3.4) one defines the following sets:

$$\begin{array}{l} I_g^+ = \{i \in I_g | \lambda_i^* > 0\}, \\ I_g^0 = I_g \setminus I_g^+, \\ J_G^+ = \{i \in I_G \cap I_H | \tau_i^* > 0\}, \\ J_G^0 = (I_G \cap I_H) \setminus J_G^+, \\ J_H^+ = \{i \in I_G \cap I_H | \nu_i^* > 0\}, \\ J_H^0 = (I_G \cap I_H) \setminus J_H^+. \end{array}$$

The set $\bar{S}$ of normalized critical directions for the RNLP (3.1) is defined as follows ($s \in \mathbb{R}^n$):

$$\bar{S} = \{s | \|s\|_2 = 1\} \cap \{s | \nabla h(x^*)^T s = 0\} \cap \{s | \nabla g_i(x^*)^T s = 0, i \in I_g^+\} \cap \{s | \nabla g_i(x^*)^T s \geq 0, i \in I_g^0\} \cap$$

$$\cap \{s | \nabla G_i(x^*)^T s = 0, i \in I_G \setminus I_H\} \cap \{s | \nabla G_i(x^*)^T s \geq 0, i \in J_G^0\} \cap \{s | \nabla G_i(x^*)^T s = 0, i \in J_G^+\} \cap$$

$$\cap \{s | \nabla H_i(x^*)^T s = 0, i \in I_H \setminus I_G\} \cap \{s | \nabla H_i(x^*)^T s \geq 0, i \in J_H^0\} \cap \{s | \nabla H_i(x^*)^T s = 0, i \in J_H^+\}.$$

The set of normalized critical directions S^* for (2.1) is:

$$S^* = \bar{S} \cap \{s | \min(\nabla H_i(x^*)^T s, \nabla G_i(x^*)^T s) = 0, i \in J_G^0 \cap J_H^0\}$$

This is obtained by enforcing the additional condition that either $\nabla H_i(x^*)^T s = 0$ or $\nabla G_i(x^*)^T s = 0$, $i \in J_G^0 \cap J_H^0$.

Definition 3. Let x^* be a strongly stationary point. The MPCC-SOSC holds at x^* if there is $\sigma > 0$ such that for every $s \in S^*$, there are multipliers $(\lambda_i^*, \mu_i^*, \tau_i^*, \nu_i^*)$ satisfying (3.3-3.4) such that

$$s^T \nabla_{xx}^2 L(x^*, \lambda_i^*, \mu_i^*, \tau_i^*, \nu_i^*) s \geq \sigma.$$

The RNLP-SOSC holds at x^* if for every $s \in \bar{S}$, there are $(\lambda_i^*, \mu_i^*, \tau_i^*, \nu_i^*)$ such that (3.3) holds.

As theoretical support, we summarized some known results based on [Ralph and Wright, 2004] concerning constraint qualifications and first and second order optimality conditions of MPCC. In the same work some properties of regularization schemes are presented: estimating distance between their solutions and the MPCC optimum, boundedness of their Lagrange multipliers and local uniqueness of their solutions. Some alternative regularized formulations are also studied. A penalty approach similar to the elastic mode from [Anitescu et al., 2006] is also analyzed.

Based on these ideas and on [Lin and Fukushima, 2005] and [Kadrani et al., 2009], a computational implementation of a modified regularized scheme was developed. Details of the corresponding algorithm are in next section.

4 Algorithm details

An algorithm was implemented (Algorithm 1) to iteratively solve problem (2.3) for specific values of t and ρ, with $t \to 0$ and $\rho \to \infty$. This algorithm has two iterative procedures, the inner one is performed by `fmincon` routine from MATLAB Optimization toolbox, that uses the SQP method.

Algorithm 1 Modified regularization scheme

Take initial values x_0, $t_0 > 0$, $\rho_0 > 0$ and tolerances ϵ_1, ϵ_2.
for $k = 0, 1, 2, \ldots$ **do**
    Solve the minimization problem (2.3) with x_k, t_k and ρ_k obtaining x_{k+1}.
    if $\|\nabla L(x_{k+1}, \ldots)\| \leq \epsilon_1$ and $\|f(x_{k+1}) - f(x_k))\| \leq \epsilon_2$ **then**
        STOP.
    else
        $\rho_{k+1} = r_1 \rho_k,\ r_1 > 1.$
        $t_{k+1} = r_2 t_k,\ 0 < r_2 < 1.$
    end if
end for

To evaluate the stop criterium in the algorithm, we consider the following equality in the solution x^*:

$$\nabla L(x^*, \delta, \gamma, \xi, \zeta) = \nabla f(x^*) - \sum_{i=1}^{m} \delta_i \nabla g_i(x^*) - \sum_{i=1}^{p} \gamma_i \nabla h_i(x^*) - \sum_{i=1}^{q} \xi_i \nabla \Phi_{t,i}(x^*) + \sum_{i=1}^{q} \zeta_i \nabla \Psi_{t,i}(x^*)$$

where for $i = 1, 2, \ldots q$ and $x \in \mathbb{R}^n$,

$$\Phi_{t,i}(x) = (G_i(x) + t)(H_i(x) + t) - t^2$$

$$\Psi_{t,i}(x) = (G_i(x) - t)(H_i(x) - t) - t^2,$$

we have,

$$\nabla\Phi_{t,i}(x) = (G_i(x) + t)\nabla H_i(x) + (H_i(x) + t)\nabla G_i(x),$$

$$\nabla\Psi_{t,i}(x) = (G_i(x) - t)\nabla H_i(x) + (H_i(x) - t)\nabla G_i(x).$$

The Lagrange multipliers δ, γ, ξ and ζ are an output of the `fmincon` routine from MATLAB. The tolerances used in the stop criterium are $\epsilon_1 = \epsilon_2 = 10^{-4}$. We consider $r_1 = 10$ and $r_2 = 0.1$ in the parameters update scheme. The initial choices, $t_0 = 0.25$ and $\rho_0 = 1$ were considered. Next section reports the numerical results using 70 test problems.

5 Numerical results

This section describes the experiments with an implementation of our modified regularization scheme for problem (2.1). The computational experiments were made on a *2.26 GHz Intel Core 2 Duo* with 8GB of RAM, MAC OS 10.6.8 operating system. The MATLAB version used was 7.11.0 (R2010b). The `fmincon` routine is connected to the modeling language AMPL [Fourer and Kernighan, 1993] by a MATLAB mex interface. The test problems in Table 1 are from MacMPEC database [Leyffer, 2000].

In order to evaluate and compare the performance of the Algorithm 1, we implemented another regularized formulation (Algorithm 2) proposed in [Lin and Fukushima, 2005], using the same conditions. Table 2 reports the numerical results of both algorithms. The first column indicates the test problem, next six columns refer to Algorithm 1: f^* shows the final objective function value, $\|\nabla L\|$ presents the norm of the Lagrangian function of problem (2.3), `int` presents the number of internal iterations performed by the `fmincon` routine from Matlab, `ext` shows the number of external iterations, the last two columns report the number of function evaluations and the exit status, respectively. The last six columns present the corresponding results from Algorithm 2. Table 3 summarizes the exit status of the algorithms.

Table 1: Test problems

Problem	n	m	p	q	Problem	n	m	p	q
bar-truss-3	35	6	28	6	gnash16	13	8	4	8
bard1	5	3	1	3	gnash17	13	8	4	8
bard3	6	2	3	1	gnash18	13	8	4	8
bard1m	6	3	1	3	gnash19	13	8	4	8
bard3m	6	4	1	3	hs044-i	20	10	4	10
bilevel3	11	4	6	3	jr1	2	1	0	1
dempe	3	1	1	1	jr2	2	1	0	1
desilva	6	2	2	2	kth1	2	1	0	1
df1	2	3	0	1	kth2	2	1	0	1
ex9.1.1	13	5	7	5	kth3	2	1	0	1
ex9.1.2	8	2	5	2	nash1a	6	2	2	2
ex9.1.3	23	6	15	6	nash1b	6	2	2	2
ex9.1.4	8	2	5	2	nash1c	6	2	2	2
ex9.1.5	13	5	7	5	nash1d	6	2	2	2
ex9.1.6	14	6	7	6	portfl-i-1	87	12	13	12
ex9.1.7	17	6	9	6	portfl-i-2	87	12	13	12
ex9.1.8	11	4	5	3	portfl-i-3	87	12	13	12
ex9.1.9	12	5	6	5	portfl-i-4	87	12	13	12
ex9.1.10	11	4	5	3	portfl-i-6	87	12	13	12
ex9.2.1	10	4	5	4	qpec1	30	20	0	20
ex9.2.2	9	4	4	3	qpec2	30	20	0	20
ex9.2.4	8	2	5	2	ralph1	2	1	0	1
ex9.2.6	16	6	6	6	ralph2	2	1	0	1
ex9.2.7	10	4	5	4	ralphmod	104	100	0	100
ex9.2.8	6	2	3	2	scholtes1	3	1	0	1
ex9.2.9	9	3	5	3	scholtes2	3	1	0	1
flp2	4	2	0	2	scholtes3	3	1	0	1
flp4-1	80	60	0	30	scholtes4	3	1	0	1
flp4-2	110	110	0	60	scholtes5	3	1	0	1
flp4-3	140	170	0	70	scale1	2	1	0	1
flp4-4	200	250	0	100	scale2	2	1	0	1
gnash10	13	8	4	8	scale3	2	1	0	1
gnash11	13	8	4	8	scale5	2	1	0	1
gnash12	13	8	4	8	stackelberg1	3	1	1	1
gnash15	13	8	4	8	taxmcp	15	11	3	11

The algorithms present similar behaviour with respect to the internal and external iterations, to the function evaluations and to the solution accuracy. However Algorithm 2 does not solve eight problems, whereas Algorithm 1 only fails in three problems. Some of them are ill-posed, for instance, `ralph1`, `ralphmod` and `scholtes4`. The solutions obtained by both algorithms are similar to the ones reported in MacMPEC database with good accuracy.

6 Conclusions and future work

An iterative algorithm in MATLAB language to solve MPCC was implemented. The algorithm aims to compute a local optimal solution joining a modified regularization scheme and the SQP strategy. The algorithm is still in an improvement phase but some conclusions can already be taken: the promising

Table 2: Results of Algorithm 1 and Algorithm 2

Problem	f^*	$\|\nabla L\|$	int	ext	nfe	flag	f^*	$\|\nabla L\|$	int	ext	nfe	flag
bar-truss-3	10166,512	1,36E-05	39	5	1689	1	10166,5519	2,57E-05	61	5	2606	1
bard1	17	1,71E-07	25	5	210	1	16,9999952	2,46E-07	13	3	114	1
bard3	-12,679	1,49E-15	3	2	38	1	-12,678711	1,16E-06	3	2	38	1
bard1m	17	6,77E-08	21	5	208	1	16,9999952	1,90E-06	17	3	162	1
bard3m	-12,679	1,28E-07	19	5	187	1	-12,678722	1,33E-07	17	3	157	1
bilevel3	-12,679	1,04E-07	31	5	469	1	-12,678752	8,77E-08	29	5	438	1
dempe	28,25	2,47E-07	53	5	358	1	28,2500113	2,82E-07	45	2	311	1
desilva	-1	2,71E-08	3	2	38	1	-1	1,75E-08	3	2	38	1
df1	0	2,98E-08	5	2	28	1	0	1,49E-08	103	100	429	0
ex9.1.1	-13	2,82E-08	57	5	925	1	-13,000062	4,22E-08	50	5	813	1
ex9.1.2	-6,25	1,67E-08	23	4	257	1	-6,2500625	1,26E-08	14	4	167	1
ex9.1.3	-29,2	1,05E-05	46	6	1294	1	-29,20002	9,05E-07	39	3	1047	1
ex9.1.4	-37	4,00E-08	48	4	550	1	-37,000078	9,10E-09	44	4	494	1
ex9.1.5	-1	1,33E-08	62	6	1064	1	-1,0000003	3,90E-06	52	7	885	1
ex9.1.6	-49	7,57E-07	90	4	1573	1	-49,00035	3,73E-08	84	4	1558	1
ex9.1.7	-26	7,03E-05	69	6	1432	1	-26,000065	9,99E-08	58	6	1249	1
ex9.1.8	-3,25	3,01E-16	7	2	115	1	-3,25	0,00E+00	9	3	141	1
ex9.1.9	3,111	3,20E-09	80	6	1444	1	3,11108333	1,08E-07	80	5	1385	1
ex9.1.10	-3,25	3,01E-16	7	2	115	1	-3,25	0,00E+00	9	3	141	1
ex9.2.1	17	1,56E-07	25	5	355	1	16,9999952	1,28E-06	14	3	201	1
ex9.2.2	100	2,60E-06	40	5	490	1	99,9998557	5,67E-06	37	5	457	1
ex9.2.4	0,5	1,53E-08	25	6	295	1	0,49999994	9,69E-07	15	4	186	1
ex9.2.6	-1	3,34E-08	17	6	374	1	-1,0000001	1,40E-06	12	4	268	1
ex9.2.7	17	1,56E-07	25	5	355	1	16,9999952	1,28E-06	14	3	201	1
ex9.2.8	1,5	3,52E-09	11	5	123	1	1,49999997	2,90E-09	9	4	93	1
ex9.2.9	2	7,68E-10	10	2	131	1	2	0,00E+00	9	2	119	1
flp2	0	5,03E-07	19	2	124	1	0	1,53E-07	56	17	415	1
flp4-1	0	1,32E-25	2	2	326	1	0	0,00E+00	4	4	571	1
flp4-2	0	6,17E-25	2	2	446	1	0	0,00E+00	3	3	558	1
flp4-3	0	7,02E-16	2	2	566	1	0	0,00E+00	3	3	708	1
flp4-4	0	8,01E-25	2	2	806	1	0	0,00E+00	3	3	1008	1
gnash10	-230,824	3,39E-07	27	5	475	1	-230,82343	1,47E-07	14	2	238	1
gnash11	-129,913	2,19E-07	28	5	490	1	-129,91193	3,60E-06	17	3	297	1
gnash12	-36,933	2,30E-07	29	5	505	1	-36,933107	3,59E-06	15	3	267	1
gnash15	-354,7	4,00E-07	41	5	685	1	-256,38936	2,90E-07	69	8	1322	1
gnash16	-241,442	4,22E-07	44	5	737	1	-129,91192	3,70E-05	87	9	2000	1
gnash17	-90,749	2,45E-07	52	5	855	1	-36,933107	1,02E-05	78	9	1981	1
gnash18	-25,698	4,25E-07	58	5	941	1	-25,698216	1,01E-06	29	4	497	1
gnash19	-6,117	4,42E-07	62	6	1018	1	-	-	-	-	-	-
hs044-i	15,618	1,88E-06	39	6	984	1	15,617653	4,24E-06	36	5	897	1
jr1	0,5	2,64E-07	29	6	138	1	0,4999975	2,14E-08	30	6	142	1
jr2	0,5	1,83E-07	27	6	129	1	0,49999988	4,63E-07	18	4	88	1
kth1	0	4,71E-16	3	2	18	1	-3,548E+15	1,41E+00	124	100	821	0
kth2	0	1,63E-08	5	2	26	1	0	1,07E-08	4	2	19	1
kth3	0,5	1,76E-07	26	6	122	1	0,49999994	5,20E-08	22	4	101	1
nash1a	0	2,87E-07	6	1	55	1	0	2,36E-07	19	5	181	1
nash1b	0	1,32E-07	17	2	150	1	0	2,36E-07	21	5	196	1
nash1c	0	4,68E-07	19	2	167	1	0	2,36E-07	20	5	188	1
nash1d	0	5,24E-07	16	2	142	1	0	2,40E-07	21	5	190	1
portfl-i-1	0	2,55E-07	12	2	1244	1	1,424E-05	1,26E-06	43	9	4379	1
portfl-i-2	0	7,72E-07	11	2	1156	1	1,4573E-05	2,54E-07	37	9	4004	1
portfl-i-3	0	6,17E-07	11	2	1155	1	0	6,10E-07	11	2	1155	1
portfl-i-4	0	6,54E-07	11	2	1156	1	0	3,85E-07	10	3	1066	1
portfl-i-6	0	1,75E-06	13	2	1334	1	2,3417E-06	3,86E-07	40	10	4112	1
qpec1	80	1,72E-07	8	2	318	1	80	8,64E-07	9	2	350	1
qpec2	45	6,33E-07	62	6	2143	1	44,9999	3,51E-06	61	6	2110	1
ralph1	0	0	52	7	229	2	-6,475E-07	0,00E+00	309	100	1536	0
ralph2	0	2,88E-07	38	5	167	1	-1E-10	9,10E-07	544	100	2476	0
ralphmod	-683,033	6,69E-05	282	10	31101	2	-683,03302	8,49E-05	182	10	20040	1
scholtes1	2	4,55E-08	9	2	53	1	2	5,43E-08	8	2	44	1
scholtes2	15	1,33E-07	12	2	61	1	15	1,19E-07	8	2	49	1
scholtes3	0,5	1,68E-08	39	6	177	1	0,49999994	5,77E-08	33	4	145	1
scholtes4	0	0	58	7	318	2	-3,694E-07	0,00E+00	312	100	1960	0
scholtes5	1	1,46E-06	30	6	176	1	0,999995	1,46E-06	30	6	176	1
scale1	1	2,69E-06	36	10	212	1	0,99999925	2,23E+00	123	100	2626	0
scale2	1	1,75E-06	24	6	145	1	0,99999988	1,49E-06	17	4	94	1
scale3	1	3,67E-05	35	8	258	1	1	1,49E-04	131	100	930	0
scale5	99,999	4,08E-05	34	6	209	1	99,9999875	4,60E-05	27	4	146	1
stackelberg1	-3266,672	3,75E-06	16	4	96	1	-3266,6725	2,11E-06	16	4	96	1
taxmcp	0,819	1,29E-08	31	6	623	1	0,81870494	8,65E-07	29	8	591	1

Table 3: Algorithms exit flags

flag	status
0	Terminated by iteration limit (100)
1	Found an optimal solution of MPCC
2	`fmincon` stopped because step length to small (10^{-6})

numerical results present good accuracy of the solutions when compared with the ones provided from the MacMPEC test problem database. The implemented approach proved to be competitive when compared with other regularized scheme proposed in literature.

As future work, it is intended to study what kind of stationary point is achieved by this regularization scheme. Another future work idea is to implement other schemes to update the regularization and penalization parameters, based on a feasibility test.

Acknowledgement

This work is funded by FEDER funds through Operational Programme for Competitiveness Factors - COMPETE and National Funds through FCT - Foundation for Science and Technology in Project scope: FCOMP-01-0124-FEDER-022674. The work is also supported by PEst-OE/MAT/UI4080/2011. The authors are very grateful to all the financial support and to the referees for valuable remarks and comments, which significantly contributed to the quality of the paper.

References

[Anitescu, 2005] Anitescu, M. (2005). On using the elastic mode in nonlinear programming approaches to mathematical programs with complementarity constraints. *SIAM Journal on Optimization*, 15(4):1203–1236.

[Anitescu et al., 2006] Anitescu, M., Tseng, P. and Ralph, D. (2006). Elastic mode algorithms for mathematical programs with equilibrium constraints: global convergence and stationarity proprieties. *Mathematical Programming*, 110(1):337–371.

[Chen and Florian, 1995] Chen, X. and Florian, M. (1995). The nonlinear bilevel programming problem: formulations, regularity and optimality conditions. *Optimization*, 32(4):193–209.

[Demiguel et al., 2005] Demiguel, V., Friedlander, M., Nogales, F. and Scholtes, S. (2005). A two-sided relaxation scheme for mathematical programs with equilibrium constraints. *SIAM Journal on Optimization*, 16(2):587–609.

[Facchinei et al., 1996] Facchinei, F., Jiang, H. and Qi, L. (1996). A smoothing method for mathematical programs with equilibrium constraints. Technical Report 03.96, Universita di Roma, "La Sapienza".

[Ferris and Pang, 1997] Ferris, M. and Pang, J. (1997). Engineering and economic applications of complementarity constraints. *SIAM Review*, 39(4):669–713.

[Fletcher et al., 2006] Fletcher, R., Leyffer, S., Ralph, D. and Scholtes, S. (2006). Local convergence of SQP methods for mathematical programs with equilibrium constraints. *SIAM Journal on Optimization*, 17(1):259–286.

[Fourer and Kernighan, 1993] Fourer, R. and Kernighan, B. (1993). AMPL: A modeling language for mathematical programming. Duxburg Press, Massachusetts.

[Fukushima and Pang, 1999] Fukushima, M. and Pang, J. (1999). Convergence of a smoothing continuation method for mathematical programs with complementarity constraints. Lectures Notes in Economics and Mathematical Systems, Springer-Verlag, 447:99–110.

[Hu and Ralph, 2004] Hu, X. and Ralph, D. (2004). Convergence of a penalty method for mathematical programs with complementarity constraints. *Journal of Optimization Theory and Application*, 123(2):365–390.

[Leyffer, 2000] Leyffer, S. (2000). MacMPEC AMPL collection of mathematical programs with equilibrium constraints. `http://wiki.mcs.anl.gov/leyffer/index.php/MacMPEC`

[Lin and Fukushima, 2005] Lin, G. and Fukushima, M. (2005). A modified relaxation scheme for mathematical programs with complementarity constraints. *Ann. Oper. Res.*, 133(1):63–84.

[Kadrani et al., 2009] Kadrani, A., Dussault, J. and Benchakroun, A. (2009). A new regularization scheme for mathematical programs with complementarity constraints. *SIAM Journal on Optimization* , 20(1):78–103.

[Monteiro and Meira, 2011] Monteiro, M., and Meira, J. (2011). A penalty method and a regularization strategy to solve MPCC. *International Journal of Computers Mathematics*, 88(1):145–149.

[Monteiro and Rodrigues, 2011] Monteiro, M., and Rodrigues, H. (2011). Combining the regularization strategy and the SQP to solve MPCC – A MATLAB implementation. *Journal of Computational and Applied Mathematics*, 235(18):5348–5356.

[Outrata et al., 1998] Outrata, J., Kocvara, M. and Zowe, J. (1998). Nonsmooth approach to optimization problems with equilibrium constraints. Kluwer Academic Publishers, Dordrecht.

[Raghunathan and Biegler, 2003] Raghunathan, A. and Biegler, L. (2003). Interior point methods for mathematical programs with complementarity constraints. CAPD technical report, Department of Chemical Engineering, Carnegie Mellon University, June 2003

[Ralph and Wright, 2004] Ralph, D. and Wright, S. (2004). Some proprieties of regularization and penalization schemes for MPECs. *Optimization Methods and Software*, 19(5):527–556.

[Scheel and Scholtes, 2000] Scheel, H. and Scholtes, S. (2000). Mathematical programs with complementarity constraints: Stationarity, optimality and sensitivity. *Mathematics of Operations Research*, 25(1):1–22.

[Scholtes, 2001] Scholtes, S. (2001). Convergence proprieties of a regularization schemes for mathematical programs with complementarity constraints. *SIAM Journal on Optimization*, 11(4):918–936.

Extending the Resource-Task Network (RTN) for industrial scheduling problems

Samuel Moniz*‡, Ana P. Barbosa-Póvoa*, Jorge P. Sousa†‡

* Centro de Estudos de Gestão, Instituto Superior Técnico, Universidade Técnica de Lisboa, 1049-001 Lisboa
E-mail:dli09005@fe.up.pt

† INESC TEC (formerly INESC Porto), Porto,
E-mail:apovoa@ist.utl.pt

‡ Faculdade de Engenharia da Universidade do Porto, Porto.
E-mail:jsousa@inescporto.pt

Abstract

Multipurpose batch plants are general purpose facilities in which production resources can be shared to produce several products [Barbosa-Póvoa, 2007]. These plants are being used by the process industry (*e.g.*, chemical and pharmaceutical) due to their operational flexibility to simultaneously manufacture products with arbitrary production sequences, and to accommodate demands and recipes that change quite often. Consequently, production scheduling is required in order to run efficiently such plants and to ensure responsiveness need to accommodate new customers' orders. The purpose of this paper is to present several extensions of the Resource-Task Network (RTN) discrete-time model that were motivated by a real-world application at Hovione FarmaCiencia SA. We discuss traceability of the production lots, task-unit assignment, sequence dependent changeovers and temporary storage in the processing units.

Keywords: multipurpose batch plants, Resource-Task Network, scheduling, MILP models.

1 Introduction

Optimization models are being adopted by the process industry with the purpose of increasing operational results and efficiency. The application of such models in real world planning and scheduling problems clearly creates improvement opportunities for logistics and manufacturing activities. However relevant implementation challenges arise compromising the models effectiveness. Grossmann [2005] summarized the major challenges for the process industry being as: generic modeling; multiscale optimization; uncertainty; and algorithm and computational efficiency. The first two challenges are related to the capability of the production planning and scheduling models to address complex requirements of the manufacturing processes and to tackle large spatial and temporal scales (*e.g.*, long-term *versus* short-term planning, planning and scheduling and control integration). The third challenge, uncertainty, must be taken into account to approximate the theoretical models to stochastic process variations such as demand variations or equipment breakdown events. The fourth challenge has to do with efficient modeling techniques capable of solving large-scale optimization models.

The development and implementation of optimization models requires a clear understanding of the real scheduling requirements and of the limitations that the models have. In practice, the industry is expecting to have a perfect integration of these models with their decision-making processes. Therefore, models must be efficient in getting reliable solutions and the time required to obtain solutions must take into account the timing of the planning and scheduling decisions. In this work, we present some extensions of the RTN discrete-time model for addressing some scheduling requirements that were found in a real-world application.

This paper is organized as follows. In the next section we present a brief literature review. We then present the problem definition and the base of the RTN discrete-time formulation. In the following sections, we introduce the developed extensions. Then, we describe an illustrative example together with the results obtained by our model. Finally, some conclusions are presented.

2 Literature review

The development of a general model capable of efficiently solving all batch plants scheduling problems has proved to be very difficult, mainly because we need to consider a variety of aspects that depend on the plant structure, process topology and product characteristics [Grossmann, 2005]. Nevertheless, the combination of different problem characteristics motivated the development of several new models and solution approaches. Both State-Task Network (STN) representation suggested by Kondili et al. [1993] and the RTN representation introduced by Pantelides [1994] have been widely used for the design of alternative formulations, see [Barbosa-Póvoa, 2007, Mendez et al., 2006]. In their discrete-time formulations, models deal easily with material balances and inventory costs, multiple delivery dates and result into compact formulations. On the other hand, they present some difficulties when modeling variable processing times and sequence-dependent changeovers. Moreover, we need to be aware of the tradeoffs between accuracy of the scheduling solutions, time discretization and scheduling horizon, since computational performance strongly depends on the number of time intervals considered. Alternatively, continuous-time formulations have been developed in order to overcome the drawbacks of discrete-time formulations. Some relevant work in this field was developed by Castro et al. [2001], Maravelias and Grossmann [2003] and Schilling and Pantelides [1996] in which the authors defined a common time grid for all production resources, and Ierapetritou and Floudas [1998] and Janak et al. [2004] that developed unit-specific time formulations. Continuous-time formulations generally result into large integrality gaps that tend to deteriorate computational times.
Several other authors have used either the STN or the RTN models. For example, Barbosa-Póvoa and Macchietto [1994] developed the maximal state-task network (m-STN) representation that simultaneously considers the operational and design characteristics. Castro et al. [2001] developed a continuous RTN formulation for the scheduling of multipurpose batch plants. Pinto et al. [2005] modified the RTN to address design and retrofit of batch plants with periodic mode operation. Wassick and Ferrio [2011] extended the RTN for addressing multiple deliveries and material transfers and Moniz et al. [2012] introduced the units redesign scheduling problem.
RTN is a generic representation that can deal with different types of problems. In particular, discrete-time formulations have been widely used in models for scheduling batch plants because they can address the majority of the scheduling complexities and are a good basis for the development of additional requirements.

3 Problem definition

In the problem under investigation, we consider that the following information is given: *(i)* the RTN representation of the processes (tasks and resources); *(ii)* the processing units available, and their maximum and minimum capacity; *(iii)* the scheduling granularity and time horizon; *(iv)* the production lots requirements and delivery dates; *(v)* and the value of the products and the raw materials, storage, changeovers and operational costs. The objective is to maximize the overall profit of the schedule by determining: *(i)* the task-unit assignment and respective batch size; *(ii)* the amount of each production lot, in which blending of materials belonging to different lots is defined; *(iii)* the temporary storage in the processing units, and changeovers tasks between production lots of the same product and between different products.

4 The RTN discrete-time formulation

The RTN representation has two types of entities: tasks and resources, see [Pantelides, 1994]. A task refers to an operation that consumes or produces any type of production resource that needs to be scheduled. One main advantage of RTN is that resources are uniformly treated through the set of resources balance constraints (4.1). The formulation developed by Pantelides [1994] uses three types of decision variables. The assignment of tasks to resources is done through binary variables N_{kt} that take the value 1 if task k starts at time t, the continuous variables ξ_{kt} that indicate the task k batch size at the time t, and the continuous variables R_{rt} that state the availability of resource r at time t. The amount of resources consumed or produced (allocated or released if we are referring to processing units) in each time interval is expressed by the integer and continuous part $(\mu_{kr\theta}N_{k,t-\theta}+\nu_{kr\theta}\xi_{k,t-\theta})$ of constraints (4.1). For simplicity, resource consumption and production will be used in a similar way for all types of resources, *i.e.*, materials and processing units. Parameters $\mu_{kr\theta}$ and $\nu_{kr\theta}$ represent the fixed and variable resource

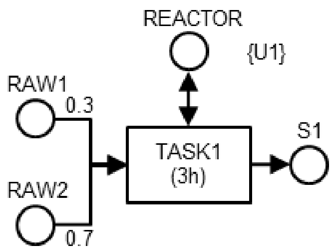


Figure 1: Task representation in RTN model.

consumption/production, respectively. For instance in Figure 1, parameter $\mu_{TASK1,U1,\theta}$ will have the value -1 for $\theta = 0$, 1 for $\theta = 3$, and 0 for $\theta = 1$ and $\theta = 2$. The materials consumption/production will be proportional to the task batch size. In the example, we will have the parameters $\nu_{TASK1,RAW1,\theta}$ equal to -0.3 and $\nu_{TASK1,RAW2,\theta}$ equal to -0.7 for $\theta = 0$, $\nu_{TASK1,S1,\theta}$ equal to 1 for $\theta = 3$ and 0 for $\theta = 1$ and $\theta = 2$. Note that the resource consumption is denoted by a negative sign and the resource production is signed positively.
The RTN discrete-time model was introduced by Pantelides [1994] as follows:

$$R_{rt} = (R_r^{init}|_{t=0}, R_{r,t-1}|_{t>0}) + \sum_{k \in K_r} \sum_{\theta=0}^{\tau_k} (\mu_{kr\theta} N_{k,t-\theta} + \nu_{kr\theta} \xi_{k,t-\theta}) + \Pi_{rt} \quad \forall r \in R, t \in H. \tag{4.1}$$

$$0 \leq R_{rt} \leq R_r^{max} \quad \forall r \in R, t \in H. \tag{4.2}$$

$$V_{kr}^{min} N_{kt} \leq \xi_{kt} \leq V_{kr}^{max} N_{kt} \quad \forall r \in E, k \in K_r, t \in H. \tag{4.3}$$

Constraints (4.1) express the resource balance for all production resources (*e.g.*, processing units, raw materials, intermediaries, final products, people, *etc.*) The resource availability R_{rt} for each time interval is equal to the availability of that resource in the previous time interval $R_{r,t-1}$ plus the amount of resource that is consumed and produced at time t. The minimum and maximum resources allowed are limited by constraints (4.2). For resources of the type *material* the parameter R_r^{max} takes the value 0 if there is Non-Intermediate Storage (NIS) and takes a value greater than zero if there is Finite Intermediate Storage (FIS) or Unlimited storage (UIS). In the latter case the value should be sufficiently large to account for unlimited storage capacity. Constraints (4.3) determine task k batch size which must be within the minimum V_{kr}^{min} and maximum V_{kr}^{max} capacities of resource r for task k. E is the subset of R that groups the processing units, K_r is the set of tasks k that use resource r, τ_k gives the processing time of task k and Π_{rt} specifies the deliveries or receipts of materials.
In the following sections, we discuss the RTN extensions for production lots, task-unit assignment, changeovers and temporary material storage.

4.1 Production lots

In the chemical-pharmaceutical industry, we can define production lot as the fixed quantity of material produced by a set of tasks executed in a known production sequence (referred as recipe). In this way, our schedule must determine the tasks batch sizes in order to achieve the quantity specified by each production lot. Although this requirement can be implemented by constraints (4.1) of the above model, it is not obvious how the relations between production lots, resources and tasks batch sizes are modeled. Moreover, two additional requirements are related with production lots that cannot be modeled in a straightforward manner with RTN, which are blending and traceability of lots. The scheduling model should allow blending of materials belonging to different production lots, while keeping record of the blending process. For example, a task producing a distinct lot may consume an amount of material that was originated from different lots and we need to keep a track record of the individual quantities of each one. To implement these requirements constraints (4.4-4.8) are proposed:

$$R_{rlt}^p = \sum_{k \in K_r} \sum_{\theta=0}^{\tau_k} (\nu_{kr\theta}^p \xi_{kl,t-\theta}) \quad \forall r \in I \cup P, l \in L_r, t \in H. \tag{4.4}$$

$$\sum_{l' \in B_l} R^c_{rl't} = \sum_{l' \in B_l} \sum_{k \in K_r} \sum_{\theta=0}^{\tau_k} (\nu^c_{kr\theta} \xi_{kl',t-\theta}) \quad \forall r \in M, l \in L_r, t \in H. \tag{4.5}$$

$$R_{rlt} = (R^{init}_{rl}|_{t=0}, R_{rl,t-1}|_{t>0}) + R^p_{rlt} + R^c_{rlt} + \Pi_{rlt} \quad \forall r \in M, l \in L_r, t \in H. \tag{4.6}$$

$$R_{rlt} = (R_{rlt}|_{t=0}, R_{rl,t-1}|_{t>0}) + \sum_{k \in K_r} \sum_{\theta=0}^{\tau_k} (\mu_{kr\theta} N_{kl,t-\theta}) \quad \forall r \in E, l \in L_r, t \in H. \tag{4.7}$$

$$\sum_{l \in L_r} R_{rl0} \leq R^{init}_r \quad \forall r \in E. \tag{4.8}$$

Material resources are modeled in constraints (4.4-4.6), separately from the processing units balance constraints (4.7). Constraints (4.4) are only related to the produced intermediaries and final products that are associated to unique production lots. R^p_{rlt} are continuous variables that define the amount of resource r of production lot l produced at time t. Constraints (4.5) model materials consumption given by continuous variables R^c_{rlt} . In this case, blending of production lots is allowed through the set B_l. In practice, blending lots of stable intermediaries or products is allowed to originate other lots of intermediaries or final products. To model lot deliveries we use the continuous variables Π_{rlt} to define the amount of resource r of production lot l delivered at time t. Constraints (4.6) define the material resources balance given by the sum of the resource amount in the previous time interval with the production R^p_{rlt}, consumption R^c_{rlt} and the delivery Π_{rlt} variables. Remember that variables R^c_{rlt} and Π_{rlt} take only negative values because they are related to materials consumption and deliveries respectively. No external material receipts are expected during the scheduling horizon. Finally, Constraints (4.7) define the resources balance just for the processing units and Constraints (4.8) perform the initial assignment of processing units to lots.

4.2 Lot size, demand, task batch size and resource availability

The amount of product r of lot l delivered at delivery d is limited by a minimum Q^{min}_{rld} and a maximum quantity Q^{max}_{rld}. A delivery time window for each delivery d is settled through parameters T^{ed}_{rd} that define the earliest time interval, and T^{dd}_{rd} that define the latest time interval. Moreover, slack variables Π^{slack}_{rld} were added to treat the demand as soft constraints, see (4.9). Constraints (4.10) and constraints (4.11) set the delivery variables to zero for the time intervals out of the delivery time window and for other resources than final products. Bounds on the tasks batch size and resources maximum availability are defined by constraints (4.12) and (4.13), respectively.

$$Q^{max}_{rld} \geq \sum_{T^{dd}_{rd} \geq t \geq T^{ed}_{rd}} (-\Pi_{rlt}) \geq Q^{min}_{rld} - \Pi^{slack}_{rld} \quad \forall r \in P, d \in D_r, l \in L_r. \tag{4.9}$$

$$\sum_{T^{ed}_{rd} > t > T^{dd}_{rd}} \Pi_{rlt} = 0 \quad \forall r \in P, d \in D_r, l \in L_r. \tag{4.10}$$

$$\Pi_{rlt} = 0 \quad \forall r \in R \setminus P, l \in L_r, t \in H. \tag{4.11}$$

$$V^{min}_{krl} N_{klt} \leq \xi_{klt} \leq V^{max}_{krl} N_{klt} \quad \forall r \in E, k \in K_r, l \in L_r, t \in H. \tag{4.12}$$

$$0 \leq \sum_{l \in L_r} R_{rlt} \leq R^{max}_r \quad \forall r \in R, t \in H. \tag{4.13}$$

4.3 Task unit assignment

Production processes are defined by recipes that specify the tasks sequence and the alternative processing units that each task may have. In many situations it is desirable to assign just one unit to each task. This happens frequently in the chemical-pharmaceutical industry, where a significant number of alternative units need to be assessed during the scheduling process. This requirement can be implemented through constraints (4.14). Thus, if the first term of the constraint is one, the second term must be zero for all alternative tasks k' of task k for all time intervals t'.

$$\sum_{l\in L} N_{klt} + \sum_{l\in L}\sum_{k'\in A_k} N_{k'lt'} \leq 1 \quad \forall k\in K, t\in H, t'\in t..T. \tag{4.14}$$

4.4 Changeovers

Changeovers cannot be neglected since they often occupy processing units during long time periods. We may have unit and sequence dependent changeovers, the latter being usually more significant in terms of time. Sequence dependent changeovers can be modeled in the original RTN formulation through the creation of changeover tasks or through a set of constraints that inhibit the occurrence of the production tasks for certain sequences.

4.4.1 Option 1 - changeover constraints

If it is not relevant to determine the exact time of the changeover occurrences, we can use constraints (4.15). Therefore, if task k of lot l occurs at time t, the first term of the constraint is equal to one and the second term is forced to be zero for all tasks k' belonging to lots l' and time intervals corresponding to $t-\tau_{k'}-\theta$.

$$\sum_{k\in f_l^r} N_{klt} + \sum_{k'\in f_{l'}^r} N_{k'l',t-\tau_{k'}-\theta} \leq 1 \quad \forall r\in E, l,l'\in L^{chg}, \theta=0..C_{l'l}-1, t\in H. \tag{4.15}$$

4.4.2 Option 2 - changeover tasks

When changeovers timing and costs need to be taken into account in the model, we consider changeovers tasks k^{chg}, see constraints (4.16). One possible way to specify cleaning timings is to use a changeover parameter $\mu^{chg}_{k^{chg}rl\theta}$ that defines the changeover time between products and lots. In our case, changeovers between lots of the same product take 8 hours and changeovers between different products take 24 hours.

$$\begin{aligned} R_{rlt} = (R_{rlt}|_{t=0}, R_{rl,t-1}|_{t>0}) + \sum_{k\in K_r}\sum_{\theta=0}^{\tau_k}(\mu_{kr\theta}N_{kl,t-\theta}) + \\ \sum_{k^{chg}\in K_{rl}}\sum_{\theta=0}^{\tau_k^{chg}}(\mu^{chg}_{k^{chg}rl\theta}C_{k^{chg},t-\theta}) \quad \forall r\in E, l\in L_r, t\in H. \end{aligned} \tag{4.16}$$

4.5 Temporary material storage

The RTN formulation deals easily with different inventory policies, but at high computational costs if certain storage policies are considered. The UIS policy can be modeled by setting the maximum availability of resource parameter R_r^{max} to a sufficient large value for the materials for which this policy applies. The FIS policy may require storage tasks, if the storage units need to be planned as well. In the NIS policy there are no dedicated units available for storage, but processing units may temporarily hold materials after the task is finished, which also requires the creation of storage tasks. Two modeling alternatives are presented to implement this requirement.

4.5.1 Option 1 - Intermediate availability

One possible way to model temporary storage is to define a binary variable to control the existence of the intermediate when a NIS policy is considered. Therefore, Z_{rt} is equal to one if intermediate r at time interval t is greater than zero, see constraints (4.17).
Constraints (4.18-4.19) implement this behavior, while constraints (4.20) define that if Z_{rt} is equal to one, then production or changeover tasks must be equal to zero, since the processing unit is still occupied with the intermediate. Note that, in this formulation R_r^{max} must take a value equal to the maximum capacity among the processing units suitable to produce intermediate r. Constraints (4.20) are valid just for the case of single task-unit assignment, which is imposed by constraints (4.14).

$$Z_{rt} = \begin{cases} 1 & if\ \sum_{l\in L_r} R_{rlt}>0 \\ 0 & if\ \sum_{l\in L_r} R_{rlt}=0 \end{cases} \quad \forall r\in I^{NIS}, t\in H. \tag{4.17}$$

$$\sum_{l \in L_r} R_{rlt} \geq Z_{rt} \quad \forall r \in I^{NIS}, t \in H. \tag{4.18}$$

$$\sum_{l \in L_r} R_{rlt} \leq R_r^{max} Z_{rt} \quad \forall r \in I^{NIS}, t \in H. \tag{4.19}$$

$$\sum_{k \in K_r^p} N_{klt} + \sum_{k^{chg} \in K_{rl}^{chg}} C_{k^{chg}t} + Z_{rt} \leq 1 \quad \forall r \in I^{NIS}, l \in L_r, t \in H. \tag{4.20}$$

4.5.2 Option 2 - Intermediate tasks

Another possible approach is to use storage tasks for all tasks producing intermediaries for which the NIS policy is applied. Kondili et al. [1993] suggested this approach for the STN, which can be applied directly in the RTN. So, constraints (4.21) define that the batch size of a storage task is less or equal than the previous amount stored plus the amount produced at each time interval. If the batch size of a storage task is greater than zero, then the assignment binary variable for the storage task must be one by constraints (4.12). Constraints (4.21) ensure that the intermediate is held by the unit in which it was produced. Note that, storage tasks have duration equal to one since materials availability needs to be checked at every time interval.

$$\xi_{k_r^{sto}lt} \leq \xi_{k_r^{sto}l,t-1} + \sum_{\theta=0}^{\tau_k} (\nu_{kr\theta}^p \xi_{kl,t-\theta}) \quad \forall r \in I^{NIS}, k \in K_r^p, l \in L_r, t \in H. \tag{4.21}$$

4.6 Objective function

The objective function targets the profit maximization and it is modeled by expression (4.22). The first term determines the product value, and the following terms are related to storage, changeover and operational costs and a penalty factor associated with the slack variable Π_{rld}^{slack}. It is assumed that all products are delivered; therefore the resource availability at the end of the scheduling horizon R_{rT} is equal to zero.

$$\begin{aligned} max\Bigg[\sum_{r \in P} \sum_{l \in L_r} \sum_{t \in H} \big((v_r - c_r^{raw})(-\Pi_{rlt})\big) - \sum_{r \in I \cup P} \sum_{l \in L_r} \sum_{t \in H} c_r^{sto} R_{rlt} - \sum_{k^{chg} \in K^{chg}} \sum_{t \in H} c_{k^{chg}}^{chg} C_{k^{chg}t} - \\ \sum_{k \in K_r} \sum_{l \in L_r} \sum_{t \in H} c_k^{op} N_{klt} - \sum_{r \in P} \sum_{l \in L_r} \sum_{d \in D_r} c^{slack} \Pi_{rld}^{slack} \Bigg] \end{aligned} \tag{4.22}$$

5 Illustrative example

Consider the determination of a production schedule with three products PA, PB and PC. Task sequences and respective alternative units are depicted in Figure 2. Products PA and PB are produced from raw materials, while product PC is produced from PA and PB. The objective is to maximize the overall profit by determining one week schedule that keeps track record of the production lots, changeovers between products and lots, and considers temporary storage in the processing units. Table 1 shows the demand for the three products, the earliest and due date, the minimum and maximum lot size and the number of production lots. The model was defined by resources balance constraints (4.4-4.6); processing units initial lot-assignment constraints (4.8); lot size, demand, task batch size and resource availability constraints (4.9-4.13); task-unit assignment constraints (4.14); processing units balance and changeovers tasks constraints (4.16); and temporary storage constraints (4.17-4.20). The time intervals were assumed to have duration of 8 hours. The model was written using ILOG/CPLEX version 12.5 and was solved in an Intel Core i7 at 2.67GHz with 4 GB of RAM.

5.1 Results

The model solution and statistics are given in Table 2. The model has a profit of 205.42 rmu (relative monetary units) with the following value structure: income equal to 215.4 rmu; storage costs are 1.03 rmu; operational costs are 3.95 rmu; and changeover costs are 5 rmu. This solution consists in producing

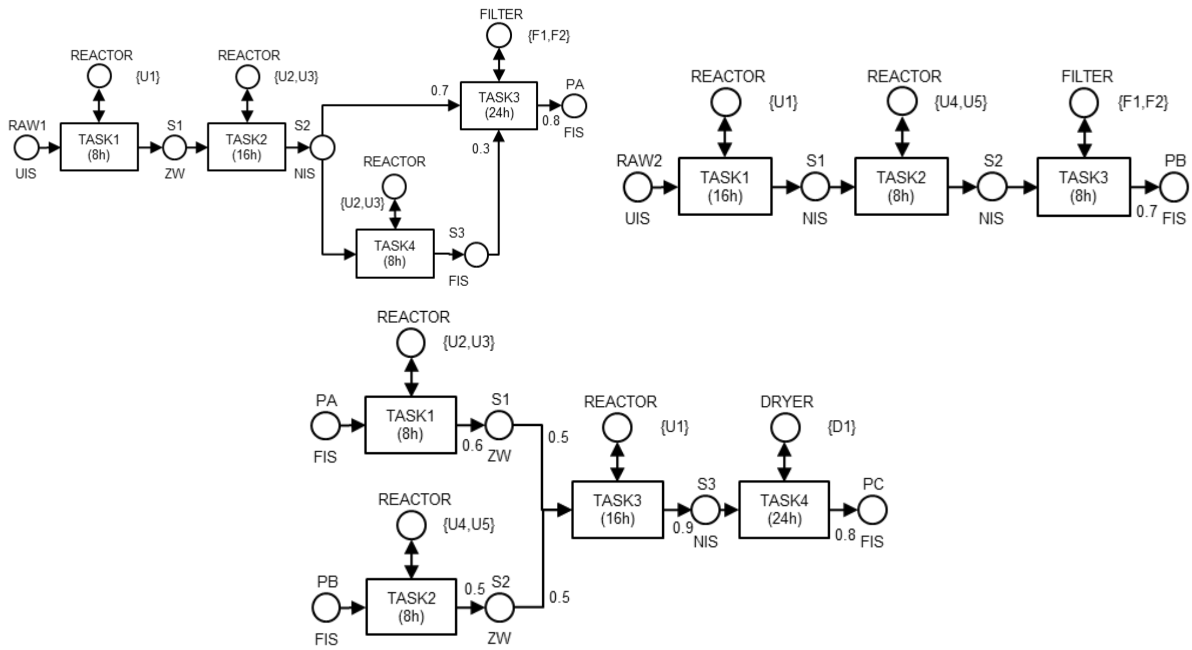


Figure 2: Recipes of products PA, PB and PC.

Table 1: Products PA, PB and PC demand (one week schedule).

Product	Earliest date	Due date	Min. lot size [tons]	Max. lot size [tons]	Lots [#]
PA	16	21	1	2	2
PB	16	21	1	2	2
PC	18	21	1	2	1

3.15 tons of PA, 4 tons of PB and 2 tons of PC. The required computational time was 40 CPU sec with 0% of integrality gap. Looking in more detail to the solution, we would expect to have two lots of PA with an amount between 1 and 2 tons, to be available between time intervals 16 and 21 (see Table 1). The model retrieved 2 tons and 1.15 tons for the first and second lot of PA, respectively. The first lot is entirely consumed by PC, while the second lot is just partially consumed. PC requires 0.32 tons of the second lot and, therefore, the remaining 0.83 tons are available for delivery. Regarding PB, both production lots have 2 tons. PC consumes a total of 2.78 tons of PB and the remaining 1.22 tons are available for delivery.

Table 2: Model solution and statistics.

Instance	Int. variable/ cont. variables/ constraints	Nodes	LP relaxation	Gap %	Objective	CPU time (s)
One week	2882/3084/9845	22931	219.79	0	205.42	40
Two weeks	5633/6027/27338	54808	476.18	5	453.28	1813

A two weeks instance is also shown in Table 2, and it can be seen that the solution time increases significantly indicating the difficulty of solving this problem when a large instance is considered. Assuming an integrality gap within 5% the model obtained a solution of 453.28 rmu in 1813 CPU sec.
The production schedule for one week (see Figure 3) shows production and changeover tasks and temporary storage occurrences of the three products. The first lot of PA initiates the production in unit U1 at $t = 1$, and then a changeover occurs at $t = 3$ to initiate the second lot at $t = 4$. At $t = 5$, unit U1 is prepared to produce PB and at $t = 8$ the first lot of this product starts being produced. Among the task-units assignments the model retrieved the following production lines: $PA = \{U1, U2, U3, F1\}$, $PB = \{U1, U5, F2\}$ and $PC = \{U1, U3, U4, D1\}$. Moreover, it can be seen that U3 temporarily held intermediate S2 of PA from $t = 7$ to $t = 11$, and that two more occurrences of temporary storage took

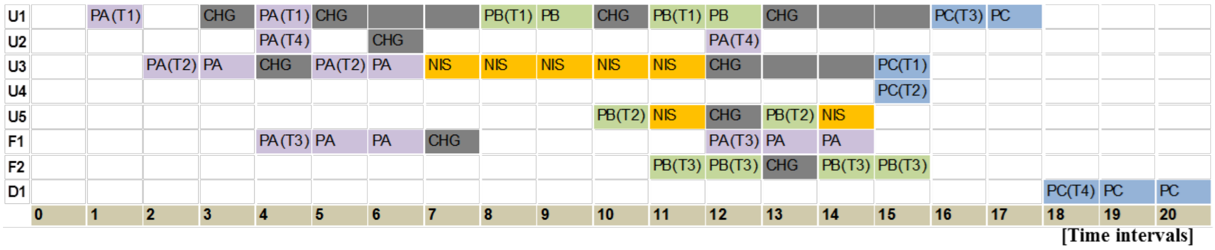


Figure 3: Production schedule for one week horizon.

place during the production of PB at $t = 11$ and $t = 14$.

6 Conclusions

This paper has discussed some scheduling requirements often found in the chemical-pharmaceutical industry and proposed extensions to the RTN discrete-time model. In this way, we address the traceability of production lots, temporary storage within the processing units, sequence dependent changeovers and task-unit assignment. An illustrative example considering three products is solved to show the applicability of these extensions for scheduling horizons of one and two weeks. The one week problem was solved in 40 CPU sec, while for the two weeks problem, a solution with a 5% gap was reached in 1813 CPU sec. The extensions presented in this paper lead to a more general formulation, since the model is capable to address additional scheduling features. However, this is done through the increase of the number of constraints and binary variables, thus complicating the model and often leading to instances that are more difficult to solve. So the development of model extensions for industrial scheduling problems must be followed by the development of non-exact methods as a way to solve real-size practical instances.

Acknowledgment

The authors gratefully acknowledge the financial support of Hovione FarmaCiencia SA and Fundação para a Ciência e Tecnologia, under the grant SFRH/BD/33970/2009.

References

[Barbosa-Póvoa, 2007] Barbosa-Póvoa, A. P. (2007) A critical review on the design and retrofit of batch plants. *Computers and Chemical Engineering*, 31: 833-855.

[Barbosa-Póvoa and Macchietto, 1994] Barbosa-Póvoa, A. P., and Macchietto, S. (1994). Detailed design of multipurpose batch plants. *Computers and Chemical Engineering*, 18: 1013-1042.

[Castro et al., 2001] Castro, P., Barbosa-Póvoa, A., and Matos, H. (2001). An improved RTN continuous-time formulation for the short-term scheduling of multipurpose batch plants. *Industrial and engineering chemistry research*, 40: 2059-2068.

[Grossmann, 2005] Grossmann, I. (2005). Enterprise wide optimization: A new frontier in process systems engineering. *AIChE Journal*, 51: 1846-1857.

[Ierapetritou and Floudas, 1998] Ierapetritou, M., and Floudas, C. (1998). Effective continuous-time formulation for short-term scheduling. 1. Multipurpose batch processes. *Industrial and engineering chemistry research*, 37: 4341-4359.

[Janak et al, 2004] Janak, S. L., Lin, X., and Floudas, C. A. (2004). Enhanced continuous-time unit-specific event-based formulation for short-term scheduling of multipurpose batch processes: Resource constraints and mixed storage policies. *Industrial and engineering chemistry research*, 43: 2516-2533.

[Kondili et al., 1993] Kondili, E., Pantelides, C., and Sargent, R. (1993). A general algorithm for short-term scheduling of batch operations-I. MILP formulation. *Computers and Chemical Engineering*, 17: 211-227.

[Maravelias and Grossmann, 2003] Maravelias, C. T., and Grossmann, I. E. (2003). New general continuous-time state-task network formulation for short-term scheduling of multipurpose batch plants. *Industrial and engineering chemistry research*, 42: 3056-3074.

[Mendez et al., 2006] Mendez, C. A., Cerda, J., Grossmann, I. E., Harjunkoski, I., and Fahl, M. (2006). State-of-the-art review of optimization methods for short-term scheduling of batch processes. *Computers and Chemical Engineering*, 30: 913-946.

[Moniz et al., 2012] Moniz, S., Barbosa-Póvoa, A. P.,and Pinho de Sousa, J. (2012). Scheduling with equipment redesign in multipurpose batch plants. *Foundations of Computer-Aided Process Operations - FOCAPO 2012*.

[Pantelides, 1994] Pantelides, C. C. (1994). Unified frameworks for optimal process planning and scheduling. *Cache Publications New York*: 253-274:

[Pinto et al., 2005] Pinto, T., Barbosa-Póvoa, A. P., and Novais, A. Q. (2005). Optimal design and retrofit of batch plants with a periodic mode of operation. *Computers and Chemical Engineering*, 29: 1293-1303.

[Schilling and Pantelides, 1996] Schilling, G., and Pantelides, C. C. (1996). A simple continuous-time process scheduling formulation and a novel solution algorithm. *Computers and Chemical Engineering*, 20: S1221-S1226.

[Wassick and Ferrio, 2011] Wassick, J. M., and Ferrio, J. (2011). Extending the Resource Task Network for Industrial Applications. *Computers and Chemical Engineering.*

Investment Projects: Evaluation Tools and Methods

Nuno Moutinho*, Helena Mouta*

* *Escola Superior de Tecnologia e Gestão, Instituto Politécnico de Bragança, Portugal*
E-mail: `nmoutinho@ipb.pt`, `hmouta@ipb.pt`

Abstract

Given the need to study different areas of analysis during the investment decision process, it is imperative to know which tools and methods are used by companies to assess various financial and non-financial aspects. As regards tools, we find that firms use checklists of analysis for non-financial aspects, whereas they use their past experience in risk assessment, gathered from other projects. Records of past evaluation tend to be maintained in companies and those that use external advisors to evaluate projects tend to perform political analysis. As for the methods, companies use the identification of risk factors and assessment of effects and risk probabilities, the discussion and assessment of favorable and unfavorable factors to the project's execution, a coordinated analysis of financial and non-financial aspects, and the creation of lists of risk indices, attributing a qualitative weight to each item. We have also analyzed the relationship between these tools and methods and each area of analysis in project evaluation.

Keywords: Real Investment Projects; Evaluation Tools and Methods; Non-Financial Analysis.

1 Introduction

Financial theory has analyzed investment projects based on cash flows and has used NPV, IRR and Payback. In this case, investment decision making is relatively simple since only financial criteria are taken into account (Lopes and Flavell, 1998). Moutinho and Lopes (2011) show the importance of the analysis of various non-financial aspects and how some of those aspects have a greater relevance than that attributed to financial elements. As the most relevant elements, they find the strategic and technical aspects, followed by the analysis of financial and commercial aspects. The least relevant in firms' project appraisal are the social and political areas. In these analyses, there are many aspects that are not easily measured.

As non financial analysis can provide additional information to the decision making, it is important to take into account all financial and non financial areas: strategic, technical, commercial, political, social, environmental, organizational, human resource and management (Skitmore, et. al., 1989; Adler, 2000; Chen, 1995; Meredith and Mantel, 2000; Love et. al., 2002; Kendra and Taplin, 2004; Moutinho and Lopes, 2011). The incorporation of these areas of analysis can pose new problems to project appraisal. However, Moutinho (2011) presents the main tools and methods used in the analysis of all financial and non financial areas together in a project evaluation. As main tools for the evaluation of non financial aspects, he finds that companies choose the incorporation of people from several backgrounds, the free-format qualitative evaluation, the experience of evaluators/decision-makers in other projects, and the use of external advisors. About the methods used he finds that companies prefer the discussion and assessment of favourable and unfavourable factors to the project's execution, followed by the coordinated analysis of financial and non-financial aspects, the identification of risk factors, and the assessment of effects and risk probabilities. So, we propose to analyze which tools and methods companies can use to incorporate each area of analysis into their decision-making process.

This paper is organized as follows. In the next section, we present our hypotheses and the research methodology used. Then we present and discuss the results. Finally, we present our conclusions.

2 Research Hypotheses and Methodology

Lopes and Flavell (1998) show the most important tools for the incorporation of non-financial aspects in project evaluation: decision-makers' experience in other projects, maintaining records of past evaluations to verify the credibility of management opinion; creating a record of its own past experience and generating checklists of analysis of all aspects; incorporating people from several backgrounds; using external advisors in project appraisal; use of a free-format qualitative evaluation; definition of alternative strategies, and; considering all risks. In addition, Nardini (1997) suggests that for non financial aspects companies can use a multi-criteria analysis of the projects. In project analysis several tools of financial evaluation are known. However, it should be addressed how the assessment is made when the projects take into account non-financial aspects. As such the following hypothesis is studied:

Hypothesis 1: The tool used to evaluate non-financial elements is associated with each dimension of evaluation.

In respect to the methods used to quantify non-financial aspects in project evaluation, Lopes and Flavell (1998) consider that these methods are dependent on personal opinions and perspectives and there should be a combination between personal analysis and personal opinion. It is necessary to employ common sense to assess and identify risk factors, assess their effects and probabilities. The decisions are justified through discussion of arguments of all aspects analysed in inter-connection. Companies can create lists of risk indices, attributing a qualitative weight to each item. The success of these processes depends on the personal experience, the feedback of past projects, the systematic and detailed analysis of expert support, and the inclusion of people from several backgrounds and its connection to decision making. Companies can also create checklists as risk warnings to decision-makers. Then, the methods used to quantify non-financial aspects can be tested by the hypothesis below:

Hypothesis 2: The method used to incorporate and quantify non-financial aspects is related to each dimension of evaluation.

3 Tools and Methods

This research has used the same questionnaire of Moutinho and Lopes (2011) and Moutinho (2011). There were 96 responses to the questionnaire addressed to the 1000 largest Portuguese companies (in terms of sales) in 2005, representing a response rate of 9,6%.

Regarding the first hypothesis mentioned in Section 2, Table 1 identifies the methods for evaluating non-financial aspects, used in each dimension of analysis. Companies tend to adopt checklists of analysis of non-financial aspects when considering the social and the project manager aspects in the project evaluation. The appraisal based on past experience in risk assessment, gathered by several companies tends to be used by companies engaged in environmental analysis and by those who do not consider the human resources area. Maintaining records of past evaluation tends to be used in companies engaged in political analysis. The experience of evaluators/decision-makers in other projects tends to be used by companies engaged in technical analysis and by those who consider the social analysis in project evaluation. Companies that use external advisors to evaluate projects tend to perform political analysis. Considering management's errors of judgment in risk assessment tends not to be used by companies that analyze strategic aspects.

Table 1 shows the results of multivariate analysis, by logit, between the probability of performing each form of assessment of financial and non-financial analyses and the performance of each dimensions of evaluation. Panel A analyses the free-format qualitative evaluation. Panel B analyses the checklists of analysis of non-financial aspects. Panel C analyses past experience in risk assessment, gathered from other projects. Panel D analyses the maintaining records of past evaluations. Panel E analyses the experience of evaluators/decision-makers in other projects. Panel F analyses the use of external factors. Panel G analyses the incorporation of people from several backgrounds. Panel H analyses the consideration of management's errors of judgment in non-financial evaluation. Panel I analyses the consideration of management's errors of judgment in risk assessment. Each column shows the regression for each form of assessment. In addition to the coefficient of the variable, in brackets is the standard deviation associated with it. ***, ** and * show the existence of statistical significance at 1

Next, with respect to hypothesis 2 defined in Section 2, we study the methods used to incorporate and quantify non-financial aspects. These results are presented in Table 2. Companies that use the identification of risk factors and assess effects and risk probabilities are carrying out financial, social and project manager analyses. The discussion and assessment of favorable and unfavorable factors to the project's execution is used in companies engaged in human resources analysis and those that do not perform organizational analysis. A coordinated analysis of financial and non-financial aspects tends to be chosen by companies engaged in the analysis of organizational aspects. Requesting opinions from each area is common in companies engaged in human resources analysis. The creation of lists of risk indices, attributing a qualitative weight to each item tends not to be used in companies engaged in organizational analysis and those that do not consider the strategic dimension in the evaluation of projects. Dividing

Table 1: Evaluation tools by area of analysis

Analyses	***Panel A***	***Panel B***	***Panel C***	***Panel D***	***Panel E***	***Panel F***	***Panel G***	***Panel H***	***Panel I***
C	-0.99	-5.92	-0.86	-2.12	-5.43	0.27	-0.52	-0.82	-0.75
	(-0.72)	(-3.32)	(-0.55)	(-1.20)	(-2.84)	(0.19)	(-0.38)	(-0.54)	(-0.54)
Financial	-0.58		0.16	-0.29	1.29	-0.54	0.87	-0.04	-0.32
	(-0.84)		(0.18)	(-0.29)	(1.02)	(-0.69)	(1.10)	(-0.03)	(-0.35)
Strategic	1.15		-1.36	-0.97		-1.61	-0.87	-2.51 *	-2.08
	(0.91)		(-0.99)	(-0.73)		(-1.40)	(-0.74)	(-1.75)	(-1.55)
Technical	0.22	1.86	0.86	0.77	1.87 *	0.23	-0.69	0.01	-0.05
	(0.35)	(1.34)	(1.02)	(0.66)	(1.95)	(0.28)	(-1.01)	(0.01)	(-0.04)
Commercial	0.25	1.26	0.34	-0.21	0.73	-0.56	0.78		
	(0.42)	(1.44)	(0.43)	(-0.25)	(0.77)	(-0.82)	(1.18)		
Political	-0.64	-0.56	-0.47	1.52 **	0.06	1.43 **	-0.20	-0.67	-0.49
	(-1.07)	(-0.79)	(-0.68)	(1.99)	(0.09)	(2.21)	(-0.32)	(-0.82)	(-0.57)
Social	-0.11	1.52 **	-0.18	-0.09	1.04 *	0.50	0.72	0.09	-0.03
	(-0.19)	(2.06)	(-0.25)	(-0.14)	(1.65)	(0.81)	(1.21)	(0.09)	(-0.03)
Environmental	0.04	0.42	1.16 ***	0.75	0.56	0.18	-0.43	1.16	0.96
	(0.07)	(0.71)	(1.78)	(0.95)	(0.88)	(0.30)	(-0.74)	(1.08)	(0.83)
Organizational	0.02	1.08	0.28	-0.40	0.66	0.04	0.69	1.08	
	(0.04)	(1.38)	(0.37)	(-0.49)	(0.87)	(0.05)	(1.07)	(1.07)	
Hyman Resourses	0.25	0.26	-1.72	0.40	0.48	-0.04	0.51	-1.63	-0.92
	(0.36)	(0.30)	(-2.01) **	(0.42)	(0.53)	(-0.05)	(0.68)	(-1.52)	(-0.85)
Project Manager	-0.58	1.11 *	0.55	0.07	-0.45	0.55	-0.82	-0.52	-0.30
	(-1.2)	(1.94)	(0.96)	(0.11)	(-0.65)	(0.99)	(-1.56)	(-0.55)	(-0.32)
McFadden R2	0.04	0.28	0.12	0.13	0.22	0.14	0.10	0.13	0.12
LR statistic	5.21	33.00	12.75	13.14	28.29	17.86	13.04	4.37	3.87
Prob. (LR stat)	0.88	0.00	0.24	0.22	0.00	0.06	0.22	0.89	0.87

factors into levels of importance for subsequent evaluation is little used by companies that conduct environmental analysis in project evaluation.

In Table 2, Panel A analyses the identification of risk factors, assessment of effects and risk probabilities. Panel B analyses the discussion and assessment of favorable and unfavorable factors to the project's execution. Panel C analyses the coordinated analysis of financial and non-financial aspects. Panel D analyses the request of opinions from each area and verification of consistency with the company's strategy. Panel E analyses the creation of lists of risk indices, attributing a qualitative weight to each item. Panel F analyses the attribution of positive and negative considerations to each factor in the qualitative analysis. Panel G analyses the separation of factors into levels of importance for subsequent evaluation. Each column shows the regression for each method. In addition to the coefficient of the variable, in brackets is the standard deviation associated with it. ***, ** and * show the existence of statistical significance at 1

Table 2: Evaluation methods by area of analysis

Analyses	***Panel***	***Panel***	***Panel***	***Panel***	***Panel***	***Panel***	***Panel***
	A	***B***	***C***	***D***	***E***	***F***	***G***
C	-2.058	-2.049	-3.000	-0.342	-2.225	-1.895	-2.185
	(-1.202)	(-1.293)	-1.981	(-0.248)	(-1.033)	(-1.094)	(-3.447)
Financial	1.903 **	1.045	1.253	0.153			
	(2.066)	(1.175)	(1.474)	(0.198)			
Strategic	-0.253	2.244	0.415	-0.343	-3.360 ***	-0.769	
	(-0.182)	(1.576)	(0.327)	(-0.286)	(-3.794)	(-0.553)	
Technical	-0.927	0.697	0.655	-1.102	-1.413	0.154	
	(-1.260)	(0.922)	(0.912)	(-1.613)	(-1.367)	(0.161)	
Commercial	0.321	-0.652	-0.443	0.134			
	(0.453)	(-0.775)	(-0.627)	(0.201)			
Political	-0.039	-0.138	0.556	-0.424		-0.496	-0.928
	(-0.055)	(-0.171)	(0.864)	(-0.643)		(-0.534)	(-0.964)
Social	1.873 ***	0.869	-0.006	0.137		-0.105	1.398
	(2.675)	(1.031)	(-0.009)	(0.226)		(-0.127)	(1.376)
Environmental	-0.154	-0.311	0.051	-0.067	1.168	0.012	-2.267 ***
	(-0.235)	(-0.463)	(0.088)	(-0.116)	(0.990)	(0.012)	(-2.665)
Organizational	-0.838	-2.582 **	1.623 **	-0.182	2.276 ***	0.763	
	(-1.054)	(-1.984)	(2.424)	(-0.282)	(2.692)	(0.829)	
Hyman Resourses	0.268	2.827 **	-0.081	1.492 *		0.692	
	(0.317)	(2.118)	(-0.111)	(1.901)		(0.569)	
Project Manager	1.637 ***	0.011	0.659	-0.136	1.471	-0.250	
	(2.797)	(0.018)	(1.211)	(-0.259)	(1.681)	(-0.428)	
McFadden R2	0.262796	0.209014	0.202143	0.075609	0.199910	0.049491	0.145387
LR statistic	34.23416	20.87265	26.12615	9.212466	8.921399	3.739057	5.679960
Prob. (LR stat)	0.000169	0.021999	0.003574	0.512066	0.112241	0.879866	0.128263

4 Conclusion

In the process of evaluating investment projects, because of the importance of combining the analysis of financial and non financial aspects, it becomes important to determine how to evaluate each aspect. In this article, we show which tools and methods are used by companies in the analysis of each individual assessment areas.

Companies use checklists of analysis of non-financial aspects, and use their past experience in risk assessment, gathered from other projects. Maintaining records of past evaluation tends to be used in companies and those that use external advisors to evaluate projects tend to perform political analysis. About the methods used, companies use the identification of risk factors and assessment of effects and risk probabilities, the discussion and assessment of favorable and unfavorable factors to the project's execution, a coordinated analysis of financial and non-financial aspects, and the creation of lists of risk indices, attributing a qualitative weight to each item.

The scope and the methodology adopted for data collection were the main problems encountered during the work. In fact, the lack of empirical work on the evaluation of non-financial aspects implies the construction of a survey, which may bring some complications, including interpretation of the universal questions. Another limitation of this study concerns the representativeness of the sample, since the findings only represent the reality of the practices of Portuguese companies if the answers for each project are representative of the universe of projects. In addition, the studies are based on questionnaires measuring beliefs or opinions and not necessarily actions or occurrences, since there is no way to verify if these match the company's actions.

References

[Adler, 2000] Adler, A. (2000). Strategic Investment Decision Appraisal Techniques: The Old and the New. *Business Horizons*, 43(6):15–22.

[Chen, 1995] Chen, S. (1995). An Empirical Examination of Capital Budgeting Techniques: Impacts of Investment Types and Firm Characteristics. *The Engineering Economist*, 40(2):145–167.

[Kendra and Talpin, 2004] Kendra, K. and Taplin, L. J. (2004). Project Success: A Cultural Framework. In *Project Management Journal*, 35(1):30–45.

[Lopes and Flavell, 1998] Lopes, M. and Flavell, R. (1998). Project Appraisal - A Framework to Assess Non-Financial Aspects of Projects During the Project Life Cycle. In *International Journal of Project Management*, 16(4):223–233.

[Love et al., 2002] Love, P., Holt, G., Shen, L., Li, H. and Irani, Z. (2002). Using Systems Dynamics to Better Understand Change and Rework in Construction Project Management Systems. In *International Journal of Project Management*, 20(6):425–436.

[Moutinho, 2011] Moutinho, N. (2011). Decision-Making Process in Investment Projects. In *International Conference on Project Evaluation - ICOPEV*, 28-29 April, Guimarães.

[Moutinho and Lopes, 2011] Moutinho, N. and Lopes, M. (2011). The Relative Importance of Financial and Non-Financial Analysis in Project Evaluation - Evidence from Portuguese Firms. In *2011 FMA European Conference*, 8-10 June, Porto.

[Nardini, 1997] Nardini, A. (1997). Proposal for Integrating Environmental Impact Assessment, Cost-Benefit Analysis and Multicriteria Analysis in Decision-Making. In *Project Appraisal*, 12(3):173–184.

[Skitmore et al., 1989] Skitmore, R., Strading, S. and Tuohy, A. (1989). Project Management Under Uncertainty. In *Construction Management and Economics*, 7(2):103–113.

Planeamento de rotas marítimas e estiva de contentores

Jorge Oliveira*, Ana Moura*†

* *Departamento de Economia, Gestão e Engenharia Industrial da Universidade de Aveiro*
† *INESC COIMBRA - Instituto de Engenharia de Sistemas e Computadores de Coimbra,*
e-mail:`jorge. antonio@ua. pt`
e-mail:`ana. moura@ua. pt`

Resumo

Este trabalho pretende contribuir para uma melhoria da gestão de frotas de porta-contentores no transporte marítimo de pequenas distâncias (portos europeus), por forma a reduzir os custos de distribuição. Comparativamente com outros meios de transporte, a distribuição por via marítima tem várias vantagens reconhecidas pelos países da UE. Esta distribuição pode ser abordada como uma combinação de dois problemas NP-Difíceis: o CSP - *Container Stowage Problem* e o VRP - *Vehicle Routing Problem.* A integração destes dois problemas (CSSRP - *Container Stowage and Ship Routing Problem*) é também um problema de otimização combinatória NP-Difícil. O principal objetivo é resolver o problema de carga de contentores e o planeamento das rotas dos porta-contentores de uma forma integrada. Neste trabalho é apresentado um trbalho de progrmação inteira e testado com instâncias geradas a partir de dados reais. Para os resultados obtidos encontrou-se sempre a solução ótima num tempo computacional bastante reduzido.

Palavras chave: Problema de rotas para veículos, Empacotamentos bidimensionais, Programação linear inteira.

1 Introdução

Existem vários meios de transporte para movimentação de contentores entre diferentes localizações. O transporte marítimo é um deles e está entre as formas de transporte que mais tem crescido ao longo da última década. Até aos dias de hoje, o transporte marítimo de curta distância é responsável por grande parte do fluxo de frotas na costa europeia. O crescimento da distribuição marítima nas últimas décadas, deve-se ao uso crescente de contentores. De acordo com *Drewry Shipping Consultants*, mais de 70% do transporte internacional marítimo é efetuado em contentores. As frotas de transporte de contentores expandiram-se a uma taxa anual de 9% ao longo do período 2002-2006 e o total da frota foi de aproximadamente 23.2 milhões de TEUs (*Twenty-foot Equivalent Units*) em 2006. Segundo estatísticas da UE 3800 milhões de toneladas foram transportadas por via marítima em 2006 e espera-se que até 2018, depois de ultrapassada a crise económica actual, sejam alcançadas as 5300 milhões de toneladas. Por outro lado, neste tipo de distribuição as desvantagens prendem-se com: burocracia nos clientes; eficiência e custo dos serviços nos portos; duração das viagens; inflexibilidade das rotas; e dependência de fatores ambientais. Com este estudo pretende-se contribuir para a redução e otimização de algumas destas desvantagens. O objetivo principal é a minimização dos custos totais relacionados com a duração das viagens, duração dos tempos de serviço e eficiência nos portos. Desta forma foi desenvolvido um modelo de programação linear inteira (PLI) que tem como função objetivo minimizar o custo de distribuição marítima de curta distância. A ideia surge da comparação deste tipo de transporte com o transporte terrestre. No transporte terrestre rodoviário o problema apresentado por [Moura e Oliveira, 2009], Problema de Rotas e Carga de Veículos (VRLP - *Vehicle Routing and Loading Problem*), é a integração entre dois problemas bastante reconhecidos na literatura VRP (*Vehicle Routing Problem*) e CLP (*Container Loading Problem*). Neste problema, as rotas e a carga das caixas (ou paletes) em camiões, são determinadas considerando não só a regra LIFO (*Last-In-First-Out*), assim como todas as restrições inerentes a estes dois problemas. O estudo do VRLP iniciou-se em 2006 e desde então têm sido publicados alguns trabalhos. [Gendreau et al., 2006] and [Moura e Oliveira, 2009], foram os primeiros autores a formular este problema, a apresentar abordagens para a sua resolução e publicar problemas de teste. Seguindo a mesma linha de raciocínio, neste trabalho abordamos o problema de distribuição marítima para curtas distâncias e consideramo-lo como um caso particular do VRLP. As rotas são planeadas considerando: distâncias, procura, e deadlines

de entrega para cada porto. Além disso, considera-se a interligação entre a otimização das rotas e a colocação dos contentores nos porta-contentores, de forma a evitar a existência de overstows[1]. Assim, estamos perante dois problemas de otimização combinatória que devem ser resolvidos de uma forma integrada: Problema de rotas dos navios e problema de estiva de contentores. A integração destes dois problemas foi denominada por [Moura et al., 2012], como CSSRP (*Container Stowage and Ship Routing Problem*).
Este artigo tem a seguinte estrutura: secção 2 consiste numa descrição detalhada do CSSRP; na secção 3 é apresentada a revisão da literatura relevante para o trabalho desenvolvido; o modelo de programação linear inteira é apresentado na secção 4; na secção 5 apresentam-se os problemas de teste e resultados obtidos assim como uma análise de sensibilidade para validação do modelo; e para concluir, na secção 6, são sumariadas algumas conclusões globais do trabalho e sugestões de trabalho futuro.

2 Descrição do CSSRP

Os porta-contentores são navios especificamente construídos para o transporte de contentores. O espaço de carga está dividido em compartimentos, posicionados por cima e por baixo do convés (*deck*) do navio e separados por uma escotilha ([Wilson and Roach, 1999]). Cada compartimento ou bay (secção longitudinal) consiste num conjunto de pilhas ou *Tiers* (colunas - secções transversais) e níveis ou *stacks* (linhas - secções verticais), que têm a dimensão adequada para posicionamento de um contentor (Figura 1). O espaço definido para o posicionamento de um contentor será de agora em diante denominado por slot. Os porta-contentores têm restrições de capacidade, isto é, número máximo de contentores, peso máximo admissível e dimensões das slots. Cada contentor tem um determinado peso que não pode ser

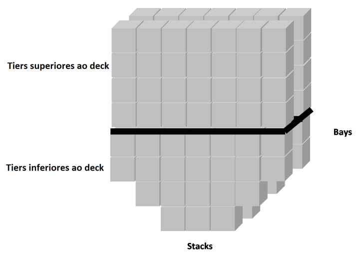


Figura 1: Secções de posicionamento.

excedido e dimensões standard. Essas dimensões são dadas em pés (ft) e podem ser TEU (*Twenty-foot Equivalent Unit*) ou FEU (*Forty-foot Equivalent Unit*). Em cada slot pode ser alocado um contentor de 40' ou dois contentores de 20'. Existem outros tipos de contentores que também vamos considerar neste problema que são, os refrigerados, que necessitam de uma ligação à corrente eléctrica (tomada eléctrica na slot) e outros que são considerados "perigosos"uma vez que transportam matéria nociva. O peso máximo dos contentores é 24 toneladas para os de 20' e 40 toneladas para os de 40'. No entanto existem outros contentores com outras características e dimensões que não estão descritos neste artigo, uma vez que não são considerados no nosso problema. Assim, o CSSRP pode ser caracterizado por um conjunto de portos, cada um com uma procura (conjunto de contentores) que tem de ser entregue dentro de um determinado período de tempo (deadlines). As decisões operacionais que devem ser consideradas são: (I) que portos devem ser visitados por cada porta-contentor e a respectiva sequência de visita; (II) como colocar os contentores a bordo considerando os requisitos de carga e a regra LIFO. O objetivo é minimizar o custo total de transporte. Como explicado em [Moura et al., 2012], os custos mais relevantes na distribuição marítima estão relacionados com o manuseamento dos contentores nos portos. Assim, o posicionamento dos contentores dentro do navio é crucial para reduzir este manuseamento, uma vez que o manuseamento depende diretamente do plano de estiva. O manuseamento é caraterizado por movimentos carga ou descarga de contentores num porto. Logo, consideremos que existe movimentação de contentores quando estes:

1. São descarregados porque chegam ao seu destino;

[1]Entenda-se por overstow, quando existem contentores (para um dado porto) que devem ser movimentados porque estão a bloquear o acesso a outros contentores que têm de ser descarregados nesse porto.

2. Têm de ser movimentados porque estão a bloquear o acesso a outros contentores que têm de ser descarregados (overstow);
3. Têm de ser reposicionados devido a garantir a estabilidade do navio e/ou porque facilitam o manuseamento no porto seguinte.

A movimentação de contentores é uma atividade que consome bastante tempo, por isso o posicionamento dos contentores a bordo é crucial para melhorar a eficiência operacional. É fácil de entender que quando é efetuado um movimento (ponto 2 ou 3) que não seja de carga ou descarga, o contentor é sempre colocado numa posição diferente de onde se encontrava e por isso o plano de carga é sempre alterado.
Outros custos relacionados com os portos são as taxas e custos de utilização. Existem também custos, relacionados com os navios, por exemplo custos de viagem que dependem dos custos de combustível/milha e custo de utilização dos navios que dependem da tripulação/dia. Tendo em consideração esta descrição do CSSRP facilmente se concluí que existem várias restrições que devem ser consideradas. Restrições essas relacionadas com os portos (restrições de planeamento de rotas), com os navios (restrições de carga e posicionamento) e restrições relacionadas com estes dois problemas em simultâneo (restrições planeamento de rotas e carga dos navios).

(R1) Restrições de rotas:

- Deadlines - Estas restrições podem ser modeladas como restrições lineares dos VRP com janelas temporais, onde o instante de chegada de um navio a um porto deve ser menor ou igual ao menor deadline dos contentores para esse porto.

(R2) Restrições de carga e posicionamento dos contentores:

- Capacidade dos navios - Relacionadas com o peso e número máximo de contentores: Os limites de peso e o número de slots, podem ser modelados como restrições lineares do problema da mochila, onde o somatório dos pesos e número de contentores nunca deve exceder a capacidade do navio;
- Posicionamento - Relacionadas com o posicionamento dos contentores nas slots e estabilidade da carga, onde os contentores podem ser directamente posicionados no convés ou escotilha do navio ou em cima de outros contentores.

(R3) Interligação dos problemas:

- Esta restrição implica que se um navio visita um determinado porto, então a procura desse porto encontra-se alocada no porta-contentores.

O desafio de combinar o planeamento das rotas com deadlines e o planeamento da estiva dos contentores, é poder aplicar esta abordagem a problemas reais obtendo soluções ótimas em tempos computacionais reduzidos, sendo esta a maior contribuição do presente trabalho. No entanto no problema apresentado neste trabalho, não são contabilizados considerados os overstows. Sendo as únicas movimentações: a carga e descarga dos contentores.

3 Revisão bibliográfica

O CSSRP tem recebido pouca atenção da comunidade científica. A grande maioria das publicações só consideram o CSP, algumas considerando a regra LIFO, mas com rotas pré-definidas. No entanto, a primeira publicação que resolve o CSP integrado com o VRP com deadlines de uma forma integrada é de [Martins et.al., 2010]. Mais tarde em 2012, [Moura et al., 2012], propõem um modelo programação linear inteira mista para o CSSRP mas com algumas considerações e simplificações ao problema. Essas simplificações estão relacionadas com os contentores. A distinção entre contentores de 20' e 40', normais, refrigerados e danger não é considerada. A procura de um porto é vista como um único contentor cujas dimensões e pesos é igual ao somatório das dimensões e pesos dos contentores da procura. Assim a matriz de posicionamento é consideravelmente simplificada e reduzida. No trabalho que aqui se apresenta e com o objetivo de adaptar o modelo apresentado em [Moura et al., 2012] a problemas reais, essas simplificações não são consideradas.
Relativamente ao que concerne o escalonamento e planeamento de rotas de navios, [Christiansen and Nygreen, 1998] apresentam e resolvem um problema, caracterizado pelos autores como: (*Inventory Pickup and Delivery Problem with Time Windows*) (IPDPTW). O modelo de programação matemática é resolvido usando a decomposição de Dantzig -Wolfe, decompondo a formulação inicial em sub-problemas para cada porto e para cada navio. A relaxação linear do problema principal é resolvida por geração de colunas, onde estas representam as rotas dos navios ou a sequência de visita dos portos. Por forma a fazer com que a solução inteira seja ótima o processo de iteração é embebido numa pesquisa em árvore. Mais tarde [Agarwal and Ergun, 2008] apresentam um modelo misto de programação linear inteira, para resolução do escalonamento de navios e problemas carga-rota em simultâneo. Neste trabalho, é desenvolvida

uma heurística gulosa, baseada em geração de colunas e um algoritmo de duas fases. Esta abordagem gera boas soluções para o problema em questão. Como já referido, [Martins et.al., 2010] apresenta um algoritmo genético para melhorar a flexibilidade da distribuição marítima de curta distância e aumentar a sua competitividade com outros meios de transporte. É apresentado um modelo logístico para gestão de uma frota de dois ou mais navios que transportam carga para e de vários portos, tendo em consideração o posicionamento da carga e respectivos deadlines. Ao contrário do CSSRP, o CSP também conhecido por (*Master Bay Problem*) (MBP) tem tido vários trabalhos publicados durante as três últimas décadas. O CSP é um problema NP-Difícil [Avriel, et al., 2000] e prende-se com a forma de alocação de contentores em porta-contentores. Pode ser categorizado como um problema de atribuição, onde um conjunto de contentores com um determinado porto destino e diferentes características, devem ser atribuídos a slots de um navio com o objetivo de minimizar o custo de transporte. Desde finais dos anos 80, foram publicados vários trabalhos abordando este problema. [Aslidis, 2000] foi o primeiro autor a resolver este problema considerando overstows e usando um algoritmo de programação dinâmica, abordagem esta utilizada posteriormente por outros autores. [Avriel and Penn, 1993] propuseram uma abordagem, *Whole Columns Heuristic Procedure* que permite encontrar a solução ótima para o problema de estiva num único bay retangular, considerando só restrições de acessibilidade. Esta heurística envolve um modelo de programação inteira mista após algum processamento dos dados. Este método provou ser limitado devido ao elevado número de variáveis binárias e restrições. Mais tarde, [Avriel et al., 1998], desenvolveu uma heurística (*Suspensory Heuristic Procedure*) que obtém resultados bastante satisfatórios num tempo computacional reduzido. Formulações de programação linear binária para o CSP com restrições de estabilidade, acessibilidade, etc., podem ser encontradas em [Botter and Brinati, 1992], [Ambrosino et al., 2004] e [Ambrosino et al., 2006]. Em todos estes trabalhos os autores concluem que é impossível obter soluções ótimas através de modelos de programação linear mista para o CSP. Um trabalho recente que envolve o CSP, de [Delgado et al., 2012] contradiz estas afirmações. Estes autores decompõem o problema e apresentam um modelo de programação linear inteira e programação dinâmica para carga de um conjunto de contentores numa única bay e resolvem problemas reais num tempo computacional razoável. Métodos de pesquisa tais como algoritmos genéticos [Dubrovsky et al., 2002], [Martins et.al., 2010] e pesquisa tabu [Wilson and Roach, 2000] foram também aplicados ao CSP. A vantagem de usar abordagens heurísticas e meta-heurísticas para resolver este tipo de problema foi provada através dos resultados obtidos com estes trabalhos. [Wilson and Roach, 1999] e mais tarde em [Wilson and Roach, 2000], apresentam um método de duas fases. Na primeira fase com o objetivo de reduzir overstows e movimentações de contentores nos portos destino, os contentores são agrupados pelo destino usando um algoritmo de pesquisa *Branch-and-Bound.* Em seguida, na segunda fase por forma a reduzir rearranjos da carga, tendo em consideração a estabilidade da carga, é aplicado um algoritmo de pesquisa tabu à solução geral que tenta mover contentores a alocá-los a uma slot específica. A primeira vez que o CSP foi comparado e caracterizado com um problema de empacotamentos (Bin Packing Problem) foi por [Wei-Ying et al., 2005]. Neste trabalho, o CSP é visto como um problema de empacotamento bidimensional, onde as bays são bins, o número de slots em cada bay é a capacidade dos bins e contentores com diferentes características são tratados como itens a empacotar. Foi desenvolvida uma abordagem em duas fases: na primeira fase são consideradas duas funções objetivo, uma para minimizar o número de bays e outra para minimizar o número de overstows. Em seguida os contentores atribuídos a cada bay (na primeira fase) são na segunda fase alocados a slots aplicando um algoritmo de pesquisa tabu e considerando restrições de peso, estabilidade e overstows.

4 Modelo matemático do CSSRP

Nesta secção é apresentado o modelo de programação inteira mista que tem como objetivo a geração de soluções ótimas para o CSSRP. Foi desenvolvido um modelo matemático para gestão de uma frota de porta-contentores que transporta carga para vários portos, tendo em consideração a estiva e deadlines de entrega. Cada porto está caraterizado pela sua localização geográfica e é representado por um conjunto de nós de um grafo $G(P, A)$, onde $P = \{1, ..., p\}$ representa o conjunto de portos e $A = \{(i, j) : i, j \in P, i \neq j\}$ o conjunto de arestas em G. O comprimento de cada arco d_{ij} corresponde à distância entre portos em milhas. u_{ik} é o custo em euros de visita a um porto i pelo navio k, que depende das taxas portuárias e dos custos de utilização de cada porto. Assume-se que o porto inicial é o porto 1 e só neste porto se efetuam cargas, nos restantes só descargas, o custo de visita deste porto é também considerado. Os dados dos navios estão relacionados com as suas características: $V = \{1, ..., v\}$ é o conjunto de navios. Assume-se que todos os navios possuem uma velocidade de navegação constante,(vel_k). c_k é o custo operacional do navio k (por unidade de distância, em euros) que depende do custo de combustível/milha; e um custo com tripulação para a utilização do navio k (uc_k) (por dia, em euros). Q_k é a capacidade de peso do navio k e αMax_k é a capacidade em número de contentores do navio k (TEU e FEU). t_{ik} é uma estimativa do tempo de serviço do navio k no porto i. Tal como explicado na secção 1, considera-se que os contentores têm dimensões standard (TEU e FEU) e três tipos de contentores, normais, refrigerados e perigosos. Cada porto i tem uma determinada procura definida pelo número de diferentes tipos de contentores e inferior à capacidade do navio. Para cada porto a procura é $\alpha_i = \alpha t_i + \alpha f_i + \alpha r t_i + \alpha r f_i + \alpha d t_i + \alpha d f_i$ onde $\alpha t_i, \alpha f_i, \alpha r t_i, \alpha r f_i, \alpha d t_i, \alpha d f_i$ são o número de contentores de 20', 40', refrigerados de 20', refrigerados de 40', perigosos de 20' e perigosos de 40', respetivamente, para o porto i. $\alpha Total = \sum_{i \neq 1}^{p} \alpha_i$ é a procura total de cada porto i. Cada tipo de contentor numa determinada procura possui um determinado peso $qt_i, qf_i, qrt_i, qrf_i, qdt_i, qdf_i$ para 20', 40', refrigerados de 20', refrigerados de 40', perigosos de 20' e perigosos de 40', respetivamente. Além disto, a procura de cada porto tem um deadline associado que é igual ao menor dos deadlines dos contentores para esse porto (dl_i). Devido à configuração irregular do espaço de carga de um navio, as várias bays são associadas a matrizes $posM_{kcs}$ de slots (Figura 2), onde cada slot está posicionada numa fila $C = \{1, ..., t\}$ e uma coluna $S = \{1, ..., r\}$. Cada slot de valor igual a zero corresponde a um local possível de posicionamento de um contentor de 40' ou de dois contentores de 20'. Uma slot com valor igual a 1 é uma slot que não pode ser usada. A matriz $Plugs_{kcs}$ é usada para definição dos posicionamentos das tomadas elétricas para colocação de contentores refrigerados (Figura 3). Uma terceira matriz $Danger_{kcs}$ é usada para definir as posições atribuídas ao posicionamento de contentores perigosos (Figura 4). Os contentores normais podem ser

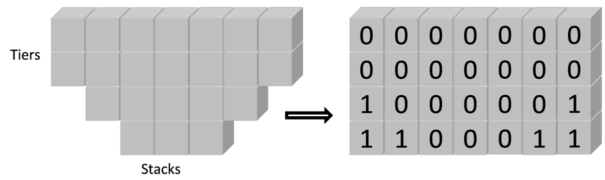


Figura 2: Matriz de posicionamentos.

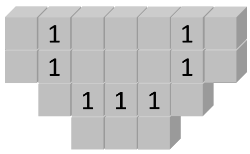


Figura 3: Matriz de tomadas elétricas.

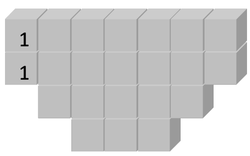


Figura 4: Matriz de contentores perigosos.

colocados em qualquer uma destas posições. No entanto, quando o número de contentores de refrigerados e de danger é igual ao número de posições dedicadas a estes, os normais não podem ser colocados nestas células. As slots com tomadas elétricas só podem levar um contentor refrigerado uma vez que só existe em cada slot uma única tomada. Assim neste caso, esta slot pode ser ocupada ou por um contentor refrigerado de 40', ou por um refrigerado de 20' e outro normal de 20'. No caso dos contentores perigosos, estas têm áreas específicas nos porta-contentores. Foi criada também uma matriz para o posicionamento final de qualquer tipo de contentor $FinalM_{kics}$ que irá conter o conjunto de variáveis binárias usadas para indicar em que slot de um navio é colocado um contentor para um determinado porto, $t_{kics}, f_{kics}, rt_{kics}, rf_{kics}, dt_{kics}, df_{kics}$. Onde, t_{kics} é igual a um se um contentor normal de 20' da procura do porto i é carregado na slot (C, S) do navio k, e igual a zero senão. As restantes variáveis estão relacionadas com os contentores normais de 40', refrigerados de 20' e de 40' e perigosos de 20' e de 40', respetivamente. $(t_{kics} + f_{kics} + rt_{kics} + rf_{kics} + dt_{kics} + df_{kics})$. Como já foi mencionado, o custo de um movimento simples (carga ou descarga) é bastante importante para o custo final da distribuição, logo na função objetivo, m_α representa o custo incorrido num movimento simples de um contentor. Apresentados os dados, vamos agora definir as variáveis de decisão e variáveis auxiliares do problema. Este modelo considera a variável binária x_{ijk} que assume valor um se o navio k percorre o arco (i,j) e zero se não. O navio deve chegar aos portos dentro de um determinado período de tempo, por isso, s_{ik} é uma variável de decisão que indica o instante de chegada do navio k ao porto i. O modelo de programação inteira mista é definido como:

$$Min \left[\sum_{k=1}^{v} \sum_{i=1}^{p} \sum_{j=1 \wedge j \neq i}^{p} (d_{ij} \times c_k \times x_{ijk} + x_{ijk} \times u_{jk} + x_{ijk} \times \alpha_j \times m_\alpha) + alphaTotal \times m_\alpha + \sum_{k=1}^{v} uc_k \times s_{1k} \right] \tag{4.1}$$

s.a: Restrições das rotas:

$$\sum_{i=1, i \neq j}^{p} x_{ijk} \leq 1 \qquad \forall j \neq 1 \in P, \forall k \in V \tag{4.2}$$

$$\sum_{j=1, j \neq i}^{p} x_{ijk} \leq 1 \qquad \forall i \neq 1 \in P, \forall k \in V \tag{4.3}$$

$$\sum_{i=1, i \neq j}^{p} (x_{ijk} - x_{jik}) = 0 \qquad \forall k \in V, \forall j \in P \tag{4.4}$$

$$\sum_{j=1}^{p} x_{1jk} \leq 1 \qquad \forall k \in V \tag{4.5}$$

$$\sum_{k=1}^{v} x_{ijk} \leq 1 \qquad \forall i, j \neq i \in P \tag{4.6}$$

$$s_{ik} + t_{ik} + \frac{d_{ij}}{vel_k} \leq s_{jk} + M\,(1 - x_{ijk}) \qquad \forall k \in V, \forall i, j \in P \tag{4.7}$$

$$s_{ik} \leq dl_i \qquad \forall k \in V, \forall i \in P \tag{4.8}$$

Restrições de carga e posicionamento de contentores:

$$\sum_{j \neq 1}^{p} (\alpha_j \times x_{ijk}) \leq \alpha Max_k \qquad \forall k \in V, \forall i \in P \tag{4.9}$$

$$\textstyle\sum_{j \neq 1 \wedge j \neq i}^{p} ((qt_j \times \alpha t_j + qf_j \times \alpha f_j + qrt_j \times \alpha rt_j + qrf_j \times \alpha rf_j + qdt_j \times \alpha dt_j + qdf_j \times \alpha df_j) \times x_{ijk}) \leq Q_k \quad \forall k \in V, i \in P \tag{4.10}$$

$$t_{kics} + 2 \times posM_{kcs} \leq 2 \qquad \forall k \in V, \forall i \neq 1 \in P, \forall c \in C, \forall s \in S \tag{4.11}$$

$$f_{kics} + posM_{kcs} \leq 1 \qquad \forall k \in V, \forall i \neq 1 \in P, \forall c \in C, \forall s \in S \tag{4.12}$$

$$rt_{kics} - Plugs_{kcs} \leq 0 \qquad \forall k \in V, \forall i \neq 1 \in P, \forall c \in C, \forall s \in S \tag{4.13}$$

$$rf_{kics} - Plugs_{kcs} \leq 0 \qquad \forall k \in V, \forall i \neq 1 \in P, \forall c \in C, \forall s \in S \tag{4.14}$$

$$dt_{kics} - 2 \times Danger_{kcs} \leq 0 \qquad \forall k \in V, \forall i \neq 1 \in P, \forall c \in C, \forall s \in S \tag{4.15}$$

$$df_{kics} - Danger_{kcs} \leq 0 \qquad \forall k \in V, \forall i \neq 1 \in P, \forall c \in C, \forall s \in S \tag{4.16}$$

$$\sum_{i \neq 1}^{p} (t_{kics} + 2 \times f_{kics}) \leq 2 \qquad \forall k \in V, \forall c \in C, \forall s \in S \tag{4.17}$$

$$\sum_{i \neq 1}^{p} (rt_{kics} + rf_{kics}) \leq 1 \qquad \forall k \in V, \forall c \in C, \forall s \in S \tag{4.18}$$

$$\sum_{i \neq 1}^{p} (dt_{kics} + 2 \times df_{kics}) \leq 2 \qquad \forall k \in V, \forall c \in C, \forall s \in S \tag{4.19}$$

$$\sum_{i\neq 1}^{p}(t_{kics} + 2 \times rf_{kics}) \leq 2 \qquad \forall k \in V, \forall c \in C, \forall s \in S \tag{4.20}$$

$$\sum_{i\neq 1}^{p}(rt_{kics} + f_{kics}) \leq 1 \qquad \forall k \in V, \forall c \in C, \forall s \in S \tag{4.21}$$

$$\sum_{i\neq 1}^{p}(t_{kics} + 2 \times rt_{kics}) \leq 2 \qquad \forall k \in V, \forall c \in C, \forall s \in S \tag{4.22}$$

$$\sum_{i\neq 1}^{p}(f_{kics} + rf_{kics}) \leq 1 \qquad \forall k \in V, \forall c \in C, \forall s \in S \tag{4.23}$$

$$\sum_{i\neq 1}^{p}(t_{kics} + 2 \times df_{kics}) \leq 2 \qquad \forall k \in V, \forall c \in C, \forall s \in S \tag{4.24}$$

$$\sum_{i\neq 1}^{p}(t_{kics} + dt_{kics}) \leq 2 \qquad \forall k \in V, \forall c \in C, \forall s \in S \tag{4.25}$$

$$\sum_{i\neq 1}^{p}(rt_{kics} + df_{kics}) \leq 1 \qquad \forall k \in V, \forall c \in C, \forall s \in S \tag{4.26}$$

$$\sum_{i\neq 1}^{p}(rf_{kics} + df_{kics}) \leq 1 \qquad \forall k \in V, \forall c \in C, \forall s \in S \tag{4.27}$$

$$\sum_{i\neq 1}^{p}(rt_{kics} + dt_{kics}) \leq 1 \qquad \forall k \in V, \forall c \in C, \forall s \in S \tag{4.28}$$

$$\sum_{i\neq 1}^{p}(rf_{kics} + dt_{kics}) \leq 1 \qquad \forall k \in V, \forall c \in C, \forall s \in S \tag{4.29}$$

$$\sum_{i\neq 1}^{p}(t_{kics} + dt_{kics}) \leq 1 \qquad \forall k \in V, \forall c \in C, \forall s \in S \tag{4.30}$$

$$\sum_{i\neq 1}^{p}(f_{kics} + df_{kics}) \leq 1 \qquad \forall k \in V, \forall c \in C, \forall s \in S \tag{4.31}$$

$$\begin{gathered}\textstyle\sum_{i\neq 1}^{p}[(rt_{kics} + t_{kics} + 2 \times f_{kics} + 2 \times rf_{kics} + dt_{kics} + 2 \times df_{kics} + 2 \times posM_{kcs}) - \\ -(rt_{kias} + t_{kias} + 2 \times f_{kias} + 2 \times rf_{kias} + dt_{kias} + 2 \times df_{kias} + 2 \times posM_{kas})] \geq 0 \\ \forall k \in V, \forall c, a < c \in C, \forall s \in S\end{gathered} \tag{4.32}$$

Restrições de rotas e carga dos navios:

$$\sum_{i=1}^{p} x_{ijk} \times \alpha_j + \sum_{l=1}^{p} x_{jlk} \times \alpha_j = 2 * \times \sum_{c=1}^{t}\sum_{s=1}^{r} finalM_{kjcs} \qquad \forall k \in V, \forall j \neq 1 \in P \tag{4.33}$$

$$x_{ijk} \in \{0,1\}, s_{ik} \geq 0, t_{kics} \in \{0,1\}, f_{kics} \in \{0,1\}, dt_{kics} \in \{0,1\}, \tag{4.34}$$

$$df_{kics} \in \{0,1\}, rt_{kics} \in \{0,1\}, rf_{kics} \in \{0,1\} \tag{4.35}$$

O objetivo é a minimização do custo total da distribuição (equação 1). O primeiro conjunto de restrições, estão relacionadas com as rotas. Restrições 2, 3 e 6 asseguram que um porto é visitado por um único veículo. Equação 4 garante que se um navio chega a um porto então tem de sair desse porto para um outro. A equação 5 garante que se um navio é utilizado, deve iniciar a sua rota no porto inicial. O grupo de restrições R1 (secção 2), é obtido com as restrições 7 e 8. Estas garantem que um serviço num porto nunca inicia antes do navio chegar a esse porto, onde M é uma constante Big-M e que os deadlines dos contentores não são violados, respetivamente. O conjunto seguinte de equações está relacionado com as restrições de carga e posicionamento. Grupo de restrições R2 (secção 2) estão definidas pelas equação 9 e 10. A primeira garante que a procura total de uma rota não excede a capacidade do navio em número de contentores e a segunda em peso. As restantes restrições deste grupo estão relacionadas com o posicionamento dos contentores no navio. Como já foi mencionado existem três matrizes de posicionamento diferentes, uma relativa às possíveis posições para todos os contentores, outra que indica as slots com energia eléctrica e a última com indicação das slots para colocação dos contentores perigosos. Assim, um conjunto de restrições, desde equação 11 até à 31 são necessárias para garantir que os contentores só são colocados em slots apropriadas, contentores refrigerados em slots com tomadas eléctricas (13 -14), perigosos em local específico (15-16) e normais em todas as posições(11-12) [Delgado et al., 2012]. As restrições 17-31 garantem as possibilidades de combinações de diferentes tipos de contentores na mesma célula da slot. O conjunto de restrições definidas pelas equações 17-31 juntamente com a restrição 32 garantem que a estiva dos contentores é feita corretamente garantindo a estabilidade da mesma, através da colocação da carga de baixo para cima, não colocando carga suspensa. Os dois problemas, planeamento de rotas e estiva de contentores (restrição R3, secção 2), são integrados com a equação 33, que liga as variáveis de estiva de contentores com as variáveis do problema de rotas, isto é, se um navio visita um dado porto então a procura desse porto tem de estar nesse mesmo navio.

5 Resultados computacionais e análise de sensibilidade

As instâncias usadas para testar e validar o modelo apresentado anteriormente, foram desenvolvidas com base nos dados apresentados por [Martins et al., 2010] e posteriormente por [Moura et al., 2012], trabalhos estes que abordam problemas de *"Short Sea Shipping"*. As instâncias consideram portos Europeus cada um deles com uma determinada procura que deve ser satisfeita dentro de um limite de tempo. As distâncias entre portos são conhecidas assim como as taxas aplicadas em cada um deles. Para a distribuição existe uma frota de porta-contentores com características idênticas em termos de capacidade e custos de utilização que variam com a tripulação necessária e com o tempo de navegação. Os diferentes cenários em termos de dados utilizados por [Martins et al., 2010] foram recolhidos do histórico do porto marítimo de Viana do Castelo - Portugal. Uma vez que existem diferenças entre os problemas tratados nestes dois trabalhos e no presente, foram feitas algumas alterações e adaptações aos dados. Ao contrário de [Martins et al., 2010] e tal como em [Moura et al., 2012], no nosso problema os navios são carregados com os contentores no primeiro porto e nos outros portos só são efetuados movimentos de descarga. Além disso, para provar a robustez deste modelo, foram geradas instâncias maiores. Tal como em [Martins et al., 2010] e em [Moura et al., 2012] o problema foi simplificado assumindo que: todas as taxas portuárias são iguais; o consumo de combustível e a velocidade dos navios é sempre o mesmo independentemente da carga dos mesmos. Os problemas de teste diferem no número de portos e de navios. Cada instância tem uma capacidade fixa para todos os navios em termos de peso e número de contentores e a procura é sempre igual ou superior à capacidade de um navio. Todos os portos têm deadlines dl_i definidos. Os problemas de teste foram resolvidos usando o CPLEX, e testados num Intel CORE i7 vPro 2,2GHz com 8Gb de memória. Figura 5, mostra a solução ótima obtida para um exemplo com 5 portos e um navio num tempo computacional de 0,58 segundos. Por forma a analisar o comportamento e robustez

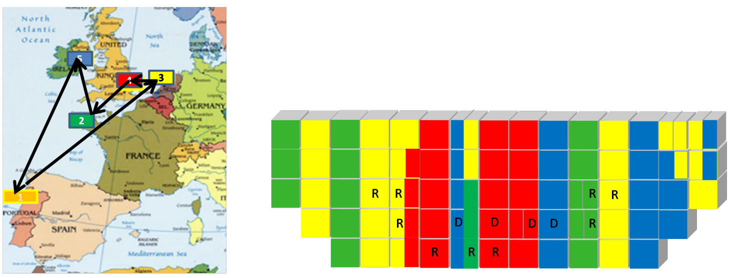


Figura 5: Exemplo da solução de um problema cinco portos e um navio.

do modelo, foram criadas novas instâncias onde se varia o número de portos e navios (Tabela 1). A Tabela 1 apresenta alguns resultados computacionais onde as primeiras duas colunas representam as características das instâncias em termos de número de navios e número de portos. Nas restantes colunas consideraram-se várias medidas de performance. Nas colunas 3 e 4 são apresentados os valores da função objetivo da solução inteira ótima e tempos computacionais necessários para obter essa solução. Na coluna 5 é apresentado o GAP. O GAP é determinado através da equação: $GAP = (Z^*_{pi} - Z^*_{rl})/Z^*_{pi}$, onde Z^*_{pi} é o valor da solução ótima inteira e Z^*_{rl} está associado com a solução ótima da relaxação linear. Com

Navios	Portos	Sol. óptima (euros)	Tempo comput. (seg)	GAP	Variáveis	Restrições
1	5	120085,36	1,09	65,1	1831	12026
2	5	158057,16	45,10	98,8	3661	22208
1	10	197481,78	1,06	56,5	4161	25361
2	10	-	+15min	-	8321	46528

Tabela 1: Resultados computacionais

o objectivo de mostrar a dimensão do modelo MIP, em termos de número de variáveis e restrições, as duas últimas colunas apresentam os valores para o conjunto de instâncias testadas. Para o conjunto de problemas testados o modelo encontrou sempre a solução ótima. O tempo computacional aumenta com o aumento do número de portos e de navios. Para os problemas com um navio, o aumento do número de portos não resulta num aumento significativo do tempo computacional. No entanto para o mesmo número de portos aumentando o número de navios, o aumento do tempo computacional já é bastante significativo. De facto, para dez portos e dois navios, o tempo computacional foi superior a 15 minutos. De acordo com [Delgado et al.,2012] em problemas de distribuição marítima, mais concretamente CSP, tempos de resposta superiores a dez minutos são considerados excessivos. Assim sendo para a última instância testada não é adequada a utilização deste modelo. Em particular com estes problemas testados verifica-se que o GAP diminui com o aumento do número de portos. O mesmo comportamento não se verifica quando se aumenta o número de navios. Foi também estudado o impato da variação da distribuição da quantidade de cada uma das procuras e da variação dos tempos de entrega (deadlines). Para isso, foi feita uma análise de sensibilidade ao modelo. A intenção é explorar os efeitos destas variações nos tempos computacionais. Se, por exemplo, a procura para um dado porto é igual ou muito próxima da capacidade de um navio, esse navio só visitará esse (um) porto, tornando o problema potencialmente mais simples de resolver devido à diminuição de combinações possíveis para as rotas. Por outro lado, quando as cargas são muito semelhantes ou mesmo iguais, as combinações possíveis para o seu posicionamento aumenta significativamente. Assim, foram criadas mais duas distribuições de cargas por procura além da já testada (e que consideramos homogénea) para problemas com 5 portos e 2 navios: fracamente heterogénea e fortemente heterogénea. A distribuição de carga homogénea é quando o número de contentores e respectivos pesos é quase igual em todas as procuras. Por outro lado fortemente heterogéneas são quando essas diferenças são bastante significativas, e fracamente heterogéneas quando existem diferenças ligeiras entre elas. Relativamente aos deadlines, também se consideram além dos deadlines normais dos problemas anteriores, deadlines pequenos e grandes. As tabelas 2 e 3, apresentam os resultados obtidos com estas variações. Como se pode verificar, o tempo computacional necessário para encontrar a solução

Procura	Sol. óptima (euros)	Tempo comput. (seg)
Fortemente Heterogénea	190232,37	10,25
Homogénea	158057,16	45,10
Fracamente Heterogénea	164098,43	8,78

Tabela 2: Alteração da procura

Deadlines	Sol. óptima (euros)	Tempo comput. (seg)
Pequenos	158057,16	30,05
Normais	158057,16	45,10
Grandes	158057,16	45,13

Tabela 3: Alteração dos deadlines

ótima dos problemas é bastante inferior para cargas fracamente heterogéneas. Além disso o maior tempo computacional está sempre relacionado com os problemas com cargas homogéneas. No que concerne aos deadlines, as variações não são muito significativas, no entanto é de realçar que para grandes deadlines, o tempo computacional é sempre superior, o que se explica pelos mesmos motivos descritos anteriormente para o caso das distribuições das cargas.

6 Conclusões

Neste trabalho foi apresentado um modelo de programação linear mista para problemas de distribuição de curta distância caracterizados como CSSRP. Este problema pode ser abordado como uma integração de dois problemas NP-Difíceis: VRP e CSP. No entanto no CSSRP a procura tem deadlines e existem locais fixos para a colocação dos contentores nos navios. Apesar da complexidade deste problema e do modelo matemático apresentado, foi provado que este pode ser aplicado na resolução de problemas de distribuição marítima de curta distância. O modelo foi testado com vários problemas de teste (dados reais) e provou-se que para estes problemas com estas dimensões, se obtém sempre a solução ótima num tempo computacional bastante reduzido. No entanto não é considerado no modelo a minimização dos overstows. Esta consideração é de grande importância, uma vez que a existência de overstows implica um aumento substancial nos custos finais da distribuição. De qualquer forma, pelas soluções encontradas e devido à forma que é indicada no modelo para colocação da carga nos navios, verificamos em todas as instâncias a não existência de overstows. O que de qualquer forma nunca poderemos afirmar que isto se verificará sempre em qualquer problema testado, a não ser que esta restrição seja incluída no modelo.

Referências

[Agarwal and Ergum, 2008] Agarwal,R. and Ergum,O., (2008). Ship scheduling and network design for cargo routing in linear shipping. *Transportation Science*, 42:175–196.

[Ambrosino et al., 2004] Ambrosino,D., Sciomachen,A. and Tanfani,E., (2004). Stowing a containership: the master bay plan problem. *Transportation Research Part A*, 38:81–99.

[Ambrosino et al., 2006] Ambrosino,D., Sciomachen,A. and Tanfani,E., (2006). SA decomposition heuristics for the container ship stowage problem. *Journal of Heuristics*, 12:211–233.

[Alidis, 2000] Aslidis,A.H., (2000). Combinatorial algorithms for stacking problems. *in PhD dissertation, MIT*.

[Avriel and Penn, 1993] Avriel,M., and Penn,M., (1993). Exact and approximate solutions of the container ship stowage problem. Computers and Industrial Engineering, 25:271–274.

[Avriel et al., 2000] Avriel,M., Penn,M., and Shpirer,N., (2000). Container ship stowage problem: complexity and connection to the coloring of circle graphs. Discrete Applied Mathematics, 103:271–279.

[Avriel et al., 1998] Avriel,M., Penn,M., Shpirer,N., andWitteboon,S., (1998). Stowage planning for container ships to reduce the number of shifts. Annals of Operations Research, 76:55–71.

[Botter and Brinati, 1992] Botter,R.C., and BrinatiM.A., (1992). Stowage container planning: a model for getting an optimal solution. Computer Applications in the Automation of Shipyard Operation and Ship Design, VII, North-Holland: 217–229.

[Christian and Nygreen, 1998] Christiansen,M., and Nygreen,B., (1998). A method for solving ship routing problems with inventory constraints. Annals of Operational Research, 81:357–378.

[Delgado et al., 2012] Delgado,A., Jensen,R.M., Janstrup,K., Rose,T.H. and Andersen,K.H., (2012). A Constraint Programming model for fast optimal stowage of container vessel bays. European Journal of Operational Research, 220:251–261.

[Dubrovsy et al., 2002] Dubrovsy,O., Levitin,O.G. and Penn,M., (2002). A genetic algorithm with a compact solution encoding for the container ship stowage problem Journal of Heuristics, 8:585–599.

[Gendrau et al., 2006] Gendreau,M., Iori,M., Laporte,G., and Martello,S., (2006). A tabu search algorithm for a routing and container loading problem. Transportation Science, 9(3):342–350.

[Martins et al., 2010] Martins,T., Moura,A., Campos,A.A., and Lobo,V., (2010). Genetic algorithms approach for containerships fleet management dependent on cargo and their deadlines. IAME 2010: Annual Conference of the International Association of Maritime Economists, Lisbon 7-9 July.

[Moura and Oliveira, 2009] Moura, A. and Oliveira, J.F., (2009). An integrated approach to the vehicle routing and container loading problems. Operations Research Spectrum, 31:775–800.

[Moura et al., 2012] Moura,A., Oliveira,J., Pimentel,C., (2012). A Mathematical Model for the Container Stowage and Ship Routing Problem. Journal of Mathematical Modeling and Algorithms, pp. 1–15; DOI: 10.1007/s10852-012-9207-3.

[Wilson and Roach, 1999] Wilson,I.D., and Roach,P.A., (1999). Principles of combinatorial optimization applied to container-ship stowage planning Journal of Heuristics 5(4):403–418.

[Wilson and Roach, 2000] Wilson,I.D., and Roach,P.A., (2000). Container stowage planning: a methodology for generating computerized solutions. Journal of the Operational Research Society, 51:1248–1255.

[Wei-ying, et al., 2005] Wei-ying, Z., Yan, L. and Zhuo-shang, J.T., (2005). Model and algorithm for container ship stowage planning based on Bin-packing problem. Journal of Marine Science and Application, 4(3):30–36.

Tactical and Operational Planning in Reverse Logistics Systems with Multiple Depots

Tânia Ramos*, Maria Isabel Gomes†, Ana Paula Barbosa-Póvoa‡

* *Instituto Universitário de Lisboa (ISCTE-IUL), Av. das Forças Armadas, 1649 - 026 Lisbon, Portugal*
E-mail:`tania.ramos@iscte.pt`

† *CMA – FCT, Universidade Nova de Lisboa, Campus da Caparica, 2829 - 516 Caparica, Portugal,*
E-mail(segundo autor):`mirg@fct.unl.pt`

‡ *CEG – IST, Universidade Técnica de Lisboa, Av. Rovisco Pais, 1049 - 001 Lisbon, Portugal.*
E-mail (terceiro autor):`apovoa@ist.utl.pt`

Abstract

This work develops new mixed-integer linear programming models and new solution's approaches to support tactical and operational planning decisions in reverse logistics systems involving multiple depots. Depots' service areas delimitation, routes' definition and scheduling, CO_2 emissions quantification and drivers labour hours balance were addressed. With all these aspects in mind, the contribution of this work is to build the basis for a solution tool that supports a sustainable operation of reverse logistics networks. Namely, by increasing efficiency of recyclable waste collection systems, while diminishing their environmental impacts and increasing social concerns. The models were applied to different real case studies.

Keywords: Reverse logistics, Waste collection, Multiple depots, Sustainability, Service areas, Routing, MILP models.

1 Introduction

Reverse logistics can be defined as "the process of planning, implementing and controlling backward flows of raw materials, in-process inventory, packaging and finished goods, from a manufacturing, distribution or use point, to a point of recovery or point of proper disposal" (de Brito and Dekker, 2004). Following this definition, reverse logistics concentrates on streams where there is some value to be recovered, entering the outcome into a new/existent supply chain. Among other recovery options one is the recycling which, imposed by the European Union (EU), has forced member states to develop new collection systems. The traditional routes defined for undifferentiated waste do not fit the particularities of recyclable materials and different vehicles, different collection rates and different bin locations are required. This situation motivated the surge of two different waste collection systems: selective and undifferentiated. The costs involved in selective collection are higher than in undifferentiated collection, but recycling targets have anyway to be accomplished. On the other hand, recycling helps to protect the environment since it mitigates resource scarcity, decreases demand for landfill space and involves savings in energy consumption. But, the activity of collecting the recyclable waste is mainly a transportation activity, implying Greenhouse Gas emissions (GHG), resource consumption, and noise, amongst other negative impacts to the environment. Moreover, several human resources take part of the collection activities and an equity workload distribution among them should be planned. To respond to this challenge waste collection companies must invest on the effective planning and operation of their logistics structures while considering economic, environmental and social objectives. Due to the complexity involved at the planning and operation of such systems, tools that may support the planners' decision, within such companies, are required and represent a challenge to the academic community. Since this research was triggered off by three waste collection systems case-studies where strategic decisions have already been taken, the focus of the work is on tactical and operational decisions and the aim is to contribute to the development of solution approaches that may help decision makers of reverse logistics systems.

The innovative aspects of this research rely on the study of three characteristics of logistics networks that have been seldom studied: type of routes (closed *versus* open), number of products (single *versus* multiple) and objective function (economical *versus* environmental *versus* social). The majority of the published works has focused on closed routes, single-product and deal with an economical objective function. This work goes further by studying also open routes, multiple products and by dealing with environmental and social objectives in addition to the traditional economical objective.

Given the goal of defining service areas and vehicle routes simultaneously, the baseline model is the so-called Multi-Depot Vehicle Routing Problem (MDVRP). Different variants are then addressed, such as open routes, inter-depot routes, periodic, multiple products, and environmental and social objective functions. New models and solution's approaches are developed to address each problem. The models developed are validated by literature instances and by real case studies, where the collection of end-of-life products is planned in three real reverse logistics systems operating in Portugal.

The remainder of the paper is structured as follows. In section 2 a brief review of literature on routing problems and its variants is presented. In the following sections the main results of each studied variant are described. Section 3 addresses the MDVRP with Open Routes, section 4 the MDVRP with Inter-Depot Routes, section 5 the MDVRP with Multiple Products and Economic and Environmental concerns. Section 6 addresses a multi-objective Multi-Product, Multi-Depot Periodic Vehicle Routing Problem with Inter-Depot Routes. Final remarks and future work directions are drawn in section 7.

2 Literature Review

Routing problems consist on defining the optimal delivery or collection routes from a central depot to a set of geographically scattered customers, subject to various constraints. Such problems are common to a large range of logistics systems and have a significant economic impact on the planning and operation of these systems (Laporte, 2007). However, the logistics systems can have diversified features giving rise to different variants of routing problems. It may involve a single or multiple depots, homogeneous or heterogeneous vehicle fleet, stochastic or deterministic demand, closed or open routes, among others.

The MDVRP appears as a generalization of the Vehicle Routing Problem (VRP) where beyond the definition of vehicle routes, the allocation of customers to depots must also be determined. Therefore, the MDVRP simultaneously establishes the service areas of each depot and the associated vehicle routes. The vehicle routes are defined such that: (1) each route starts and ends at the same depot; (2) each customer is visited exactly once by a vehicle; (3) the total demand of each route does not exceeds the vehicle capacity; (4) the total duration of each route, including travel and service times, does not exceeds a pre-set time limit. The best solution is the one that minimizes total routing cost.

Several models have been developed for the MDVRP, exploring both exact and approximate approaches. However, since this is a NP-hard combinatorial problem, the models proposed in the literature are predominantly heuristics-based. Few exact algorithms have been present in the literature. Laporte et al. (1984), as well as Laporte et al. (1988), developed exact branch and bound algorithms for solving the symmetric and asymmetric versions of the MDVRP, respectively. More recently, Baldacci and Mingozzi (2009) developed an exact method for solving the Heterogeneous Vehicle Routing Problem (HVRP) that is capable of solving, amongst other problems, the MDVRP. On the other hand, when analysing the heuristic algorithms to solve MDVRP, several ones have been proposed such as: Tillman and Cain (1972), Golden et al. (1977), Chao et al. (1993), Renaud et al. (1996), Cordeau et al. (1997), Salhi and Sari (1997), Lim and Wang (2005), Parthanadee and Logendran (2006), Crevier et al. (2007), Pisinger and Ropke (2007), Ho et al. (2008), Dondo and Cerda (2009), among others.

When considering that the vehicle may not return to the route starting point, an open route is defined (Schrage, 1981). In the Open Vehicle Routing Problem (OVRP), vehicle routes start at the depot and end at one of the customers, defining routes as paths and not cycle tours. This problem often appears when the vehicle fleet is hired and the contractors are paid based on the kilometres driven. Since this problem appears as a common problem in real logistics systems, its study by academia has been intensified in the last decade. Several methods have been proposed where again heuristics and exact approaches have been explored. Sariklis and Powell (2000), Tarantilis and Kiranoudis (2002), Brandao (2004), Fu et al. (2005), Repoussis et al. (2007), Salari et al. (2010) are examples of some works where heuristic approaches have been proposed for solving the OVRP or its variants. On the exact methods, Bektas and Elmastas (2007) developed an integer formulation for solving the OVRP with capacity and distance constraints, which is applied to a real-life school bus routing problem. Also, Letchford et al. (2007) looked into the exact approaches and presented a branch-and-cut algorithm for the capacitated open vehicle routing problem.

While the MDVRP considers a planning horizon of just one time unit, the Periodic Vehicle Routing Problem (PVRP) considers a planning horizon of several time units giving that customers have different delivery (or collection) patterns. In this problem, a customer specifies a service frequency and a set of allowable delivery patterns, and the company has to decide on which day the delivery will occur. These two problems (the MDVRP and PVRP) have received a great deal of attention, but the combination of them – the Multi-Depot Periodic Vehicle Routing Problem (MDPVRP) has rarely been studied in the literature, and consequently only few models have been developed (see the works of Hadjiconstantinou and Baldacci (1998), Parthanadee and Logendran (2006) and Vidal et al. (2012)).

In all the above works, the objective function was defined as the minimization of either the total distance travelled, total time or the total routing cost, being the latter objective in most cases a linear function of distance or time. Nonetheless, some recent works have also explored environmental issues in vehicle routing problems (Bektas and Laporte (2011), Erdogan and Miller-Hooks (2012)). The combination of the three dimensions of sustainability has never, to the best of our knowledge, been approached in vehicle routing problems with multiple depots.

3 The Multi-Depot Vehicle Routing Problem with Open Routes

The Multi-Depot Vehicle Routing Problem with Open Routes (MDVRP-OR) has not yet been studied in the literature. In a logistics system with multiple depots, route's starting location can be different from the ending one, but all routes have to start and end at one of the network depots. Therefore, vehicle routes can be Hamiltonian cycles (closed routes) or just Hamiltonian paths between two depots (open routes), being both admissible. This study has been motivated by a real-life problem of a waste cooking oil collection system characterized by the existence of multiple depots with an outsourced vehicle fleet. The objective of the company is to minimize the total cost involving the number of vehicle routes required and the total distance travelled when visiting all the collection sites.

The MDVRP-OR builds k vehicle routes in such a way that: (1) each route starts and ends at a depot (not necessarily the same); (2) each collection site is visited exactly once by a vehicle; (3) the total demand of each route does not exceed the vehicle capacity Q ; (4) the total duration of each route, including travel and service times, does not exceeds a pre-set time limit T ; so that (5) the total routing cost is minimized.

To formulate the MDVRP-OR based on the two-commodity flow formulation we use the same decision variables as Baldacci et al. (2004) - x_{ij} and y_{ij} - and add two decision variables to carry out the duration constraints - e_{ij} and a_{ij}. A third variable k is also introduced allowing for the minimization of the number of vehicles or vehicle routes (which are equivalent under this context). Therefore, the decision variables in this formulation are:

• x_{ij}, a binary variable that represents the routing solution:
$= 1$, if site j is visited immediately after site i; 0, otherwise;

• y_{ij}, a flow variable that represents the load in the vehicle route when edge (i, j) is crossed. The flow y_{ji} represents the empty space on vehicle route when edge (i, j) is crossed; therefore $y_{ij} + y_{ji} = Q$, at any edge (i, j);

• e_{ij}, a continuous variable representing the exit time from site i to site j ;

• a_{ij}, a continuous variable representing the arrival time to site j from site i ;

• k, an integer variable representing the number of vehicles needed.

All routes start at one of the real depots (set V_d) and end at one of the copy depots (set V_f) (see Figure 1).

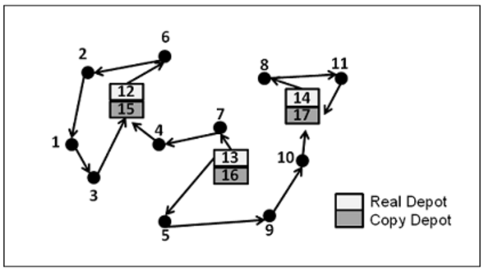


Figure 1: Routes illustration for the MDVRP-OR.

The real-life problem, with more than 100 collection sites, was solved through the CPLEX branch-and-cut algorithm, showing a 2% gap after eight hours of computational time. The proposed routing planning solution resulted in an 11% savings in total cost when compared to the company current solution. As CPLEX branch-and-cut algorithm managed to solve the MILP model applied to the real problem within a reasonable computational time for tactical decisions (eight hours), no other solution method was needed.

4 The Multi-Depot Vehicle Routing Problem with Inter-Depot Routes

The Multi-Depot Vehicle Routing Problem with Inter-Depot Routes (MDVRPI) allows routes between two different depots but imposes vehicles to return to the origin depot within a working day. A vehicle can perform multiple routes per day being those routes either closed and/or inter-depot routes. The set of routes performed by the same vehicle is called rotation (see Figure 2).

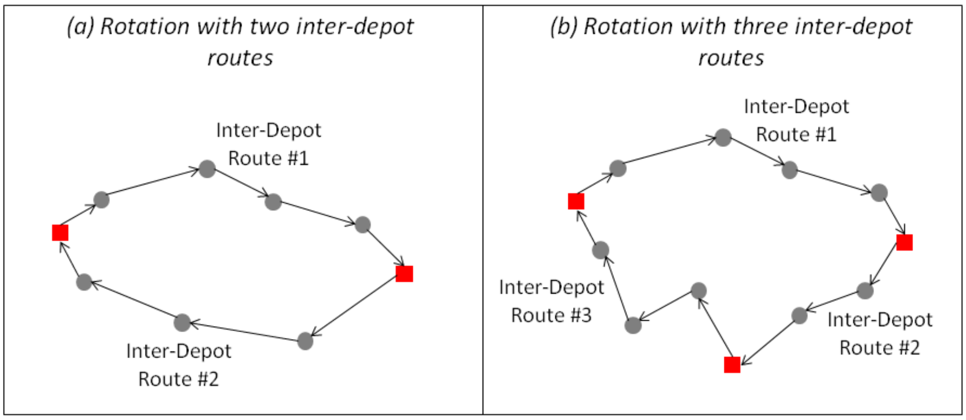


Figure 2: Examples of rotations with (a) two and (b) three inter-depot routes.

The rotation concept brought extra complexity to the problem, since the model has to define feasible rotations made by closed and/or inter-depot routes, assuring that the maximum time limit for a working day is not exceeded. A new mathematical formulation is developed where both rotations and routes are defined. The solver CPLEX was not capable of solving the model applied to literature instances of large dimension and a solution methodology was developed. This solution methodology is based on two relaxations of the original problem: the duration constraints and the number of vehicles available (see Figure 3).

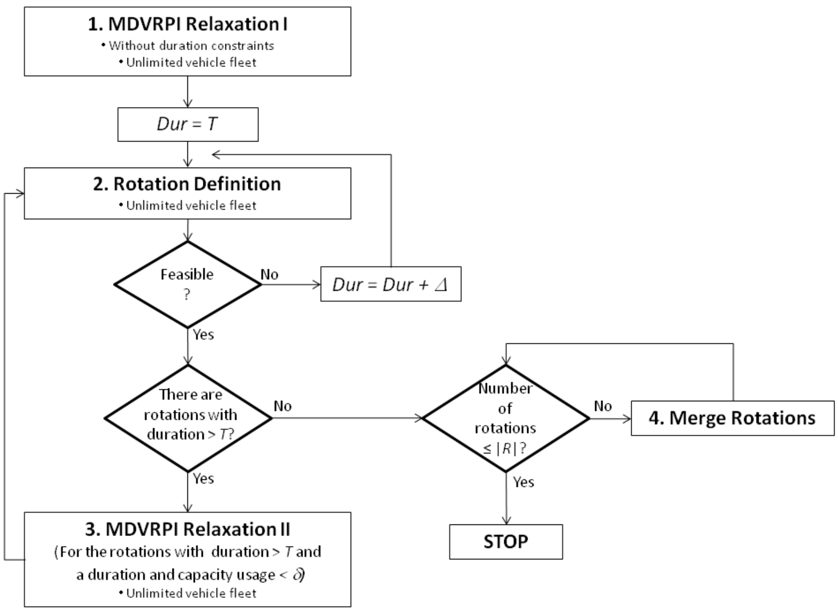


Figure 3: Solution methodology flowchart for the MDVRPI.

Half of the literature instances were then solved by the first module of the solution methodology even though the duration constraints and the number of vehicles available were not being considered. The solutions founded for all literature instances were compared with the work of Crevier et al. (2007) where the MDVRPI is tackled but assuming that all vehicles are based in a central depot rather than in multiples depots (what renders Crevier's et al. work to a Vehicle Routing Problem with Intermediate Replenishment Facilities). Economical savings in all instances up to 9.7% are attained if the MDVRPI with the vehicles based in multiple depots is solved.

5 The Multi-Product, Multi-Depot Vehicle Routing Problem with Economic and Environmental Concerns

In the classical MDVRP the objective function is distance, time or cost minimization. Here a different objective function has been analysed reflecting an environmental concern - minimization of the CO_2 emissions. A recyclable packaging waste collection system is studied where multiple products have to be collected, transported to the depots (inbound transportation) and then to a sorting station (outbound transportation). Such system manages their operations under a municipality-perspective: service areas and collection routes were defined respecting the municipalities' boundaries.

When multiple products are at stake, two alternative solutions can be created regarding service areas. The recyclable materials at each collection site: a) have to be collected from the same depot (service areas by depot); b) can be collected from different depots (service areas by recyclable materials). These two alternatives are studied and the results are compared. The Multi-Product, Multi-Depot Vehicle Routing Problem (MP-MDVRP) is formulated by a MILP model and a decomposition solution approach (see Figure 4) is developed to solve the real case.

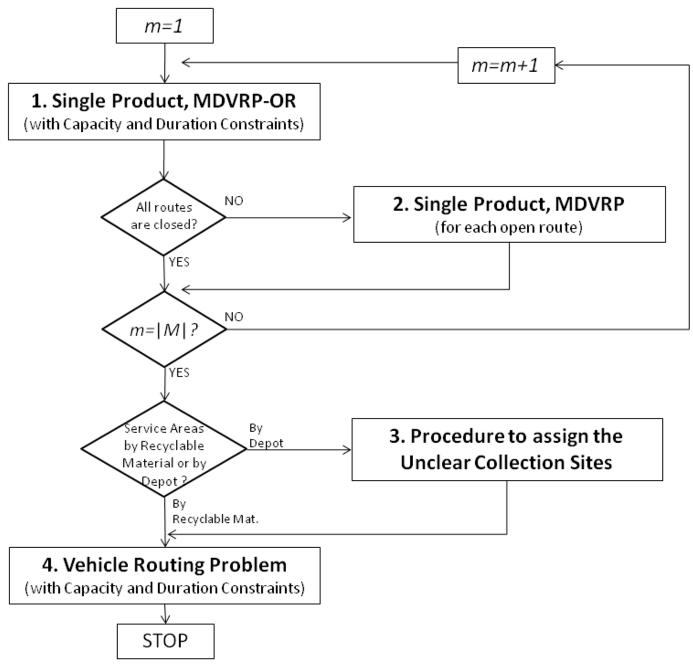


Figure 4: Solution methodology flowchart for the MP-MDVRP.

The distance and energy minimization are studied as objective functions. The latter depends on the distance, load and speed, as well as vehicle and road characteristics. The CO_2 emissions are function of the energy requirements. Effectiveness tests are performed where the optimal solution and computational times obtained by CPLEX are compared with the solution provided by the proposed decomposition method. It was concluded that the decomposition method finds a good or even the optimal solution in much less time than CPLEX alone. The maximum deviation between the optimal solution and the one proposed by the decomposition method was of 2%, inducing that an effective solution method was developed.

The decomposition method was applied to the real case study, where three service areas configurations were analysed and two objective functions compared. If the current service areas are maintained and only the vehicle routes optimized, annual savings of 13% in both cost and CO_2 emissions are observed. If service areas are restructured by depot, annual savings of 19% in cost and 24% in CO_2 emissions would be obtained. Lastly, if service areas are restructured by recyclable material, annual savings of 22% in cost and 27% in CO_2 emissions can be reached. Comparing the two objective functions (minimize distance *versus* energy), different solutions are obtained. The energy-minimizing objective produces service areas with more sites assigned to the sorting station as it minimizes the outbound transportation where vehicles transport heavier loads. Moreover, route topology also changes under the energy-minimization objective since this formulation seeks solutions where vehicles travel less distance with heavy load, so either the heaviest collection sites are placed latter in the route or more routes with lesser load are defined. However, the final results regarding total distance travelled and CO_2 emissions for the real-case study did not differ substantially. These results allow us to conclude that the distance travelled has a major contribution for CO_2 emissions. Therefore, when distance is minimized (a proxy for the economic goal), it is simultaneously contributing to mitigate the negative environmental impact of transportation.

6 The Multi-Product, Multi-Depot Periodic Vehicle Routing Problem with Inter-Depot Routes with Economic, Environmental and Social Concerns

To accomplish the main objective of this work, that is, to contribute to an increase in efficiency, diminish environmental impact and increase social concerns in reverse logistics networks, namely, in waste collection systems, the social objective is here tackled. A multi-objective approach was used to devise a solution where costs are balanced with environmental and social concerns. As it is involved the definition and scheduling of vehicle routes in a multiple depot system, where inter-depot routes are allowed, the problem is modelled as a Multi-Depot Periodic Vehicle Routing Problem with Inter-Depot Routes (MDPVRPI). This problem consists of simultaneously selecting a set of visit days for each client, defining the service areas of each depot and establishing vehicle routes for each day of the planning horizon.

The planning of a sustainable logistics system, where the three dimensions of sustainability are taken into account, is seldom studied in the literature and, therefore, this work aims at contributing to fulfil this gap. The social dimension is the less studied of the three objectives and consequently almost no metrics are proposed to deal with this aspect in the literature. In this work, this dimension is assumed linked to the promotion of equity among the human resources (in this case, the drivers) and the objective function regarding the social issue considers the minimization of the maximum number of working hours among the drivers.

Since the goal is to obtain a solution where costs are balanced with environmental and social concerns, the problem will be solved through a decomposition approach where two steps are defined (see Figure 5).

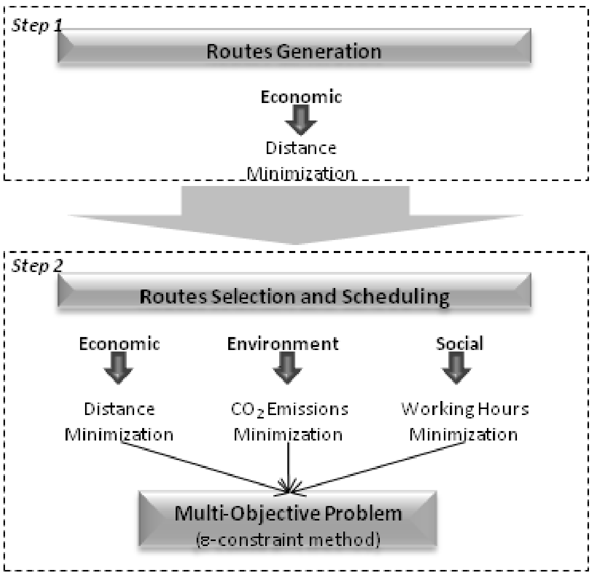


Figure 5: Solution approach overview for the MP-MDPVRPI.

In a first step, a set of feasible routes is generated considering the economic objective (using the MILP models described previously). Then, in a second step, when selecting and scheduling the routes, the three objectives are first considered individually by assessing the following metrics: total travelled distance as the economic objective (note that distance is a linear function of the variable costs); CO_2 emissions as the environmental objective; and number of working hours of each driver as the social objective. Lastly, a multi-objective approach based on the epsilon-constraint method is applied to define an approximation to the Pareto's frontier, which can be used by the decision-maker to select the solution to be adopted.

When economic and environmental objectives are minimized, unbalanced solutions are obtained in terms of working hours by driver. On the opposite side, when the social objective is minimized, a balanced solution is obtained where all drivers have the same number of driving hours (see Figure 6). However, this equity solution leads to a significant increase in distance and CO_2 emissions (see Figure 7).

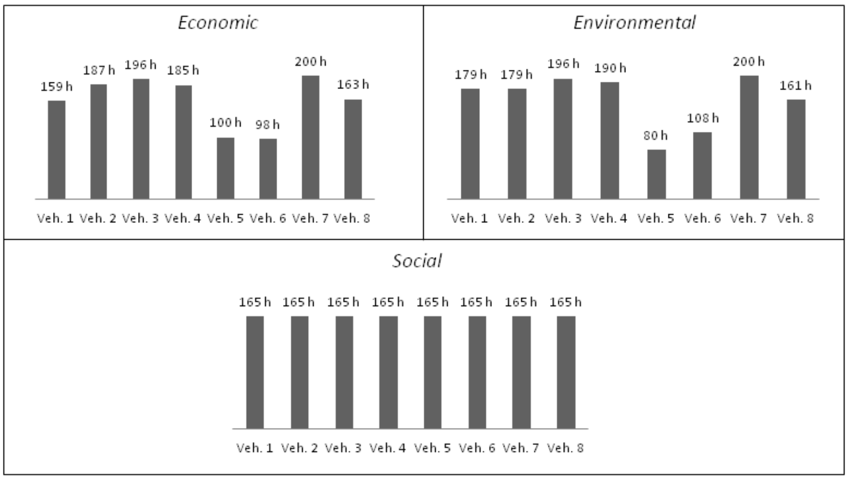


Figure 6: Number of working hours by vehicle in the three solutions.

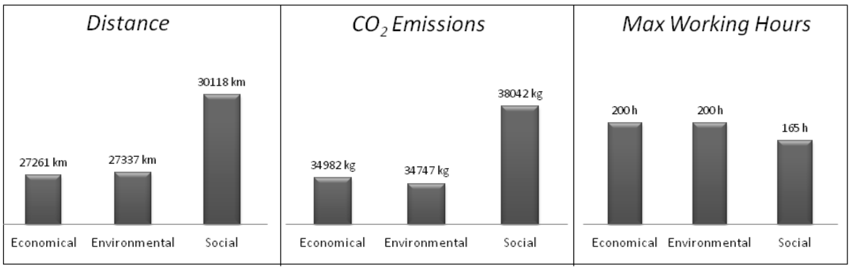


Figure 7: Distance, CO_2 emissions and maximum working hours for the three objective functions.

On the other hand, when accounting for the three objective simultaneously, an efficient solution is identified (a compromise solution, see Figure 8) where the distance to the ideal point is minimized. The Tchebycheff norm is used as distance measure. The ideal point (z^I) is defined according to the individual minima of each objective, in this case $z^I = (27\ 261$ km, $34\ 747$ kg CO_2, 165 h$)$ and the compromise solution z^C obtained is $z^C = (28\ 013$ km, $35\ 653$ kg CO_2, 174 h$)$.

As main conclusion it can be stated that an innovative approach to planning sustainable logistics systems is proposed. Tactical and operational decisions are coped and different solutions are obtained when each dimension of sustainability is addressed individually. The main contribution of this work is to integrate within a single solution the three dimensions of sustainability where innovative aspects that should be considered when planning reverse logistic systems are modelled such as service areas, routes definition as well as routes scheduling, CO_2 emissions and human resources working hours.

7 Final remarks and further work

Generic models for the tactical-operational decisions levels of reverse logistics systems with multiple depots have been developed throughout this work, which were validated by literature instances and by real case studies. As main goal, the objectives have been to increase efficiency, diminish environmental impact and increase social benefit. The economic objective function has been widely studied but the environmental and social ones have been less addressed. Some challenges were faced when dealing, for the first time, with the latter objectives. To cope with the environmental concern, a comprehensive study

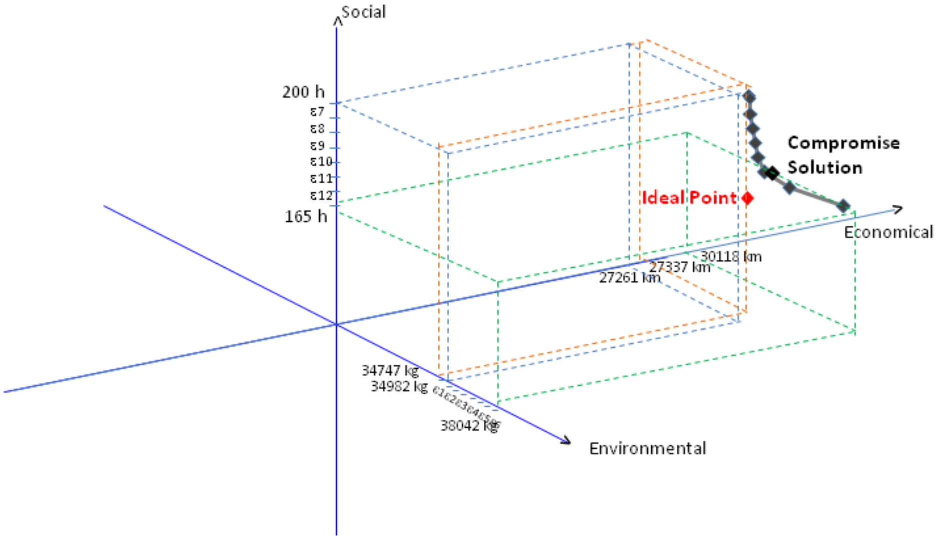


Figure 8: Approximation to Pareto's frontier considering the three objectives simultaneously, with the compromise solution and ideal point highlighted.

on vehicle emissions models was rolled out, where some mechanics and physics insights had to be studied. Regarding the social objective, the difficulty was on how to translate the social concern into metrics and decision variables.

Besides the novelty of the objective functions addressed, some key characteristics that have not being often studied in the literature were also tackled in this research, as the open routes between two depots, the inter-depot routes and the rotation concept, considering simultaneously the inbound and outbound transportation flows, and the existence of multiple products. Despite these characteristics are commonly present in real world cases, they have not been tackled profoundly by the academia.

Based on the developed work, future research topics are foreseen to generalise the problem addressed. An alternative to the single-material routes could be studied, where vehicles with compartments are used to collect two or even the three materials simultaneously without mixing them (the Multi-Compartment Vehicle Routing Problem). Also stochastic models could be developed in order to cope with more realistic scenarios regarding the quantities to be collected. Following the tactical and operational decisions tackled in this work, a future research topic could also be the inclusion of strategic decisions into the models developed, where the number and location of the depots and sorting stations are decided along with the service areas and vehicle routes. Furthermore, companies are now introducing greener vehicles (like hybrid or electric vehicles) to their fleet in order to minimize the CO_2 emissions in the transportation activity. It will be interesting to study the impact of introducing some electric vehicles in the vehicle fleet of recyclable waste collection systems and analyse which routes are assigned to such vehicles and which are assigned to diesel vehicles, by assessing the final impact in terms of costs and CO_2 emissions.

References

[Baldacci et al., 2004] Baldacci, R., Hadjiconstantinou, E., Mingozzi, A. (2004). An exact algorithm for the capacitated vehicle routing problem based on a two-commodity network flow formulation. *Operations Research*, 52, 723-738.

[Baldacci and Mingozzi, 2009] Baldacci, R. and Mingozzi, A. (2009). A unified exact method for solving different classes of vehicle routing problems. *Mathematical Programming*, 120, 347-380.

[Bektas and Elmastas, 2007] Bektas, T., Elmastas, S. (2007). Solving school bus routing problems through integer programming. *Journal of the Operational Research Society*, 58, 1599-1604.

[Bektas and Laporte, 2011] Bektas, T. and Laporte, G. (2011). The Pollution-Routing Problem. *Transportation Research Part B-Methodological*, 45, 1232-1250.

[Brandao, 2004] Brandao, J., (2004). A tabu search algorithm for the open vehicle routing problem. *European Journal of Operational Research*, 157, 552-564.

[Chao et al.,1993] Chao, I., Golden, B.L., Wasil, E. (1993). A new heuristic for the multi-depot vehicle routing problem taht improves upon best-known solutions. *American Journal of Mathematical and Management Sciences*, 13, 371-406.

[Cordeau et al., 1997] Cordeau, J.F., Gendreau, M., Laporte, G. (1997). A tabu search heuristic for periodic and multi-depot vehicle routing problems. *Networks*, 30, 105-119.

[Crevier et al., 2007] Crevier, B., Cordeau, J.F., Laporte, G. (2007). The multi-depot vehicle routing problem with inter-depot routes. *European Journal of Operational Research*, 176, 756-773.

[De Brito and Dekker, 2004] De Brito, M.P. and Dekker, R., A Framework for Reverse Logistics. *Reverse Logistics - Quantitative Models for Closed-Loop Supply Chain* (eds. R. Dekker et al.), Springer-Verlag, Berlin, 2004.

[Dondo and Cerda, 2009] Dondo, R.G.D.R.G., Cerda, J. (2009). A hybrid local improvement algorithm for large-scale multi-depot vehicle routing problems with time windows. *Computers and Chemical Engineering*, 33, 513-530.

[Erdogan and Miller-Hooks, 2012] Erdogan, S., Miller-Hooks, E. (2012). A Green Vehicle Routing Problem.. *Transportation Research Part E-Logistics and Transportation Review*, 48, 100-114.

[Fu et al., 2005] Fu, Z., Eglese, R., Li, L.Y.O. (2005). A new tabu search heuristic for the open vehicle routing problem. *Journal of the Operational Research Society*,56, 267-274.

[Golden et al., 1977] Golden, B.L., Magnanti, T.L., Nguyen, H.Q. (1977). Implementing Vehicle Routing Algorithms. *Networks*, 7, 113-148.

[Hadjiconstantinou and Baldacci, 1998] Hadjiconstantinou, E., Baldacci, R. (1998). A multi-depot period vehicle routing problem arising in the utilities sector. *Journal of the Operational Research Society*, 49, 1239-1248.

[Ho et al., 2008] Ho, W., Ho, G.T.S., Ji, P., Lau, H.C.W. (2008). A hybrid genetic algorithm for the multi-depot vehicle routing problem. *Engineering Applications of Artificial Intelligence*, 21, 548-557.

[Laporte, 2007] Laporte, G. (2007). What you should know about the vehicle routing problem. *Naval Research Logistics*, 54, 811-819.

[Laporte et al., 1984] Laporte, G., Nobert, Y., Arpin, D. (1984). Optimal solutions to capacitated multi-depot vehicle routing problems. *Congressus Numerantium*, 44, 283-292.

[Laporte et al., 1988] Laporte, G., Nobert, Y., Taillefer, S. (1988). Solving a Family of Multi-Depot Vehicle-Routing and Location-Routing Problems. *Transportation Science*, 22, 161-172.

[Letchford et al., 2007] Letchford, A.N., Lysgaard, J., Eglese, R.W. (2007). A branch-and-cut algorithm for the capacitated open vehicle routing problem. *Journal of the Operational Research Society*, 58, 1642-1651.

[Lim and Wang, 2005] Lim, A. and Wang, F. (2005). Multi-depot vehicle routing problem: A one-stage approach. *Ieee Transactions on Automation Science and Engineering*, 2, 397-402.

[Parthanadee and Logendran, 2006] Parthanadee, P. and Logendran, R. (2006). Periodic product distribution from multi-depots under limited supplies. *Iie Transactions*, 38, 1009-1026.

[Pisinger and Ropke, 2007] Pisinger, D. and Ropke, S. (2007). A general heuristic for vehicle routing problems. *Computers and Operations Research*, 34, 2403-2435.

[Renaud et al., 1996] Renaud, J., Laporte, G., Boctor, F.F. (1996). A tabu search heuristic for the multi-depot vehicle routing problem. *Computers and Operations Research*, 23, 229-235.

[Repoussis et al., 2007] Repoussis, P.P., Tarantilis, C.D., Ioannou, G. (2007). The open vehicle routing problem with time windows. *Journal of the Operational Research Society*, 58, 355-367.

[Salari et al., 2010] Salari, M., Toth, P., Tramontani, A. (2010). An ILP improvement procedure for the Open Vehicle Routing Problem. *Computers and Operations Research*, 37, 2106-2120.

[Salhi and Sari, 1997] Salhi, S. and Sari, M. (1997). A multi-level composite heuristic for the multi-depot vehicle fleet mix problem. *European Journal of Operational Research*, 103, 95-112.

[Sariklis and Powell, 2000] Sariklis, D. and Powell, S. (2000). A heuristic method for the open vehicle routing problem. *Journal of the Operational Research Society*, 51, 564-573.

[Schrage, 1981] Schrage, L. (1981). Formulation and structure of more complex-realistic routing and scheduling problems. *Networks*, 11, 229-232.

[Tarantilis and Kiranoudis, 2002] Tarantilis, C.D. and Kiranoudis, C.T. (2002). Distribution of fresh meat. *Journal of Food Engineering*, 51, 85-91.

[Tillman and Cain, 1972] Tillman, F.A. and Cain, T.M. (1972). Upperbound Algorithm for Single and Multiple Terminal Delivery Problem. *Management Science Series a-Theory*, 18, 664-682.

[Vidal et al., 2012] Vidal, T., Crainic, T.G., Gendreau, M., Lahrichi, N., Rei, W. (2012). A Hybrid Genetic Algorithm for Multidepot and Periodic Vehicle Routing Problem. *Operations Research*, 60, 611-624..

Downstream oil products distribution planning

Nuno Mota*, Susana Relvas*‡, Jorge Gonçalves†

* *Centro de Estudos de Gestão, Instituto Superior Técnico, Universidade Técnica de Lisboa, 1049-001 Lisboa*

† *PETROGAL SA, Rua Tomás da Fonseca, Torre C, Lisboa, Portugal*

‡ *corresponding author: susana.relvas@ist.utl.pt*

Abstract

In the actual competitive business environment, the oil industry requires permanent efficiency improvements, while remaining flexible to face any contingency. At the distribution level, the proper sizing and scheduling of a tanker fleet is a complex problem. In this paper, the T2S.opt decision support tool is presented. T2S.opt addresses the fleet distribution planning problem under normal and abnormal operational scenarios. The optimal planning covers short-term solutions and minimizes operational costs. A Mixed-Integer Linear Programming (MILP) was developed and implemented in a proper user interface. The software was used to schedule the secondary distribution of oil products of GalpEnergia in Portugal.

Keywords: Oil Supply Chain, Distribution, MILP, Scheduling, Decision Support System.

1 Introduction

In the actual worldwide environment, the oil industry faces fierce and growing competition. In this context, oil supply chains should be analyzed in order to improve their efficiency, while remaining flexible to successfully handle certain types of contingencies, such as lack of products or fleet. At the downstream secondary distribution level, tankers are commonly used, due to their flexibility to transport the derivative products from distribution centers to service stations [MirHassani, 2008]. This operation is usually short-term scaled so as to meet final consumers' demand. However, the availability and proper sizing of a fleet to perform the required distribution can be a complex problem due to the fact that demand is known on short notice and the distribution network may include hundreds of demand points to be satisfied.

The optimal fleet allocation solution would require a Vehicle Routing Problem (VRP), or one of its variants such as the Capacitated VRP, to be solved in short periods of time over a geographically large area with several demand locations (service stations country wide or region wide). This problem is known to be NP hard and even instances of a small size are difficult to solve. On the other hand, real word tools used to this end usually are based on geographic knowledge and experience of schedulers to perform such allocation. The gap between an optimal distribution planning tool and what is currently used in companies leaves the opportunity for the academic community to focus on feasible and good solutions in short scheduling periods. Moreover, decision support tools using such type of solutions are of great support when dealing with different types of operational contingencies.

In this paper we present a decision support system (DSS), named Tank-to-Station Optimizer (T2S.Opt), which addresses the fleet distribution planning problem under normal and abnormal operational scenarios. The system in study considers a network of oil products distribution centres that supply service stations on a daily basis. Distribution centres are supplied from refineries. The DSS is built so as to consider the distribution operation over a short-term time horizon (one to few days) minimizing operational costs. This DSS uses a mixed integer linear programming model (MILP) solved with a free solver (GLPK). The DSS includes a user interface and a flexible architecture that can be fully customized, and uses the MILP-based solution strategy reported in [Mota, 2012]. The proposed methodology is tested in a real world case study of a Portuguese company - GalpEnergia - which distributes oil products nationwide.

In the following section a literature review in the field of oil supply chains is presented. Section three describes the proposed mathematical model and a brief overview of T2S.opt. Section four presents the case study as well as the results obtained by using T2S.opt. Finally, section five encloses the conclusions and future work.

2 Literature Review

Planning might be different according to several aspects as performance measures and decision variables change. According to An et al. [2011], five planning levels may be distinguished: strategic, tactical, operational, integration of tactical and operational, integration of strategic and tactical. Gayialis and Tatsiopoulos [2004], among other authors, chooses to distinguish only the first three levels of planning mentioned above. Strategic planning uses aggregated information and regards long term decisions as investments and dimensioning of the logistics network. This may include locating and determining the capacity of the depots, physical flows, sourcing strategy, among other decisions [Gayialis and Tatsiopoulos, 2004]. The objective functions are normally related to costs and flexibility of the network that is being modified or created (see [Mota, 2012]) Tactical planning regards medium-term decisions as inventory levels, production planning and distribution related to an existing network. The objective functions normally take into consideration the costs (or profit) and the responsiveness to the customer. Operational planning considers detailed information and short-term decisions, such as daily repetitive operations, and concentrates in optimizing specific points of the network [Gayialis and Tatsiopoulos, 2004]. It may include transportation scheduling, detailed resource allocation, production decisions or distribution and routing problems as the vehicle routing problems and its variants (see [Mirabi et al., 2010] and [Nussbaum et al., 1996]). Typically, these types of problems are NP hard to solve and heuristics are proposed to solve them (see [Mirabi et al., 2010]). The integration of the strategic and tactical models typically handles tactical decision variables, such as production and/or distribution scheduling, levels of inventory, and strategic decisions, like physical allocation, sizing infrastructures and/or transports, among others [Mota, 2012]. These models focus more in the downstream operations [An et al., 2011]. The integration of operational and tactical models, when compared with operational models, tend to include more than one operation and are more concentrated in the production activity (see [Timpe and Kallrath, 2000] and [Al-Othman et al., 2008]). There are several approaches that can be used when dealing with uncertainty, risk and subsequent reactive planning. Uncertainty is related to the non-deterministic character of the variables under scrutiny. Although it is generally consensual that controlling uncertainty and risk might bring considerable benefits, there are few approaches applied to the oil supply chain. Adhitya et al. [2007] propose a framework based on a rescheduling heuristic strategy to manage failures in a refinery. However, the oil supply chain may face innumerous uncertainty factors or abnormal scenarios. To this end, decisions at different levels may be implied when finding solutions for these situations.

2.1 Literature Review Discussion

The oil supply chain is a complex system where decisions at different levels might influence the efficiency of operations. Therefore, when dealing with secondary distribution level at the downstream oil supply chain, we are covering several customer locations sourced from different refineries or distribution centres. Furthermore, under non-regular scenarios (e.g. lack of stock of a given product) tactical decisions related with allocation of customers to sourcing locations may arise at the operational level. Therefore, our aim is to develop a problem representation where the tactical level is integrated with the operational level of decision. One should consider the supply of the depots via refinery and integrate it with the distribution at a macro level. Frequent disruptions related with this oil supply chain operation are broadly defined as lack of product or lack of transportation capacity. In order to study this problem, other than stochastic models may be applied, where the deterministic nature of the problem is represent, acting on the data to represent different planning situations.

3 Problem Formulation

The problem in study consists of a distribution network with multiple depots and refineries. Each refinery supplies several depots with the respective products. The number, location and product availability of the depots are known. A fleet of trucks with different capacities and limited availability per truck type is used to transport the products from depots to customers (*e.g.* municipalities). In this way, either single customers or aggregated customers may be dealt within the proposed formulation. The time required to travel between network points (from depots to customers) is also deterministic and known. Each truck cannot visit multiple regions in one trip.

Given this setting, we aim to determine the optimal distribution planning of refined products from depots to customers in short-term time horizons that minimizes distribution costs, which include transportation costs. Penalties for unmet demand are also accounted for.

3.1 Mathematical Model Formulation

The proposed solution is divided in two phases. In the first phase we aim designing the delivery routes from a depot to customers at a minimum cost. In case it is not possible to meet all the demand, due to fleet or/and product constraints, a second phase returns the minimum time that it would be required to supply that demand. Basically, the second phase problem has no fleet limitations. The proposed formulation uses the following notation consisting of indexes, sets, parameters and variables.

Indexes:

$i-$ Depots;
$j-$ Customers;
$k-$ Products;
$p-$ Tankers;
$t-$ Time periods;

Sets:

$i \in I$ – Set of all depots;
$j \in J$ – Set of all customers;
$k \in K$ – Set of products;
$p \in P$ – Set of tankers;
$t \in T$ – Set of time periods;
$R_{k,p}-$ Set that relates products with tanker to transport it;
$S_{k,p,i}-$ Set that relates products with tankers and depots.

Parameters:

$A_{i,k,t}-$ Initial inventory of product k, at depot i, at time t;
$C_{i,j}$ – Unitary distribution cost between depot i and customer j;
$Cam_{i,t,p}-$ Available shifts from tanker p, in depot i, at time period t;
$Dem_{j,k,t}$ – Demand of customer j of product k at time period t;
$IF^{min}_{i,k,t}-$ Minimum inventory of product k, in depot i, at time period t;
$IF^{max}_{i,k,t}-$ Maximum inventory of product k, in depot i, at time period t;
$Min_{i,j}$ – Travelling time between depot i and customer j;
M – Big-M value when compared with variable $Y_{i,j,k,t}$;
Pen – Penalty per unit of volume not transported between depot i and customer j;
$TP_{i,t}$ – Maximum flow from depot i at time period t;
$TR_{i,k,t}$ – Refinery flow to depot i, from product k, at time period t;
V_p – Transportation capacity of tanker p.

Positive Decision Variables:

$X_{i,j,k,t}$ – Flow of product k transported from depot i to customer j at time period t;
$Xn_{j,k,t}$ – Demand of product k from customer j not satisfied at time period t;
$IF_{i,k,t}$ – Final inventory of product k, in depot i at time period t;
$Y_{i,j,k,t}$ – Demand of product k from customer j satisfied in the second phase and allocated to depot i at time peri
$Yn_{j,k,t}$ – Demand of product k from customer j not satisfied in the second phase at time period t.

Integer Decision Variables:

$n_{i,j,t,p}$ – Number of vehicles from depot i to customer j, from tanker p, at time period t.

Mathematical Model - 1st Phase:

Objective Function:

$$MinZ = \sum_t \sum_k \sum_j (\sum_i (C_{i,j} X_{i,j,k,t})) + Pen Xn_{j,k,t}) \tag{3.1}$$

$$Pen > MaxC_{i,j} \tag{3.2}$$

Subject to:

$$Xn_{j,k,t} = Dem_{j,k,t} - \sum_i X_{i,j,k,t} \quad \forall j,k,t \tag{3.3}$$

$$\sum_{j}\sum_{k} X_{i,j,k,t} \leq TP_{i,t} \quad \forall i,t \tag{3.4}$$

$$A_{i,k,t} + TR_{i,k,t} - \sum_{j} X_{i,j,k,t} = IF_{i,k,t} \quad t = 1; \forall i,k \tag{3.5}$$

$$IF_{i,k,t-1} + TR_{i,k,t} - \sum_{j} X_{i,j,k,t} = IF_{i,k,t} \quad t > 1; \forall i,k \tag{3.6}$$

$$\frac{\sum_{k} X_{i,j,k,t}}{V_p} = n_{i,j,t,p} \quad \forall (k,p) \in R_{k,p}, j, t, i \tag{3.7}$$

$$\sum_{i}\sum_{j}(n_{i,j,t,p} Min_{i,j}) \leq Cam_{i,t,p} \quad \forall (k,p,i) \in S_{k,p,i}, t \tag{3.8}$$

$$IF_{i,k,t}^{min} \leq IF_{i,k,t} \leq IF_{i,k,t}^{max} \quad \forall i,k,t \tag{3.9}$$

$$IF_{i,k,t}, X_{i,j,k,t}, Xn_{j,k,t} \geq 0, n_{i,j,t,p} \in \mathbb{N} \quad \forall i \in I, j \in J, k \in K, t \in T \tag{3.10}$$

The objective function (3.1) minimizes the transportation costs between depots and customers (first term), as well as the penalties in case of unsatisfied demand (second term). The condition (3.2) states that the penalty cost (which is fictitious) needs to be higher than any distribution costs. Thus, it forces the model to allocate all the available resources before demand is unmet. Eq. (3.3) acts as a soft constraint that determines unmet demand. Eq. (3.4) limits throughput of each depot in each time period. Eq (3.5) defines the final inventory at the beginning of the time horizon and Eq. (3.6) the final inventory for the subsequent time periods. Eq. (3.8) restraints the fleet transportation capacity, given different trucks (or tanks) and depots. Constraints in Eq. (3.9) limit the product inventory per depot and time period. Finally, Eq. (3.10) is a non-negativity constraint for flows and inventory.

Mathematical Model - 2$^{\text{nd}}$ Phase:

Objective Function:

$$MinZ = \sum_{t}\sum_{i}\sum_{j}\sum_{p}(Min_{i,j} n_{i,j,t,p}) \sum_{t}\sum_{k}\sum_{j}(MYn_{j,k,t}) \tag{3.11}$$

The second phase returns as a solution what are the fleet requirements to meet unfulfilled demand of the first phase. In this case, the objective function minimizes the required time to fulfill the new demand (first term), by allocating the unmet demand of the 1$^{\text{st}}$ phase to new depots, as well as it minimizes the penalties when such allocation is not possible (second term).

Subject to:

$$Yn_{j,k,t} = \sum_{i} Y_{i,j,k,t} - Xn_{j,k,t} \quad \forall j,k,t \tag{3.12}$$

$$IF_{i,k,t} \geq \sum_{j} Y_{i,j,k,t} \quad \forall i,k,t \tag{3.13}$$

$$\sum_{j}\sum_{k} Y_{i,j,k,t} \leq TP_{i,t} \quad \forall i,t \tag{3.14}$$

$$\frac{\sum_{k} Y_{i,j,k,t}}{V_p} = n_{i,j,t,p} \quad \forall (k,p) \in R_{k,p}, j, t, i \tag{3.15}$$

$$Yn_{j,k,t}, Y_{i,j,k,t} \geq 0, n_{i,j,t,p} \in \mathbb{N} \quad \forall i \in I, j \in J, k \in K, t \in T \tag{3.16}$$

Eq. (3.12) allocates the unmet demand $Xn_{j,k,t}$, calculated in the 1$^{\text{st}}$ phase, to the nearest possible depot. Constraint in Eq. (3.13) ensures that there is sufficient inventory in the depots allocated in Eq. (3.12) and Eq. (3.14) limits the throughput of each depot at time t. Eq. (3.15) restricts the flow of products to the exact capacity of each tank. Finally, Eq. (3.16) is the domain definition constraint.

4 Case study and Results

4.1 Case study summary

The object of this study consists of the downstream fuel distribution planning of four refined petroleum products: gasoline, diesel, heating oil (HO) and Jet Fuel (JF). This activity is performed by Petrogal, a subsidiary company of the Group GalpEnergia, from this point onwards referred as Galp. The distribution activity includes six depots, as one can see in the Table 1 with different capacities and products being stored, which satisfy a total of roughly 270 Portuguese municipalities. The products considered are gasoline, diesel, heating oil (HO) and Jet Fuel (JF). Figure 1 illustrates depots' locations, as well as crude oil flow. The numbers presented in Figure 1 illustrate the actual distribution regions covered by outsourced carriers.

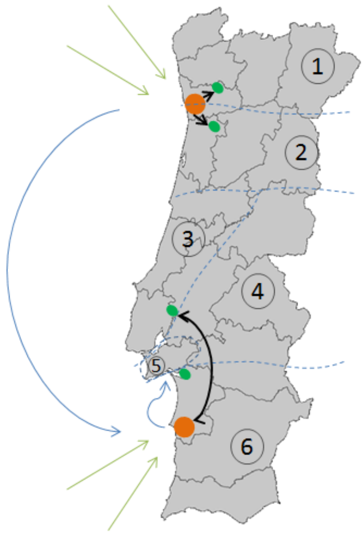


Figure 1: Galp's simplified national supply chain

Table 1: Average product availability by depot (m^3).

Depots	Gasoline	Diesel	HO	Max. JF
Boa Nova	10000	10000	-	700
Real	-	400	10000	-
CLC	10000	10000	-	10000
Mitrena	800	800	10000	-
Sines	10000	10000	-	1225
Aveiro	-	-	-	-

Table 2: Daily Trucks and shifts available by depot.

	Trucks	Shifts	JF Trucks	JF Shifts
Boa N. + Real	51	78	*(pipeline)*	-
CLC Aveiras	40	49	9	18
Mitrena	6	9	-	-
Sines	20	34	*(railway)*	-
Aveiro	-	-	-	-

The data used to test the proposed approach is referent to 2011.

Within the six depots, Aveiro is not currently in use. In the case of Jet Fuel, Boa Nova stores typically around 700 m^3, Sines roughly 1225 m^3 and CLC has practically no fuel limitation. The total daily national consumption of gasoline, diesel and HO in 2011 was mainly situated in the ranges of 0 to 6000 m^3 and 8000 to 9000 m^3. The fuel transportation is outsourced to five carriers. These carriers are allocated to depots, and each carrier manages the amount of trucks to perform the transportation service. In the present scenario, Table 2 illustrates how many trucks are available per day in each depot (1 shift equals 8 hours of transportation and some trucks perform 2 shifts per day).

The maximum capacities of each normal truck tank are 32 m^3 and 35 m^3 in the case of JF. Typically, the latter type of truck performs an average of 5 deliveries per day and the former an average of 3 deliveries.

The contingencies to which the distribution activity is exposed can be broadly categorized between absence of product or absence transport capacity. In the case of Jet Fuel there is a clear limitation of the carrying capacity given the scarcity of this type of truck.

4.2 T2S.Opt Decision Support Tool

The proposed mathematical models were implemented in the T2S.opt application. This application was developed to interpret model solutions in an effective and interactive way.

This application possesses a specific architecture, which integrates an Excel and Access file with the proposed open source optimization Solver GLPK. The Excel file is the user interface where several functionalities are available. The usage of this application will be exemplified in the results section, since all figures are generated by T2S.opt.

4.3 Case Study Results

Several contingency scenarios were implemented in the T2S.opt application. The objective of the results presented is to determine the impact of closing any of the existing depots and the subsequent solution of reallocating the available fleet (where the depot was closed). First, we are going to present a normal distribution scenario in order to compare it with the contingencies referred. The closure of distribution centres will be illustrated with the examples of CLC and Sines.

4.3.1 Normal Distribution Scenario

This scenario contemplates an ordinary day of distribution. To fulfill 7048 m^3 of diesel, 2326 m^3 of gasoline and 64 m^3 of HO, the distribution operation consumes a total of 77815 minutes, 47076 Km and a distribution cost of 118890€. In terms of JF, for a daily supply of 65 deliveries (2275m^3) from CLC to the Lisbon airport, it requires 8515 minutes of transportation time and an expense of 32258€ in transportation costs. The areas of influence of the five depots considered are shown in Figure 2.

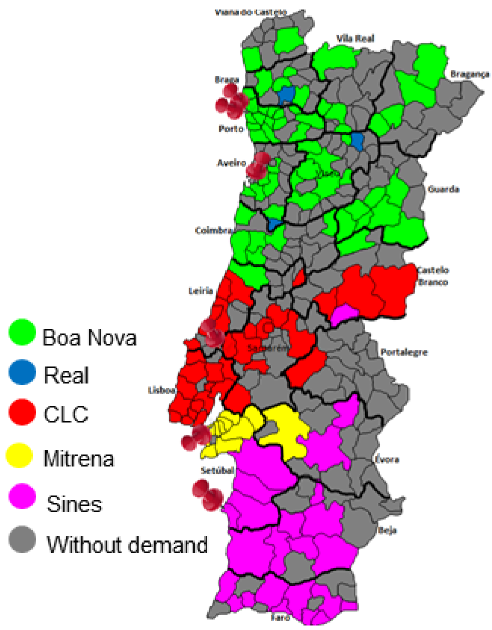


Figure 2: Normal Depot's Influence areas

The illustrated areas are consistent with the current distribution scheduling. Thus, in contingency, these areas are expected to move in the direction of the closed depot, to minimize the unfulfilled demand.

4.3.2 Abnormal Scenario: Closure of CLC

This scenario includes the closure of CLC distribution centre. 71% of the total demand was satisfied, which means that 2784 m^3 were unfulfilled. The regions whose demand was not satisfied are illustrated in Figure 3a (see next figure). In this scenario were consumed 57221 distribution minutes and the total distribution cost was of 76908€. The unitary distribution cost was 11.55€ per m^3 transported, a 7.94% decrease when compared with the normal distribution scenario. However, one can minimize the negative impacts by reallocating the available fleet in CLC to the correct depots. Figure 3b, in the following figure, illustrates the impact of reallocating the non-used fleet. With the solution computed by the second phase of the mathematical model one allocated 5 shifts to Aveiro and 44 to Sines. This measure made the unsatisfied demand decreased 69%, when compared to the scenario where no fleet was reallocated, to a total of 864 m^3. In this scenario the unitary cost of distribution increased 0.31%, from 12.59€/m^3 to 12.63€/m^3.

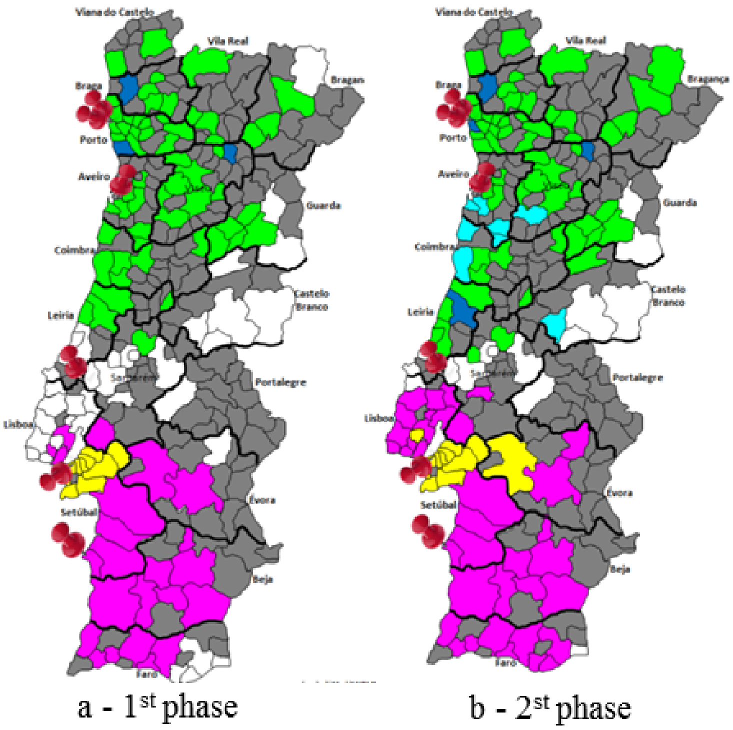


Figure 3: Closure of CLC results

4.3.3 Abnormal Scenario: Closure of Sines

This scenario includes the closure of Sines. 86% of the total demand was satisfied, which means that 1312 m^3 were unfulfilled. Figure 4a shows the unsatisfied municipalities. In this case, 65147 minutes were used for a total distribution cost of 107108€. This gives a unitary cost of 13.18€per m^3 transported, a 4.67% increase when compared with the normal distribution scenario. In the reallocation scenario one considered 7 shifts in the Mitrena and 27 in the CLC. With this measure one decreased the unsatisfied demand to 384 m^3, a 71% drop when compared with the no reallocation scenario. The unitary cost of distribution increased 10.64%, from 12.59€per m^3 to 16.01€per m^3 (when compared with the normal scenario).

4.3.4 Abnormal Scenario: Fleet Unavailability

Correct contingency impact identification will not only allow us to clarify the most critical contingencies, as well as aligning the expectations and goals of the current programming team in the company in study. Typically, the most ordinary contingency is related to the lack of fleet. So we decided to study the impact of an unexpected 25% and 50% fleet unavailability. The demand used in this study was the same of the previous sections. Table 3 summarizes each scenario impact. The first column introduces the scenario and the fleet's percentage drop in the respective depot.

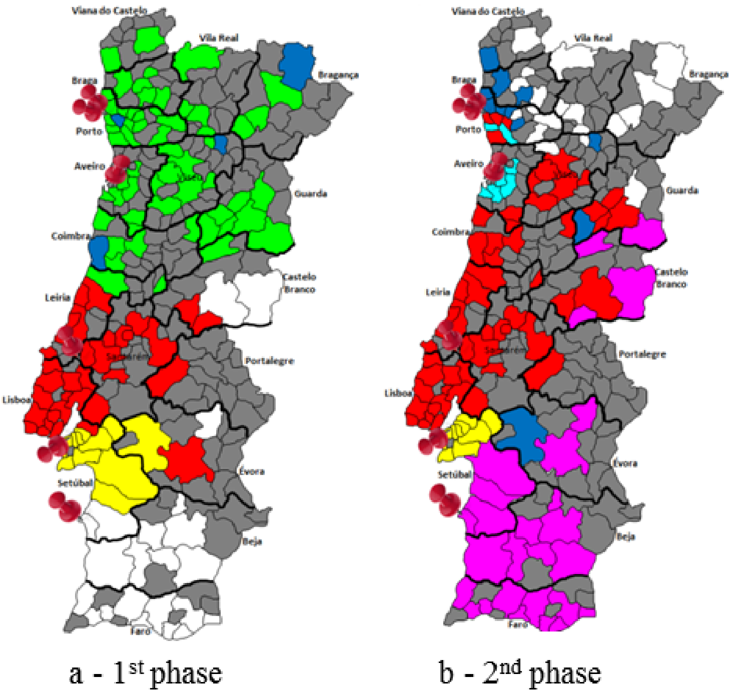


Figure 4: Closure of Sines

Table 3: Impact of 25% and 50% fleet unavailability

Scenarios	Unmet Demand	Shifts needed	U.c. of distribution	U.c. of transport
Normal	0	0	12.59€/m^3	8.90€/m^3
50% Boa N.	960m^3	35	11.82€/m^3	8.01€/m^3
50% CLC	832m^3	26	11.92€/m^3	8.80€/m^3
25% Boa N.	320m^3	14	12.26€/m^3	8.57€/m^3
50% Sines	284m^3	11	12.21€/m^3	8.42€/m^3
25% CLC	256m^3	11	12.27€/m^3	8.89€/m^3
50% Mitre.	96m^3	4	12.60€/m^3	9.11€/m^3
25% Sines	64m^3	3	12.50€/m^3	8.80€/m^3
25% Mitre.	0m^3	0	12.66€/m^3	9.15€/m^3

As one can see, and in line with contingency scenarios previously studied, a drop of 50% of the fleet in Boa Nova and CLC depots has serious consequences on the fulfilled demand. Note also that, for example Boa Nova, in order to supply the 960 m^3 of unfulfilled demand, it would require 35 additional shifts. That is, each delivery takes an average of 1.2 shifts, or approximately 9 hours, to be performed. Another important fact is that the unitary cost of transportation, and distribution, is higher than the normal unitary costs only when there is unavailable fleet in the Mitrena depot. This means that longer distances have to be covered with origin in other depots to fulfill demand otherwise supplied by Mitrena depot.

4.3.5 Fleet Resize and Reallocation

In this section we propose an alternative to the current fleet configuration, by redistributing the available fleet. The goal is to increase savings and mitigate the current operational contingency potential of Galp's operation. For this purpose we are going to use the T2S.opt application. In order to evaluate the best configuration that suits both goals we constructed a sample of 18 daily demands based in 2011's demand. For each, we calculated the optimum fleet allocation to depots, in terms of shifts, using the first phase of the mathematical model. Table 5 summarizes the results obtained, where the column named average (Avg.) represents the fleet configuration proposed as an average of the values obtained within the sample

Comparing the results obtained with the current fleet configuration (5th column of Table 4) there is a positive 10 shift deviation (6th column). However, it was calculated a deficit of 4 shifts in Mitrena.

Table 4: Proposal for shifts resizing and reallocation.

Depot	Min	Max	Avg.	Now	Dev.
Boa Nova + Real	65	81	75	78	+3
CLC	36	45	41	49	+8
Mitrena	12	15	13	9	−4
Sines	25	38	31	34	+3
TOTAL Shifts	138	179	160	170	+10

Table 5: Daily cost comparison of fleet configurations.

Costs	Actual config.	Proposed config
Transport	82854€	83335€
Product	33509€	32645€
Total	116363€	115980€

In terms of distribution costs, the proposed configuration represents an average daily gain of 383€ in relation to the current configuration. The decomposition of the average distribution costs for the current fleet and for the proposed fleet scenario is presented in Table 5. As one can see, both alternatives to the current fleet configuration represent an improvement in the average distribution costs.

4.4 Computational Results

Both models were implemented using the programing language of GUSEK and solved in an Intel Core(TM)2 Duo with 2,67GHz and 4 GB RAM computer, using GLPK solver. The stopping criteria are either the optimal solution determination or reaching the memory limit. Table 6 resumes computational data in two categories, normal distribution scenarios and contingency scenarios, in average terms, for the sake of simplicity.

Table 6: Computational Results

	Scenarios	Computational time (CPU sec.)		Relative gap (%)	
		1^{st} Ph.	2^{nd} Ph.	1^{st} Ph.	2^{nd} Ph.
Normal	Average	758.9	0.1	0.0%	0.0%
	Maximum	2901.5	0.1	0.0%	0.0%
	Minimum	1.3	0.1	0.0%	0.0%
Contingency	Average	12953.9	0.1	1.0%	0.0%
	Maximum	47945.5	0.1	5.1%	0.0%
	Minimum	184.4	0.0	0.0%	0.0%

One can conclude that the computational effort varies significantly with the scenario being tested in the first model phase. However, contingency scenarios, which are more restrictive, tend to consume more time and maintain higher relative gaps. The first phase of the model is characterized by 21360 variables, 6405 are integer variables, and 8631 equations. The second phase is characterized by 21360 variables, 6405 are integer variables, and 8605 equations. This model size is applicable to all scenarios, which consisted in a single day of time horizon (one time period).

5 Conclusions and Future Work

This paper proposes a decision support tool, based in an exact model that aims at minimizing operational costs of the secondary distribution operation. The solution procedure includes a MILP model solved using a free solver - GLPK - and managed through a user interface, the T2S.opt application. The proposed

MILP model allows the integration of tactical and operational decisions, such as the determination of areas of influence per distribution centre (frequently a tactical decision) and the daily distribution planning (operational decision). This model flexibility allows obtaining a solution to support the distribution operation decisions even when abnormal scenarios occur (as stated, lack of product or lack of fleet) as well as for fleet sizing (tactical decision). A current limitation of the proposed mathematical models lies in the fact that a vehicle is only able to perform one delivery per trip. The introduction of a condition, in future developments, that allows freight to make more than one delivery per truck will enable more accurate and real results.

References

[Adhitya et al., 2007] Adhitya, A., Srinivasan, R., Karimi, I. A. (2 de January de 2007). Heuristic Rescheduling of crude oil operations to manage abnormal supply chain events. *Wiley InterScience.*

[Al-Othman et al., 2008] Al-Othman, W. B. E. , Labadidi, H. M. S., Alatiqui, I. M., Al-Shayji, K. (2008). Supply chain optimization of petroleum organization under uncertainty in market demands and prices. *European Journal of Operational Research 189 , 822-840.*

[An et al., 2011] An, H., Wilhelm, W. E., Searcy, S. W. (2011). Biofuel and petroleum-based fuel supply chain research: A literature review. *Biomass and Bioenergy (pp. 3763-3764). Estados Unidos: Elsevier.*

[Gayialis and Tatsiopoulos, 2004] Gayialis, S. P., Tatsiopoulos I. P. (2004). Design of an IT-driven decision support system for vehicle routing and scheduling. *European Journal of Operational Research 152 , 382-398.*

[Mirabi et al., 2010] Mirabi, M., Fatemi, G. S., Jolai, F. (2010). Efficient stochastic hybrid heuristics for the multi-depot vehicle routing problem. *Robotics and Computer Integrated Manufacturing, 564-569.*

[MirHassani, 2008] MirHassani, S. A. (2008). An operational planning model for petroleum products logistics under uncertainty. *Applied Mathematics and Computation 196 , 744-751.*

[Mota, 2012] Mota, N. (2012). Análise de cenários para definição de planos de contingência na distribuição secundária de produtos petrolíferos. *Master Thesis in Industrial Engineering and Management, Instituto Superior Técnico, Oeiras, Portugal (in Portuguese).*

[Nussbaum et al., 1996] Nussbaum, M., Sepulveda, M., Cobian, A., Gaete, J., Parra, E., Cruz, J. (1996). A Fuel Distribution Knowledge-based Decision Support System. *Journal Mgmt Sci. Vol. 25, No. 2, 225-234.*

[Timpe and Kallrath, 2000] Timpe, C. H., Kallrath, J. (2000). Optimal planning in large multi-site production networks. *European Journal of Operational Research 126, 422-435.*

Distribution based artificial fish swarm in continuous global optimization

Ana Maria A.C. Rocha*, M. Fernanda P. Costa†, Edite M.G.P. Fernandes*

* *Algoritmi R&D Centre, University of Minho, Portugal*
E-mail: {arocha;emgpf}@dps.uminho.pt

† *Centre of Mathematics, University of Minho, Portugal*
E-mail: mfc@math.uminho.pt

Abstract

Distribution based artificial fish swarm (DbAFS) is a new heuristic for continuous global optimization. Based on the artificial fish swarm paradigm, the new algorithm generates trial points from the Gaussian distribution, where the mean is the midpoint between the current and the target point and the standard deviation is the difference between those two points. A local search procedure is incorporated into the algorithm aiming to improve the quality of the solutions. The performance of the proposed DbAFS is investigated using a set of small bound constrained optimization problems.

Key words: Global optimization, artificial fish swarm, Gaussian distribution, local search.

1 Introduction

The artificial fish swarm (AFS) algorithm belongs to the class of stochastic population-based methods for solving continuous global optimization problems. The artificial fish is a fictitious entity of a true fish. Its movements are simulations and interpretations of fish behavior [Wang et al., 2006, Jiang et al., 2007, Xiu-xi et al., 2010, Rocha et al., 2011, Rocha and Fernandes, 2011, Neshat et al., 2013]. The environment in which the artificial fish moves, searching for the optimal solution of an optimization problem, is the feasible search space of the problem. In nature, fishes desire to stay close to the swarm, protecting themselves from predators and looking for food, and to avoid collisions within the group. These behaviors inspire mathematical modelers that need to solve efficiently optimization problems.

The main fish swarm behavior are the following:

- *random* behavior - in general, fish swims randomly in water looking for food and other companions;
- *searching* behavior - this is a basic biological behavior since fish tends to the food; when fish discovers a region with more food, by vision or sense, it goes directly and quickly to that region;
- *swarming* behavior - when swimming, fish naturally assembles in groups which is a living habit in order to guarantee the existence of the swarm and avoid dangers;
- *chasing* behavior - when a fish, or a group of fishes, in the swarm discovers food, the others in the neighborhood find the food dangling quickly after it.

The problem to be addressed in this paper is the bound constrained problem:

$$\min_{x \in \Omega} f(x), \tag{1.1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a nonlinear function and $\Omega = \{x \in \mathbb{R}^n : -\infty < l_i \leq x_i \leq u_i < \infty, i = 1, \ldots, n\}$ is the feasible region. The objective function f may be non-smooth and may possess many local minima in the search space Ω, since we do not assume that f is convex. Here, our purpose is to compute a global minimizer, i.e., a point $x^* \in \Omega$ such that $f(x^*) \leq f(x)$ for all $x \in \Omega$. Many derivative-free algorithms and heuristics have been proposed to solve problem (1.1), namely those based on swarm intelligence [Engelbrecht, 2005]. Probably the most well-known and widely used in applications are

- the particle swarm optimization (PSO) algorithms [Kennedy and Eberhart, 1995, Ali and Kaelo, 2008, Coelho et al., 2005, Hao and Hu, 2009, Miranda et al., 2007, Pires et al., 2010, Vaz and Vicente, 2007];
- the bee colony [Bernardino et al., 2013, Karaboga and Basturk, 2007, Karaboga and Akay, 2009a), Karaboga and Akay, 2009b), Diwold et al., 2011]; and
- the ant colony algorithms [Dorigo and Stützle, 2004, Matos and Oliveira, 2004, Melo et al., 2010, Monteiro et al., 2013, Socha and Dorigo, 2008, Vilarinho and Simaria, 2006, Xiao and Li, 2011].

Our proposal for globally solving the problem (1.1) is a distribution based AFS, hereafter denoted by DbAFS, that takes the AFS algorithm presented in [Rocha et al., 2011] as a heuristic basis. The idea presented in this DbAFS algorithm is borrowed from the bare bones swarm concept [Kennedy, 2003, Omran et al., 2008]. The novelty in DbAFS algorithm is that each trial point is sampled from a Gaussian distribution, instead of being randomly generated in basis of the corresponding current point and a direction of search. The center of the distribution is given by the midpoint between the current and a target point, featured by the selected fish behavior, and the dispersion is related with the difference between those two points. The goal here is to improve the accuracy and convergence behavior.

The remainder of the paper is organized as follows. Section 2 describes the proposed DbAFS algorithm and Section 3 presents and discusses the results of preliminary experiments. Finally, the conclusions of this study and the ideas for future work are presented in Section 4.

2 Distribution based artificial fish swarm

The AFS algorithm is a stochastic method that relies on a swarm intelligence based paradigm to construct fish/point movements over the search space [Jiang et al., 2007, Neshat et al., 2013]. We will use the words 'fish' and 'point' interchangeably throughout the paper. Each point in the space is represented by $x^j \in \mathbb{R}^n$ (the jth point of a population), m is the number of points in the population, where $m < \infty$, and the component i of a vector x^j is represented by x_i^j.

2.1 The AFS algorithm

We now describe the procedure used to generate trial points in the AFS algorithm [Rocha et al., 2011]. At each iteration, a population of m solutions/points, herein denoted by $x^1, x^2, \ldots, x^m$ is used to generate a set of trial points $y^1, y^2, \ldots, y^m$. The population is initialized randomly in the entire search space Ω using the equation: $x_i^j = l_i + U(0,1)(u_i - l_i)$ for each component $i = 1, \ldots, n$ of the point x^j, where the notation $U(0,1)$ represents a random number in $(0,1)$.

Each point x^j movement is defined according to the number of points inside its 'visual scope'. The 'visual scope' is defined as the closed neighborhood centered at x^j with a positive radius α. In the herein implemented versions of the AFS algorithm, the radius is dynamically defined as a fraction of the maximum distance between x^j and the other points x^l, $l \neq j$, $\alpha^j = \gamma \max_{l \neq j} \left\| x^j - x^l \right\|$, for $\gamma \in (0,1)$. Three possible situations may occur:

- the 'visual scope' is empty;
- the 'visual scope' is crowded;
- the 'visual scope' is not crowded.

When the 'visual scope' is empty, a *random* behavior is performed, in which the trial y^j is randomly generated inside the 'visual scope' of x^j.

When the 'visual scope' is crowded, with more than 80% of the population inside the 'visual scope' of x^j, a target point is randomly selected from the visual, x_{rand}. Then, if $f(x_{\text{rand}}) \leq f(x^j)$, the *searching* behavior is implemented and the trial point y^j is randomly defined along the direction from x^j to x_{rand}. Otherwise, the *random* behavior is performed.

When the 'visual scope' is not crowded, and $f(x_{\min}) \leq f(x^j)$, where $x_{\min}$ is the best point inside the 'visual scope', the *chasing* behavior is performed. This means that y^j is randomly generated along the direction from x^j to the target point $x_{\min}$. However, if $f(x_{\min}) > f(x^j)$ and the central point of the 'visual scope', denoted by $\bar{x}$, satisfies $f(\bar{x}) \leq f(x^j)$, then the *swarming* behavior is implemented instead. This means that the trial point is randomly defined along the direction from x_j to the target point $\bar{x}$. On the other hand, if $f(\bar{x}) > f(x^j)$, a target x_{rand} is randomly selected from the 'visual scope' and if $f(x_{\text{rand}}) \leq f(x^j)$, the *searching* behavior is carried out as previously described; otherwise the *random* behavior is performed.

We note that each point x^j generates a trial point y^j, inside the set Ω, as follows:

$$y_i^j = \max\left\{ l_i, \min\left\{ x_i^j + U(0,1)\, d_i^j, u_i \right\} \right\}, \text{ for } i = 1, \ldots, n \tag{2.1}$$

where d^j is a vector defined by the direction from x^j to one of the above referred target points.

To choose which point between the current x^j and the trial y^j will be a point of the population for the next iteration, the objective function at the two points are compared and if $f(y^j) \leq f(x^j)$ the trial point is passed to the next iteration as a current point. Otherwise, the current point is preserved to the next iteration/population.

2.2 Local search

In general, algorithms for globally solving an optimization problem have two separate phases:

- the global phase performs the exhaustive exploration of the search space using mostly stochastic methods to search for a promising region where a global minimum exists;
- the local phase exploits locally the promising region, using preferentially a deterministic method.

In the described AFS algorithm, a derivative-free deterministic method that exploits the neighborhood of a point for a better approximation using an exploratory move as well as a pattern move, and known as the Hooke and Jeeves (HJ) method [Hooke and Jeeves, 1961], is implemented.

This is a variant of the well-known coordinate search method. It incorporates a pattern move to accelerate the progress of the algorithm, by exploiting information obtained from the search in previous successful iterations. At the initial iteration, the exploratory move carries out a coordinate search centered at the best point of the population, with a step size of δ_{ls}. If a new trial point, y, with a better function value than x^{best} is encountered, the iteration is successful and δ_{ls} is maintained. Otherwise, the iteration is unsuccessful and δ_{ls} is reduced. If the previous iteration was successful, the vector $y - x^{\text{best}}$ defines a promising direction and a pattern move is then implemented, i.e., the exploratory move is carried out centered at the trial point $y+(y-x^{\text{best}})$, rather than at the current point y. Then, if the coordinate search is successful, the returned point is accepted as the new point; otherwise, the pattern move is rejected and the method reduces to a coordinate search centered at y. We refer to [Hooke and Jeeves, 1961] for details.

When the step size falls below δ_{tol}, a small positive tolerance, the search terminates since convergence has been attained [Kolda et al., 2003].

The pseudo-code for the AFS algorithm is presented below in Algorithm 1.

Algorithm 1 (AFS algorithm)

Data: $m \geq 1$, $0 < \gamma < 1$, $0 < \delta_{tol} \ll 1$.

1: Randomly generate m points in Ω, evaluate x^j, $j = 1, \ldots, m$, and select x^{best}.
2: **While** stopping conditions are not satisfied **do**
3: **For all** x^j, $j = 1, \ldots, m$ **do**
4: Generate trial point y^j using (2.1), evaluate y^j, $j = 1, \ldots, m$.
5: **If** $f(y^j) \leq f(x^j)$ **then** set $x^j = y^j$.
6: **End for**
7: Select x^{best}.
8: Perform local search on x^{best}.
9: **End while**

2.3 Distribution based artificial fish swarm

The distribution based artificial fish swarm is evolved from a classical version of the AFS algorithm [Rocha et al., 2011]. A Gaussian sampling strategy is proposed to create the trial points. After selecting the fish behavior to be implemented to each current point of the population, x^j, each component i ($i = 1, \ldots, n$) of the corresponding trial point, y_i^j, is randomly selected from a Gaussian distribution $N(\mu_i, \sigma_i)$, where:

- the mean is given by the average of the corresponding components of the current and the target point, $\mu_i = \dfrac{x_i^j + t_i}{2}$;
- the standard deviation is the absolute difference between the corresponding components of the current and the target point, $\sigma_i = \left| x_i^j - t_i \right|$;

and the target point t is x_{rand} if the *searching* behavior is selected, $x_{\min}$ if *chasing* is selected, or $\bar{x}$ if *swarming* is selected. Furthermore, to maintain the trial point inside the set Ω, the following equation is used:

$$y_i^j = \max\left\{ l_i, \min\left\{ y_i^j, u_i \right\} \right\}, \text{ for } i = 1, \ldots, n. \tag{2.2}$$

However, when a point x^j follows a *random* behavior, it will have a 50% chance that the component i of the point changes to the corresponding component of the best point of the population:

$$y_i^j = \begin{cases} x_i^j & \text{if } U(0,1) > 0.5 \\ x_i^{\text{best}} & \text{otherwise} \end{cases} . \tag{2.3}$$

We note that the exploration feature of the algorithm is ensured by the standard deviation which converges to zero as the points in the population converge to the optimal solution of the problem.

3 Numerical experiments

For a preliminary practical assessment of the proposed DbAFS algorithm, numerical experiments are carried out involving nine standard benchmark small test problems with known acronyms [Ali et al., 2005]: BR (n=2), CB6 (n=2), GP (n=2), H3 (n=3), H6 (n=6), SBT (n=2), S5, S7 and S10 (n=4). The algorithm is coded in C and the results are obtained on a PC with a 2.8 GHz Core Duo Processor P9700 and 6 Gb of memory. Each problem was solved 30 times and a population of $m = 10n$ points is used. After an empirical study, we also set $\gamma = 0.8$ and $\delta_{tol} = 10^{-8}$. The initial δ_{ls} is set to $10^{-3} \max_i\{u_i - l_i\}$.

The performance of the DbAFS algorithm is compared with the version 'AFS' algorithm described in [Rocha et al., 2011], the improved particle swarm optimization (PSO) algorithms - variants 'PSO-RPB' and 'PSO-HS' - proposed in [Ali and Kaelo, 2008], the differential evolution 'DE' (originally presented in [Storn and Price, 1997]), and 'MADDF' described in [Fernandes et al., 2012]. A brief description of the notation used in this section follows:

- 'DbAFS_{HJ}' is the distribution based AFS algorithm with the HJ local procedure (Section 2.2);
- 'DbAFS_{rand}' is the distribution based AFS algorithm with a simple random local search procedure;
- 'DbAFS' is the distribution based AFS algorithm without the local search;
- 'AFS_{HJ}' corresponds to the version presented in [Rocha et al., 2011], with the HJ local procedure;
- 'PSO-RPB' is a version of PSO with randomized *pbest* (best position of a particle) in the cognitive component;
- 'PSO-HS' is a variant of the PSO algorithm with a hybrid scheme (using the point generation scheme of DE) for position update;
- 'DE' is an improved version of the original differential evolution where the scaling factor is randomly chosen, for each point, and the crossover parameter is randomly chosen, for each iteration;
- 'MADDF' is a derivative-free multistart technique based on a filter set methodology for constraint-handling. (Bound constraints are incorporated into a constraint violation function.)

The stopping conditions used to stop the algorithms are [Ali and Kaelo, 2008]:

$$\left| f(x^{\text{best}}) - f_{\text{opt}} \right| \leq 0.001 \quad \text{or} \quad nf_{\text{eval}} > 20000, \tag{3.1}$$

where $f(x^{\text{best}})$ is the best solution found so far, f_{opt} is the known optimal solution, and nf_{eval} is the number of function evaluations required to obtain a solution with the specified accuracy. However, if the optimal solution of the problem is unknown, the algorithm may use other termination conditions.

First, we aim to analyze the effect on the performance of the DbAFS algorithm when a local search procedure is included in the algorithm (see Step 8 in Algorithm 1). Figure 1 depicts the average nf_{eval} after the 30 runs, required by DbAFS and DbAFS_{HJ} for each problem. We observe that the HJ local search has improved the efficiency of the algorithm when solving eight out of the nine tested problems. Only for the problem H6, DbAFS_{HJ} required in average more function evaluations than DbAFS to reach the solution with the accuracy specified in (3.1).

With an illustrative purpose, Figure 2 aims to compare the DbAFS_{HJ} algorithm performance for different parameter values. The plot on the left shows the convergence behavior for three values of the factor γ in the definition of the 'visual scope', on the problem H6: $\gamma = \{0.1, 0.5, 0.8\}$. The largest value of γ gives the fastest reduction in f when applied to H6. The plot on the right aims to show how population size affects the convergence of the algorithm. The tested values are $m = \{10, 10n, 100\}$ and the experiment was done on the problem H6 with $\gamma = 0.8$. Convergence is rather slow with the largest value of the population size.

Finally, Table 1 summarizes the results obtained in terms of the average number of function evaluations required by the algorithms to reach the optimal solution with the accuracy defined in (3.1). (The reported results of 'PSO-RPB', 'PSO-HS', 'DE' and 'MADDF' are adopted directly from [Ali and Kaelo, 2008, Fernandes et al., 2012].) Based on the results we may conclude that DbAFS_{HJ} has a good performance. The computational effort in terms of function evaluations has been reduced when compared with the

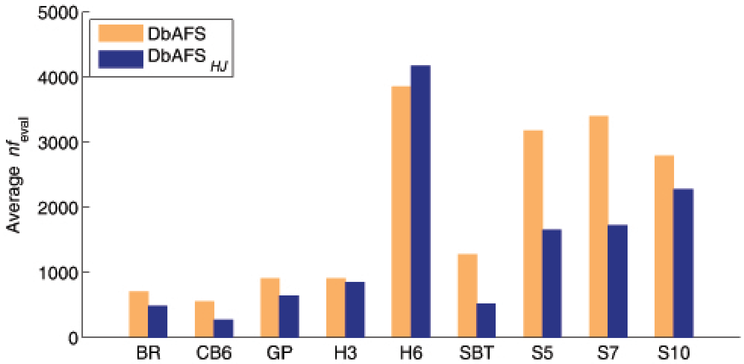


Figure 1: Average nf_{eval} for the nine problems: DbAFS vs. DbAFS$_{HJ}$.

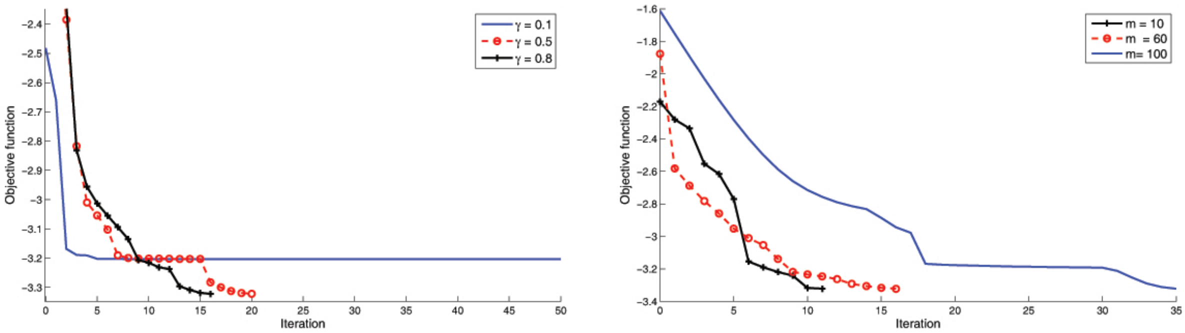


Figure 2: Factor γ effect (left) and population size effect (right), on the problem H6.

Table 1: Results of average number of function evaluations.

Prob.	f_{opt}	DbAFS		other methods				
		DbAFS$_{HJ}$	DbAFS$_{rand}$	AFS$_{HJ}$	PSO-RPB	PSO-HS	DE	MADDF
BR	0.39789	487	690	651	2652	2018	1305	506
CB6	-1.0316	274	293	246	2561	2390	1127	660
GP	3.00000	642	710	562	2817	1698	884	1063
H3	-3.86278	851	911	1573	3564	2948	1238	5845
H6	-3.32237	4167	3864	7861	8420	8675	7053	7559
SBT	-186.731	526	1256	659	4206	6216	2430	1867
S5	-10.1532	1650	1611	3773	6641	6030	5824	2929
S7	-10.4029	1723	1818	2761	6860	6078	5346	4428
S10	-10.5364	2282	1889	2721	6747	5602	4822	4489

basic AFS$_{HJ}$. The HJ local procedure incorporated into the DbAFS algorithm has improved in average the convergence to the optimal solution in six out of the nine tested problems, when compared with a simple random local search. The results also show that the DbAFS$_{HJ}$ performs better than the other methods in comparison (PSO variants, DE and MADDF).

4 Conclusions

This paper has introduced a new AFS algorithm which uses the Gaussian distribution to create the trial points, at every iteration. The main features of AFS are maintained, in particular the concept of 'visual scope' of each point and the selection of fish behavior. The proposed DbAFS algorithm performs well in a set of small benchmark problems, in terms of number of function evaluations. The effect of some parameters on the performance of the algorithm has been analyzed. A comparison with other stochastic population-based techniques available in the literature is carried out using small dimensional problems.

The results are promising in terms of reduced computational effort.

There are still some issues related to parameter setting that need further investigation. Extensive numerical experiments with larger dimensional problems remain to be done and will be reported in a future paper.

Acknowledgments

This work was financed by FEDER funds through COMPETE (Operational Programme Thematic Factors of Competitiveness) and by portuguese funds through FCT (Foundation for Science and Technology) within the projects FCOMP-01-0124-FEDER-022674 and PEst-C/MAT/UI0013/2011.

References

[Ali and Kaelo, 2008] Ali, M.M and Kaelo, P. (2008). Improved particle swarm algorithms for global optimization. *Applied Mathematics and Computation*, 196:578–593.

[Ali et al., 2005] Ali, M.M., Khompatraporn, C., and Zabinsky, Z. B. (2005). A numerical evaluation of several stochastic algorithms on selected continuous global optimization test problems. *Journal of Global Optimization*, 31:635–672.

[Bernardino et al., 2013] Bernardino, E.M., Bernardino, A.M., Sánchez-Pérez, J.M., Gómez-Pulido, J.A., and Vega-Rodríguez, M.A. (2013). Swarm optimization algorithms applied to large balanced communication networks. *Journal of Network and Computer Applications* 36(1) 504–522.

[Coelho et al., 2005] Coelho, J.P., Oliveira, P.B.M., and Cunha, J.B. (2005). Greenhouse air temperature predictive control using the particle swarm optimisation algorithm. *Computers and Electronics in Agriculture*, 49(3):330–344.

[Diwold et al., 2011] Diwold, K., Aderhold, A., Scheidler, A., and Middendorf, M. (2011). Performance evaluation of artificial bee colony optimization and new selection schemes. *Memetic Computing*, 3:149–162.

[Dorigo and Stützle, 2004] Dorigo, M. and Stützle, T. (2004). Ant Colony Optimization. MIT Press, Cambridge, Mass, USA.

[Engelbrecht, 2005] Engelbrecht, A. (2005). Fundamentals of Computational Swarm Intelligence. Wiley & Sons, New York.

[Fernandes et al., 2012] Fernandes, F.P., Costa, M.F.P., and Fernandes, E.M.G.P. (2012). A derivative-free filter driven multistart technique for global optimization. *Lecture Notes in Computer Science*, 7335, ICCSA 2012 Part III, Murgante, B. et al.(Eds.) 103–118.

[Hao and Hu, 2009] Hao, L. and Hu, L. (2009). Hybrid particle swarm optimization for continuous problems. *ISECS International Colloquium on CCCM 2009*, 3:217–220.

[Hooke and Jeeves, 1961] Hooke, R. and Jeeves, T.A. (1961). Direct search solution of numerical and statistical problems. *Journal of the Association for Computing Machinery*, 8:212–229.

[Jiang et al., 2007] Jiang, M., Wang, Y., Pfletschinger, S., Lagunas, M.A., and Yuan, D. (2007). Optimal multiuser detection with artificial fish swarm algorithm. In *CCIS Vol. 2*, ICIC 2007 Part 22, Huang, D.-S. et al.(Eds.) 1084–1093.

[Karaboga and Akay, 2009a)] Karaboga, D. and Akay, B. (2009). A comparative study of artificial bee colony algorithm. *Applied Mathematics and Computation*, 214(1):108–132.

[Karaboga and Akay, 2009b)] Karaboga, D. and Akay, B. (2009). A survey: algorithms simulating bee swarm intelligence. *Artificial Intelligence Review*, 31:61–85.

[Karaboga and Basturk, 2007] Karaboga, D. and Basturk, B. (2007). A powerful and efficient algorithm for numerical function optimization: artificial bee colony (ABC) algorithm. *Journal of Global Optimization*, 39(3):459–471.

[Kennedy, 2003] Kennedy, J. (2003). Bare bones particle swarms. In *Proc. IEEE Swarm Intelligence Symposium* 80–87.

[Kennedy and Eberhart, 1995] Kennedy, J. and Eberhart, R. (1995). Particle swarm optimization. In *Proc. IEEE Int. Conf. Neural Network 1995* 1942–1948.

[Kolda et al., 2003] Kolda, T.G., Lewis, R.M., Torczon, V. (2003). Optimization by Direct Search: New Perspectives on Some Classical and Moddern Methods. *SIAM Review*, 45(3):385–482.

[Matos and Oliveira, 2004] Matos, A.C. and Oliveira, R.C. (2004). An experimental study of the ant colony system for the period vehicle routing problem. In *Lecture Notes in Computer Science* 3172, ANTS 2004, M. Dorigo et al. (Eds.) 286–293.

[Melo et al., 2010] Melo, L., Pereira, F., and Costa, E. (2010). MC-ANT: A multi-colony ant algorithm. In *Lecture Notes in Computer Science* 5975, EA 2009, P. Collet et al. (Eds.) 25–36.

[Miranda et al., 2007] Miranda, V., Keko, H., and Jaramillo, A. (2007). EPSO: evolutionary particle swarms. *Studies in Computational Intelligence*, 66:139–167.

[Monteiro et al., 2013] Monteiro, M.S.R., Fontes, D.B.M.M., and Fontes, F.A.C.C. (2012). Concave minimum cost network flow problems solved with a colony of ants. *Journal of Heuristics*, 19(1):1–33.

[Neshat et al., 2013] Neshat, M., Sepidnam, G., Sargolzaei, M., and Toosi, A.N. (2013). Artificial fish swarm algorithm: a survey of the state-of-the-art, hybridization, combinatorial and indicative applications. *Artificial Intelligence Review*, DOI:10.1007/s10462-012-9342-2 (2013).

[Omran et al., 2008] Omran, M.G.H., Engelbrecht, A.P., and Salman, A. (2008). Bare bones differential evolution. *European Journal of Operational Research*, 196(1):128–139.

[Pires et al., 2010] Pires, E.J.S., Machado, J.A.T., Oliveira, P.B.M., Cunha, J.B., and Mendes, L. (2010). Particle swarm optimization with fractional-order velocity. *Nonlinear Dynamics*, 61(1-2):295–301.

[Rocha et al., 2011] Rocha, A.M.A.C., Fernandes, E.M.G.P., and Martins, T.F.M.C. (2011). Novel fish swarm heuristics for bound constrained global optimization problems. *Lecture Notes in Computer Science*, 6784, ICCSA 2011 Part III B. Murgante, B. et al.(Eds.) 185–199.

[Rocha and Fernandes, 2011] Rocha, A.M.A.C. and Fernandes, E.M.G.P. (2011). Mutation-Based Artificial Fish Swarm Algorithm for Bound Constrained Global Optimization. In *Numerical Analysis and Applied Mathematics ICNAAM 2011*: AIP Conf. Proc. 1389, 751–754.

[Socha and Dorigo, 2008] Socha, K. and Dorigo, M. (2008). Ant colony optimization for continuous domains. *European Journal of Operational Research*, 185(3):1155–1173.

[Storn and Price, 1997] Storn, R. and Price, K. (1997). Differential evolution – a simple and efficient heuristic for global optimization over continuous spaces. *Journal of Global Optimization*, 11:341–359.

[Vaz and Vicente, 2007] Vaz, A.I.F. and Vicente, L.N. (2007). A particle swarm pattern search method for bound constrained global optimization. *Journal of Global Optimization*, 39:197–219.

[Vilarinho and Simaria, 2006] Vilarinho, P.M. and Simaria, A.S. (2006). ANTBAL: an ant colony optimization algorithm for balancing mixed-model assembly lines with parallel workstation. *International Journal of Production Research*, 44(2):291–303.

[Wang et al., 2006] Wang, X., Gao, N., Cai, S., and Huang, M. (2006). An artificial fish swarm algorithm based and ABC supported QoS unicast routing scheme in NGI. *Lecture Notes in Computer Science*, 4331, ISPA 2006, Min, G. et al.(Eds.) 205–214.

[Xiao and Li, 2011] Xiao, J., and LiangPing Li, L. (2011). A hybrid ant colony optimization for continuous domains. *Expert Systems with Applications*, 38(9):11072–11077.

[Xiu-xi et al., 2010] Xiu-xi, W., Hai-wen, Z., and Yong-quan, Z. (2010). Hybrid artificial fish school algorithm for solving ill-conditioned linear systems of equations. In*IEEE international conference on intelligent computing and intelligent systems (ICIS)*, 390—394.

Recolha de Resíduos Sólidos Urbanos-otimização de rotas

Ana Maria Rodrigues*†‡, José Soeiro Ferreira*†

* *INESC TEC Porto, Portugal*
*E-mail:*amr@inescporto.pt

† *Faculdade de Engenharia da Universidade do Porto, Portugal*
*E-mail:*jsoeiro@inescporto.pt

‡ *Instituto Superior de Contabilidade e Administração - Instituto Politécnico do Porto.*

Resumo

Este trabalho dá a conhecer um novo problema, Problema Capacitado de Rotas em Arcos Misto, com Múltiplos Aterros Limitados. Baseado na situação de recolha/transporte de Resíduos Sólidos Urbanos no concelho de Monção, são apresentadas características que, não sendo únicas em Portugal, nunca foram mencionadas na literatura. Diferencia-se pela existência de diversos pontos de deposição que, especialmente devido às reduzidas dimensões, apresentam restrições relacionadas com o número de visitas recebidas por dia. Um novo modelo de otimização, baseado na formulação do Mixed Capacitated Arc Routing Problem é apresentado. São incluídos resultados computacionais provenientes de instâncias adaptadas da literatura e do problema real descrito.

Palavras chave: Recolha de Resíduos, Rotas, Problema Capacitado de Rotas em Arcos Misto, Modelo de Otimização, Aterros Múltiplos

1 Introdução

Uma boa gestão de Resíduos Sólidos Urbanos (RSU), depende em boa parte, da otimização de rotas de recolha. Melhores rotas traduzem-se, por exemplo, em menores gastos com combustíveis, valores estes que absorvem uma parte significativa dos custos. A situação real de recolha considerada neste trabalho, e que respeita à região de Monção, apresenta algumas particularidades: envolve aterros múltiplos (aterros ou estações de transferência) e restrições relacionadas com o número de visitas que cada aterro permite por veículo. Alguns dos pontos de receção de resíduos (estações de transferência) não permitem que os veículos façam, por exemplo, mais do que uma visita diária. Apesar de serem pontos que, muitas vezes, se encontram relativamente perto das povoações onde se pretende fazer a recolha, tornando-os alvos preferenciais no momento de esvaziar o camião, eles não podem ser usados indiscriminadamente. Para além da descrição da situação em questão, serão igualmente apresentados o Problema Capacitado de Rotas em Arcos Misto, com Múltiplos Aterros Limitados (Mixed Capacitated Arc Routing Problem with Limited Multi Landfills (MCARP-LML) e um modelo de otimização que constituem, tanto quanto nos é dado a conhecer, novas contribuições nesta área do conhecimento. O presente artigo começa por apresentar, no Capítulo 2, alguma da legislação e dos conceitos relacionados com a Recolha de RSU bem como a região que foi alvo do estudo. O Capítulo 3 descreve, de forma muito sucinta, os problemas de rotas em arcos e o conceito de problema capacitado, terminando com o Problema Capacitado de Rotas em Arcos Misto, com Múltiplos Aterros Limitados e o modelo de otimização associado. No Capítulo 4 incluem-se resultados computacionais relativos a instâncias adaptadas da literatura e a instâncias reais. Conclusões surgem no Capítulo 5.

2 Os Resíduos Sólidos Urbanos

Legislação e Conceitos

A Gestão dos Resíduos de uma maneira geral e a Recolha dos RSU, em particular, apresenta-se como uma preocupação crescente a nível mundial. Portugal e a Comunidade Europeia lançam diretivas no sentido de tentar resolver alguns problemas ou, pelo menos, reduzir o impacto negativo que os resíduos provocam no nosso planeta. "A correta gestão dos resíduos apresenta-se como um desafio inadiável para as sociedades modernas" (Decreto-Lei nº 239/97, de 9 de setembro). Entende-se por "Resíduos", segundo a

Diretiva 2006/12/CE do Parlamento Europeu e do Conselho de 5 de abril de 2006, "quaisquer substâncias ou objectos (...) de que o detentor se desfaz ou tem a intenção ou a obrigação de se desfazer". Sendo considerado "Resíduo urbano" como "o resíduo proveniente de habitações bem como outro resíduo que, pela sua natureza ou composição, seja semelhante ao resíduo proveniente de habitações" (Decreto-Lei no 178/2006, de 5 de setembro). Esta mesma Diretiva Europeia, apresenta ainda duas definições importantes para o trabalho aqui apresentado: "Gestão" e "Recolha". Fazem parte da "Gestão: a recolha, o transporte (...) dos resíduos" sendo "Recolha: a operação de apanha, triagem e/ou mistura de resíduos com vista ao seu transporte".

No sentido de uma melhor gestão dos resíduos, por forma a reduzir os percursos efetuados pelos veículos aquando da recolha, alguns municípios dispõem, para além de aterros, de estações de transferência. As estações de transferência surgem quando a distância entre o ponto que representa o centro das recolhas (centróide) e os aterros é elevada. Em muitos casos, toma-se "25km" como o valor de referência para a justificação da construção de uma estação de transferência. Não sendo esta uma imposição, há, no entanto, situações em que a distância é bastante superior e não é apresentada outra alternativa à frota senão o(s) aterro(s).

Valores apresentados no Plano Nacional de Gestão de Resíduos 2011-2020, relativos ao ano de 2008, revelam que Portugal, apesar de se encontrar abaixo da média da UE no que diz respeito à produção de resíduos urbanos e equiparados e de não urbanos, apresenta uma capitação nacional de produção de resíduos de 3,4 t/hab.ano. O mesmo plano indica que uma parte muito significativa dos resíduos urbanos e equiparados, mais de 60%, são eliminados em aterro (dados até 2009).

2.1 Concelho de Monção

O caso prático que aqui se apresenta surgiu pelo envolvimento direto com o problema da recolha de RSU no concelho de Monção. O contacto com as empresas responsáveis pela Recolha de RSU na região, também designadas para a Recolha de Resíduos Industriais, para a limpeza do arruamento e para a varredura em Monção, permitiu-nos ter acesso à realidade deste concelho.

Características Geográficas e Populacionais

O concelho de Monção, inserido no distrito de Viana do Castelo, conta com uma área de 211,51 km^2 dividida em 33 freguesias, ver Figura 1, e com uma população de 19230 habitantes (dados de 2011).

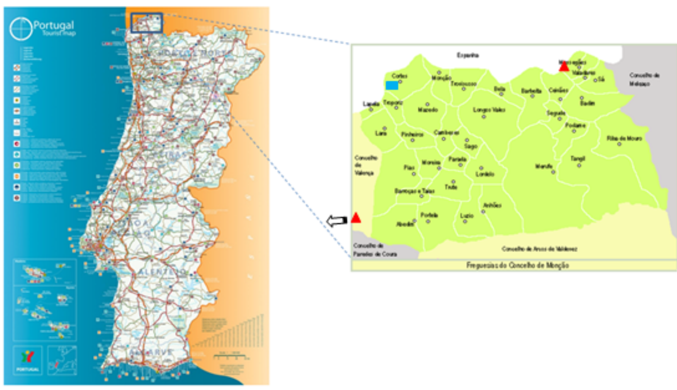


Figura 1: Localização do concelho de Monção com as suas 33 freguesias. (À direita: triângulos vermelhos representam aterro e estação de transferência e retângulo azul a garagem)

Sendo Monção uma região do interior do território muito sujeita às vagas de emigração/migração, esta região apresenta, durante o período de verão e épocas festivas, um aumento significativo da população residente. Este aumento traduz-se, de forma bem notória, num aumento na quantidade de resíduos produzidos levando a um aumento no número de visitas semanais por parte dos veículos de recolha.

Características dos dispositivos afetos ao local

Atualmente, o concelho conta com mais de 1600 contentores que se distribuem pelos diferentes tipos e capacidades. Os mais comuns são os contentores de 800 litros em polietileno.

Os contentores de maiores dimensões, Moloks ou semienterrados e enterrados, são ainda raros sendo mais usuais em regiões mais urbanas. Estes últimos contentores obrigam à existência de grua no veículo coletor.

Há ainda a salientar situações, embora em menor escala, de recolha "porta-a-porta", o vulgar saco plástico deixado na via pública, em frente à porta ou pequenos baldes que são colocados pela população ao longo do percurso de recolha.

Relativamente aos veículos de recolha, a frota é composta por dois veículos (com grua) com capacidade para $16m^3$ e um outro, mais pequeno, de $7m^3$ usado como reforço.

A garagem, local de início e de fim de percurso para todos os veículos, situa-se a sudoeste da vila de Monção. Para deposição dos seus RSU, Monção conta com um aterro em S.Pedro da Torre, Valença, e uma estação de transferência em Messegães, geridos pela empresa Valorminho - Valorização e Tratamento de Resíduos Sólidos, S.A..

3 Otimização de Rotas

Os problemas relacionados com rotas, usualmente representados por grafos, refletem a necessidade de definir um "caminho" que percorre todos ou parte dos nós e/ou ligações que formam esse mesmo grafo. Na maioria dos casos, procura-se que tal caminho minimize um dado custo. Algumas restrições, mais ou menos complexas, poderão ser adicionadas ao problema. Por exemplo, obrigar a que o caminho obedeça a uma hierarquia pré-estabelecida, um determinado conjunto de nós ou vértices só poderá ser alcançado depois de outros o serem ou associar diferentes custos consoante o sentido do percurso.

3.1 Problemas de Rotas em Arcos

Os problemas mais gerais de rotas (General Routing Problem - GRP) envolvem situações em que é necessário servir um determinado número de nós e/ou ligações entre esses mesmos nós. O objetivo passa, geralmente, como foi referido anteriormente, por cumprir essas necessidades com o menor custo. Este grupo enorme de problemas pode subdividir-se em dois grupos mais pequenos: os Problemas de Rotas em Nós (Node Routing Problem - NRP) e os Problemas de Rotas em Arestas ou Arcos (Arc Routing Problem - ARP). Como o próprio nome indica, sempre que se está perante um problema que apresenta obrigatoriedade de passagem em determinados nós ou vértices do grafo que lhe corresponde, está-se perante um NRP. Um exemplo bastante conhecido de um NRP é o Problema do Caixeiro Viajante (Traveling Salesman Problem - TSP). Imagine-se a situação em que um caixeiro viajante, partindo de uma determinada cidade C deverá visitar, com o menor custo, um determinado número n de cidades $(C_1, C_2, ..., C_n)$, regressando, no final, à cidade C.

Os ARPs estão relacionados com a determinação do conjunto (ou subconjunto) de arestas num grafo $G = (V, E)$ que apresentam menor custo de "passagem", onde V é o conjunto de vértices e E o conjunto de arestas. Este grafo poderá ser não dirigido, dirigido (com arcos em vez de arestas) ou misto (com arcos e arestas). A cada aresta, $(i, j), i \neq j$, está associado um custo $c_{ij} \geq 0$. Estes problemas aparecem numa grande variedade de situações práticas tais como entrega de correio, recolha de resíduos urbanos, limpeza de ruas, remoção de neve das ruas, definição de percursos de corte em ambiente industrial, definição de rotas em autocarros escolares entre outras.

3.2 Problemas Capacitados

Se forem adicionadas restrições às capacidades dos veículos surgem os problemas capacitados. O Problema Capacitado de Rotas em Arcos (Capacitated Arc Routing Problem - CARP) pode ser definido num grafo $G = (V, E)$, não dirigido, dirigido ou misto (MCARP), com um subconjunto E_R ($E_R \subseteq E$) de arestas requeridas, ou seja, arestas de passagem obrigatória, e um vértice especial , v_0, chamado depósito. Uma procura $q_e > 0$ está associada a cada uma das arestas requeridas $e \in E_R$. Uma frota de n veículos (não necessariamente idênticos), cada um com capacidade $Q_i > 0, i = 1, ..., n$ apresenta-se disponível. O objetivo é encontrar o conjunto de rotas de custo mínimo. Essas rotas deverão ser tais que, cada uma delas, deverá iniciar e terminar no depósito e cada aresta requerida deverá ser servida uma e uma só vez. Ainda deverá verificar-se que, a soma das procuras, em cada rota, não deverá ultrapassar a capacidade disponibilizada pelo veículo, [Letchford e Oukil(2009)], [Longo et al.(2006)]. O CARP é um problema NP-difícil e foi introduzido por [Golden e Wong(1981)].

Outros problemas relacionados com o CARP aparecem descritos na literatura, são exemplos: o CARP Periódico (Periodic CARP - PCARP), definido em [Lacomme et al.(2002)]; o CARP Estocástico (Stochastic CARP), em [Fleury et al.(2004)]; o CARP com Pontos de Recarga (CARP with Refill Points - CARP-RP), [Amaya et al.(2004)], o Problema do Carteiro Chinês Capacitado (Capacitated Chinese Postman Problem - CCPP) onde as procuras, em todas as arestas, são positivas ou o ARP com Setorização (Sectoring ARP - SARP) onde, para além da definição de rotas, se coloca a questão da separação da rede em setores; entre outros. Em [Wøhlk(2008)] pode ser encontrada uma revisão da literatura nesta área.

3.3 Problema Capacitado de Rotas em Arcos Misto, com Múltiplos Aterros Limitados

O Problema Capacitado de Rotas em Arcos Misto, com Múltiplos Aterros Limitados (Mixed Capacitated Arc Routing Problem with Limited Multi-Landfills (MCARP-*LML*)) apresenta algumas características que são comuns a alguns Problemas Capacitados. Entre estas encontram-se o facto de algumas ruas apresentarem sentidos obrigatórios (problema misto); todos os veículos deverão partir e regressar vazios, no final do turno de trabalho, a um ponto denominado depósito ou garagem; sempre que necessário, os veículos serão esvaziados num dos locais à disposição para o efeito.

No entanto, este mesmo problema tem algumas características que, até ao momento, nunca apareceram descritas na literatura. Para além de apresentar mais do que um ponto de deposição de resíduos, o que por si só não introduz qualquer novidade em termos de literatura (ver [Xing et al(2010)], há que salientar que esses pontos não têm todos a mesma capacidade de receção. Enquanto um aterro pode ser usado por um veículo coletor sempre que o veículo necessitar de efetuar uma descarga, a estação de transferência poderá, por razões relacionadas com as suas dimensões ou outras, limita o número de acessos diários por veículo. Este problema está relacionado, em particular, com a situação de Recolha de Resíduos que se verifica em Monção e que foi referida anteriormente.

3.4 Modelo de otimização para o MCARP-*LML*

Seguidamente, será apresentado um novo modelo para o MCARP-LML, que minimizará o custo das viagens entre o aterro/estação de transferência e o depósito, ao mesmo tempo que a procura é satisfeita sem que a capacidade dos veículos coletores seja ultrapassada. Este modelo baseou-se no MCARP apresentado em [Gouveia et al.(2010)]. É importante referir que a situação aqui descrita, no que ao caso real apresentado diz respeito, é antecedida pela resolução de um problema de setorização. Ou seja, a região em estudo é, numa primeira fase, dividida em zonas de menor dimensão reduzindo, assim, significativamente, a dimensão das instâncias.

Importa começar por introduzir as definições e as notações que serão usadas ao longo deste trabalho. Suponha-se um grafo misto $G = (V^*, L)$ cujo conjunto dos vértices está representado por V^* e as ligações entre eles por $L = A \cup E$ (A e E representam, respetivamente, o conjunto dos arcos e das arestas).

Não havendo motivo para confusão, aterros e estações de transferência serão tratados de forma indiferenciada por aterros. As diferenças surgirão no momento de analisar o número de visitas permitidas em cada caso.

$V^* = V' \cup V$ onde $V' = B \cup D$. B representa o conjunto dos vértices que correspondem a aterros (aterros ou estações de transferência), D representa o depósito e V representa o conjunto dos vértices *vulgares*(vértices que não representam nem depósito, nem aterros).

R é o conjunto de todas as ligações requeridas (de passagem obrigatória): Arcos requeridos A_R e arestas requeridas E_R.

T representa o número de viagens, uma viagem é considerada um caminho entre dois vértices de V'.

Q é a capacidade do veículo.

d_{ij} corresponde ao custo de passagem (sem recolha) pela ligação $(i, j) \in L$.

$c_{ij} > 0$ é o custo de servir (recolher resíduos) a ligação $(i, j) \in R$ ($c_{ij} = 0$ se $(i, j) \in \overline{R}$).

$q_{ij} > 0$ é a procura na ligação $(i, j) \in R$ ($q_{ij} = 0$ se $(i, j) \in \overline{R}$).

A primeira viagem deverá começar no depósito bem como a última deverá lá terminar. Quer no momento inicial, ao sair de depósito, quer no regresso o veículo deverá estar vazio. No decorrer de um turno de trabalho, o veículo deverá esvaziar $t - 1$ vezes. Cada aterro i apresenta um número máximo de visitas permitidas l_i. O depósito funciona, unicamente, como garagem, não é permitido fazer descargas nesse ponto.

$$x_{ijb}^t = \begin{cases} 1, & \textit{se aresta (i,j) é servida na viagem t e residuo esvaziado no aterro b} \\ 0, & \textit{caso contrário} \end{cases}$$

y_{ij}^t representa o número de vezes que uma ligação (aresta ou arco) (i,j) é atravessada sem ser servida durante a viagem t.

z_{ij}^t é a quantidade de resíduo recolhido durante a viagem t depois de atravessar a ligação (i,j).

O objetivo é o de minimizar o custo das T viagens (3.1).

Min

$$\sum_{t=1}^{T}\left(\sum_{(i,j)\in R}\sum_{b\in B} x_{ijb}^t \cdot c_{ij} + \sum_{(i,j)\in L} y_{ij}^t \cdot d_{ij}\right) \tag{3.1}$$

suj. a

$$\sum_{t=1}^{T}\left(\sum_{(i,j)\in L} y_{ij}^t + \sum_{b\in V'}\sum_{(i,j)\in R} x_{ijb}^t\right) = \sum_{t=1}^{T}\left(\sum_{(j,i)\in L} y_{ji}^t + \sum_{b\in V'}\sum_{(j,i)\in R} x_{jib}^t\right), \qquad \forall i \in V^* \tag{3.2}$$

$$\sum_{(i,j)\in L} y_{ij}^t + \sum_{b\in B}\sum_{(i,j)\in R} x_{ijb}^t = \sum_{(j,i)\in L} y_{ji}^t + \sum_{b\in B}\sum_{(j,i)\in R} x_{jib}^t, \qquad \forall i \in V; t = 1,...,T \tag{3.3}$$

$$\sum_{t=1}^{T}\sum_{b\in B} x_{ijb}^t = 1, \qquad \forall (i,j) \in A_R \tag{3.4}$$

$$\sum_{t=1}^{T}\sum_{b\in B}(x_{ijb}^t + x_{jib}^t) = 1, \qquad \forall (i,j) \in E_R \tag{3.5}$$

$$\sum_{b\in V':(i,b)\in L} y_{ib}^t = 1, \qquad t = 1,...,T \tag{3.6}$$

$$\sum_{b\in V':(b,j)\in L} y_{bj}^t = 1, \qquad t = 1,...,T \tag{3.7}$$

$$\sum_{t=1}^{T}\sum_{b\in B} x_{ijb}^t = 0, \qquad \forall (i,j) \in \overline{R} \tag{3.8}$$

$$x_{ij1}^t = 0, \qquad \forall (i,j) \in R; t = 1,...,T \tag{3.9}$$

$$\sum_{(i,b)\in L} y_{ib}^t = \sum_{(b,j)\in L} y_{bj}^{t+1}, \qquad t = 1,...,T-1; \forall b \in B \tag{3.10}$$

$$\sum_{t=1}^{T}\sum_{(b,j)\in L} y_{bj}^t \leq l_b, \qquad \forall b \in B \tag{3.11}$$

$$\sum_{(i,b)\in L} y_{ib}^t + \sum_{(b,j)\in L} y_{bj}^t \leq 2, \qquad t = 1,...,T; \forall b \in V' \tag{3.12}$$

$$\sum_{(i,j)\in L} z_{ij}^t - \sum_{(j,i)\in L} z_{ji}^t = \sum_{(i,j)\in R}\sum_{b\in B} q_{ij} \cdot x_{ijb}^t, \qquad \forall i \in V; t = 1,...,T \tag{3.13}$$

$$\sum_{(i,b)\in L} z_{ib}^t = \sum_{(i,j)\in R} q_{ij} x_{ijb}^t, \qquad t = 1,...,T; \forall b \in B \tag{3.14}$$

$$\sum_{(b,j)\in L} z_{bj}^t = 0, \qquad t = 1,...,T; \forall b \in B \tag{3.15}$$

$$z_{ij}^t \leq Q(y_{ij}^t + \sum_{b\in B} x_{ijb}^t), \qquad \forall (i,j) \in L; t = 1,...,T \tag{3.16}$$

$$x_{ijb}^t \in \{0,1\}, \qquad \forall (i,j) \in L; \forall b \in B; t = 1,...,T \tag{3.17}$$

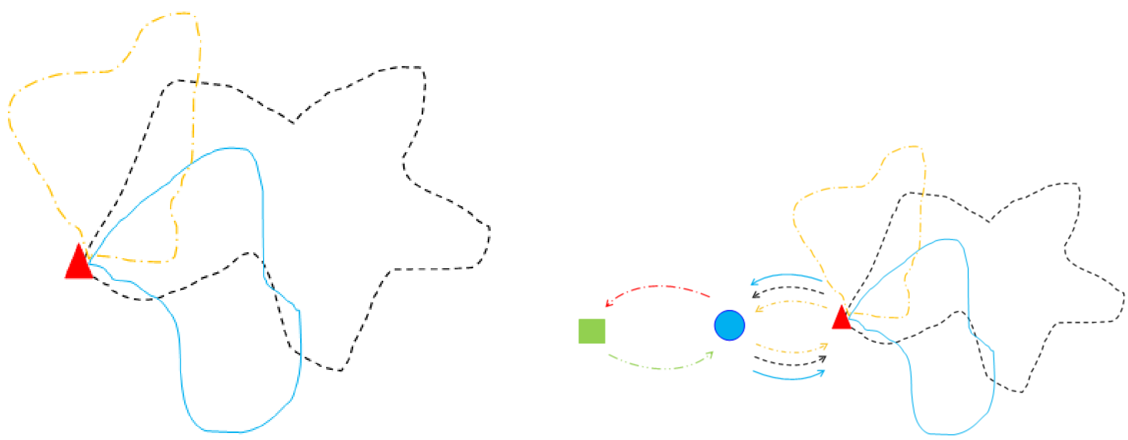
$$z_{ij}^t \geq 0, \qquad \forall (i,j) \in L; t = 1, ..., T \tag{3.18}$$

$$y_{ij}^t \geq 0, \qquad integer, \forall (i,j) \in L; t = 1, ..., T \tag{3.19}$$

O conjunto de todas as viagens deverá formar um caminho fechado (3.2, 3.3). Cada arco requerido (3.4) e cada aresta requerida (3.5) deverão ser servidos apenas uma vez. Cada viagem entra (3.6) e sai (3.7) de determinado aterro apenas uma vez. Uma aresta facultativa não é servida (3.8). A garagem não é um aterro (3.9). Quando um veículo entra num aterro numa determinada viagem, ele deverá iniciar no mesmo aterro a viagem seguinte (3.10). O número de visitas permitidas a cada aterro é limitado (3.11). Cada viagem deverá começar num aterro ou garagem e terminar num aterro ou garagem (3.12). A quantidade de resíduo recolhido numa viagem é atualizada (3.13, 3.14) depois de cada serviço. Um veículo, depois de abandonar um aterro deverá estar vazio (3.15). A quantidade de resíduo recolhido não pode ultrapassar a capacidade do veículo (3.16).

3.5 MCARP-*LML* uma extensão do MCARP

O MCARP-*LML* é uma extensão do MCARP. A Figura 2 ilustra esta afirmação. Na Figura 2(a), a título de exemplo, aparece representada uma solução para o MCARP com 3 circuitos e um depósito (representado pelo triângulo vermelho). Cada um dos três circuitos começa e termina no depósito. Na Figura 2(b) a mesma situação aparece retratada mas com: um novo depósito (quadrado verde), o antigo depósito aparece agora retratado como um vértice vulgar; um aterro (círculo azul) que permite, pelo menos, 3 visitas e duas arestas de procura nula (arestas facultativas) cujos custos de travessia são igualmente nulos. Assim sendo, a solução ótima para a nova situação gerada apresenta, exatamente, o mesmo valor da apresentada em Figura 2(a). Como uma viagem é considerada o caminho entre aterros ou entre depósito e aterro, na nova situação, serão necessárias não 3 mas 5 viagens representadas por 5 cores distintas.



(a) **(b)**

Figura 2: (a) Solução para um MCARP (b) MCARP-*LML* com um novo depósito e um novo aterro

Nestas condições, sendo o MCARP-*LML* uma extensão do MCARP que é um problema NP-Difícil, ([Belenguer et al.(2006)]), pode deduzir-se que o MCARP-*LML* também o é.

4 Resultados Computacionais

Nesta secção são apresentados os resultados computacionais. Os testes foram efetuados num computador Intel Core2 Duo 3,00 GHz with 8.00 Gb of RAM e as soluções foram obtidas com CPLEX 12.2. Os resultados apresentam-se distribuídos por dois grupos de instâncias. Ao Grupo 1, descrito na literatura como GDB ([Belenguer e Benavent(2003)]), foram feitas algumas alterações de forma a conseguir testar o MCARP-*LML*. O Grupo 2 de instâncias está relacionado com o caso real aqui apresentado. Os resultados dos dois grupos aparecem descritos nas tabelas 1 e 2.

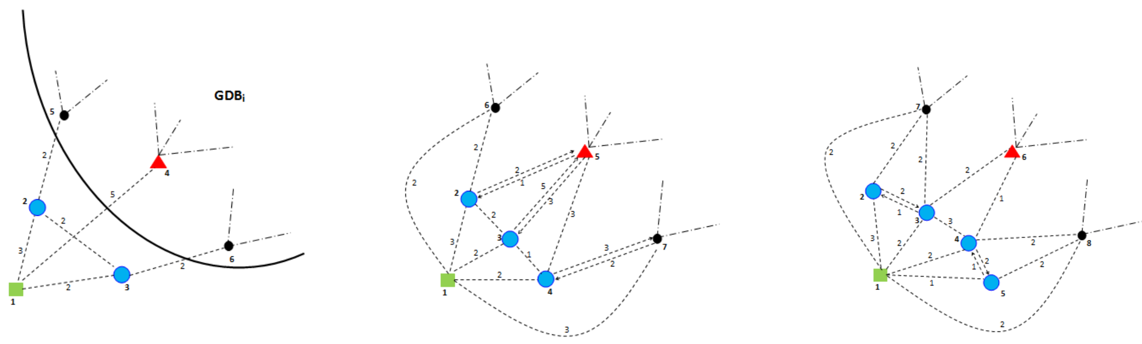
As colunas apresentadas nas duas tabelas referem-se a:

- ID: identificação da instância;
- T: número de viagens;
- $\sharp\ V$: número de vértices (depósito+aterros+vértices comuns);
- $\sharp\ R$: número de ligações requeridas ;
- $\sharp\ \overline{R}$: número de ligações não requeridas;
- Q: capacidade do veículo;
- $\sharp\ B$: número de aterros;
- $[Maxl_i]_{i=1,...,\sharp B}$: número máximo de descargas para cada aterro i;
- *Gap*(%): *gap* obtido com CPLEX 12.2;
- $Tempo(seg)$: tempo computacional em segundos;
- *FO*: função objetivo - valor ótimo (**Negrito***) ou um limite superior.

4.1 Grupo 1: GDB_i_LML e $GDB_{15}\_LML_i$

A tabela 1 mostra um conjunto de 23 instâncias GDB_i_LML, $i = 1, ..., 23$ e outras 3 instâncias desenvolvidas a partir de GDB_{15}_LML (disponíveis em `http://www.inescporto.pt/~amr/Limited_Multi_Landfills`). O grupo de instâncias GDB_i_LML foi obtido após algumas modificações aplicada às instâncias GDB ($GDB_i, i = 1, ..., 23$) apresentadas em [Belenguer e Benavent(2003)]. Estas alterações devem-se ao facto de, nas instâncias originais ser considerado apenas um depósito/aterro. As modificações em causa consistiram, como se representa na Figura 3(a), na adição de três vértices a cada uma das 23 instâncias GDB_i. Assim, o vértice 1 corresponderá ao depósito, os vértices 2 e 3 representam 2 aterros. Os vértices de cada uma das instâncias GDB_i correspondem, pela mesma ordem, aos vértices 4 e seguintes nas novas instâncias GDB_i_LML.

As instâncias $GDB_{15}\_LML_i$ provêm de novas alterações efetuadas à instância GDB_{15}, com 2, 3 ou 4 aterros de acordo com o que consta das Figuras 3(a),(b) e (c), respetivamente.



(a) **(b)** **(c)**

Figura 3: Alterações produzidas em: (a) GDB_i_LML (b) GDB_{15}_LML2 (3 aterros) (b)GDB_{15}_LML3 (4 aterros)

Os valores obtidos revelam, tal como esperado, que o processo usado revela-se eficaz para instâncias de média ou reduzida dimensão. Os piores resultados ocorreram, em geral, para instâncias com mais do que 7 viagens, o que corresponde a mais de 6 descargas. É de salientar que, na prática, tais valores estão muito acima dos praticados pelos veículos coletores. No caso aqui retratado, não é comum os veículos efetuarem, num mesmo turno de trabalho, mais do que 3 descargas, ou seja, 4 viagens.

4.2 Grupo 2 - Situação real

De maneira a melhor ilustrar o modelo do MCARP-*LML*, são apresentadas instâncias baseadas no caso real anteriormente apresentado. Os dados constam da tabela 2 (mais informações podem ser obtidas em $http://www.inescporto.pt/\sim amr/Limited\_Multi\_Landfills/RealCase$):

Tabela 1: Características e resultados de 23 instâncias GDB_i_LML e 3 $GDB_{15}\_LML_i$

ID	T	$\sharp V$	$\sharp R$	$\sharp \overline{R}$	Q	$\sharp$ B	$[Maxl_i]_{i=1,\dots,\sharp B}$	Gap(%)	Tempo (seg)	FO
GDB_1_LML	6	15	22	6	5	2	[3,3]	-	244,63	**310***
GDB_2_LML	7	15	26	6	5	2	[4,4]	-	1384,60	**347***
GDB_3_LML	6	15	22	6	5	2	[3,3]	-	30,70	**286***
GDB_4_LML	5	14	19	6	5	2	[3,3]	-	62,70	**293***
GDB_5_LML	7	16	26	6	5	2	[3,3]	1,86	Out-of-Mem	376
GDB_6_LML	6	15	22	6	5	2	[3,3]	-	47511,38	**317***
GDB_7_LML	6	15	22	6	5	2	[3,3]	-	182,40	**329***
GDB_8_LML	11	30	46	6	27	2	[5,5]	13,53	Out-of-Mem	374
GDB_9_LML	11	30	51	6	27	2	[5,5]	17,77	Out-of-Mem	378
GDB_{10}_LML	5	15	25	6	10	2	[2,2]	-	1,90	**296***
GDB_{11}_LML	6	25	45	6	50	2	[3,3]	-	367,49	**414***
GDB_{12}_LML	8	16	23	6	35	2	[4,4]	8,79	Out-of-Mem	472
GDB_{13}_LML	7	13	28	6	41	2	[4,4]	2,15	Out-of-Mem	561
GDB_{14}_LML	6	10	21	6	21	2	[3,3]	-	1,64	**123***
GDB_{15}_LML	5	10	21	6	37	2	[2,2]	-	1,28	**78***
GDB_{16}_LML	6	11	28	6	24	2	[3,3]	-	4,73	**149***
GDB_{17}_LML	6	11	28	6	41	2	[3,3]	-	1,87	**115***
GDB_{18}_LML	6	12	36	6	37	2	[3,3]	-	46,86	**187***
GDB_{19}_LML	4	11	11	6	27	2	[2,2]	-	0,33	**77***
GDB_{20}_LML	5	14	22	6	27	2	[2,2]	-	7,16	**150***
GDB_{21}_LML	7	14	33	6	27	2	[3,3]	-	166,16	**184***
GDB_{22}_LML	9	14	44	6	27	2	[5,5]	0,43	Out-of-Mem	232
GDB_{23}_LML	11	14	55	6	27	2	[6,6]	6,29	Out-of-Mem	286
GDB_{15}_LML1	5	10	21	6	37	2	[1,4]	-	1,36	**78***
GDB_{15}_LML2	5	11	21	9+6	37	3	[2,2,2]	-	0,81	**73***
GDB_{15}_LML3	5	12	21	13+4	37	4	[1,1,1,2]	-	0,62	**69***

- as instâncias "AltoMinhoX" representam uma situação rural;
- uma ligação (aresta ou arco) pode ser uma rua ou um conjunto de ruas (o que leva a uma redução da quantidade de dados a tratar);
- $\sharp\ R$ - o número de ligações requeridas, ou seja, ruas com contentores;
- $\sharp\ \overline{R}$ - número de ligações facultativas, ou ruas sem contentores;
- se numa ligação (rua ou conjunto de ruas) há mais do que um contentor, um falso contentor cuja capacidade será igual à soma da capacidade de todos os contentores nessa ligação, será adicionado;
- o custo (traduzido em tempo despendido) atribuído a cada ligação terá em atenção vários fatores: o comprimento da ligação, o número contentores a recolher, o tempo de recolha de cada tipo de contentor e a velocidade média do veículo.

Tabela 2: Características e resultados de uma situação real

ID	T	$\sharp V$	$\sharp R$	$\sharp \overline{R}$	Q	$\sharp B$	$[Maxl_i]_{i=1,\dots,\sharp B}$	Tempo (seg)	FO
AltoMinhoA	3	49	31	31	18	2	[1,3]	9,72	**12973***
AltoMinhoB	4	60	59	17	16	2	[1,3]	23,59	**29918,16***

Quando considerada apenas uma parte (um setor) do concelho o processo usado revelou-se eficaz na obtenção do percurso.

Na imagem superior da Figura 4 estão representados, a diferentes cores, os pontos de recolha das diferentes freguesias de Monção bem como o aterro, a estação de transferência e a garagem. Na imagem inferior é apresentado, um detalhe do percurso a efectuar pelo veículo coletor nas freguesias de Barbeita, Bela e Ceivães (correspondente à instância AltoMinhoB).

5 Conclusão

A situação real de recolha de Resíduos Sólidos Urbanos aqui retratada envolve algumas características que, até ao momento, nunca antes haviam sido descritas na literatura. Essas características prendem-se com o limite imposto ao número de visitas a algum (alguns) do(s) aterro(s) (aterros ou estações de

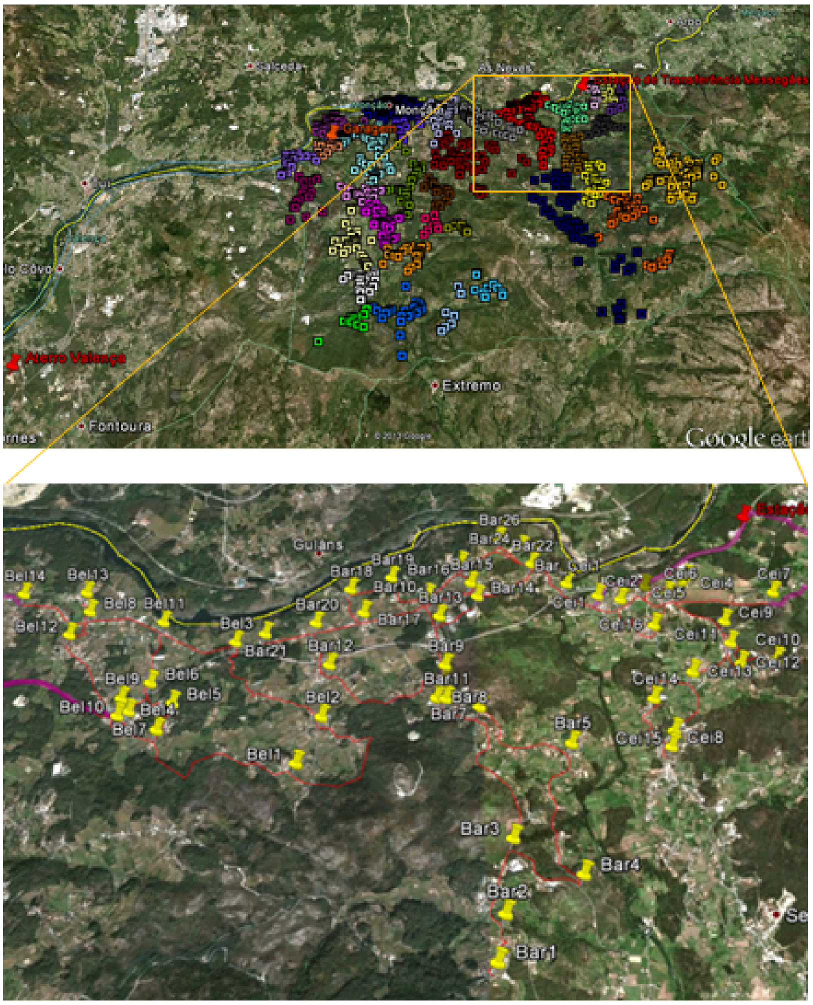


Figura 4: Detalhe da região nordeste do concelho de Monção

transferência). Essa imposição deu origem a um novo problema que foi caracterizado como Problema Capacitado de Rotas em Arcos Misto, com Múltiplos Aterros Limitados e que surgiu como uma generalização do Problema Capacitado de Rotas em Arcos Misto, bem conhecido da literatura. Para além de apresentar um novo problema, este artigo apresenta contribuições ao nível do desenvolvimento de um novo modelo de otimização. Usando o CPLEX 12.2, algumas instâncias foram resolvidas. Um primeiro grupo de instâncias foi obtido às custas de adaptações conseguidas a partir de instâncias referidas na literatura e outro de instâncias provenientes do caso real aqui retratado. Tal como esperado, o método usado na resolução revelou-se eficaz para instâncias de pequena e média dimensão.

Agradecimentos

Trabalho financiado pela FCT - Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) dentro do projeto "Sectoring and Routing Optimisation for Waste Management - theory into practice /PTDC/EGE GES/121406/2010".

Referências

[Amaya et al.(2004)] A. Amaya, A. Langevin, and M. Trépanier. The capacitated arc routing problem with refill points. *Les Cahiers du GERAD*, G-2004-100, 2004.

[Belenguer e Benavent(2003)] J. M. Belenguer and E. Benavent. A cutting plane algorithm for the capacitated arc routing problem. *Computers & Operations Research*, 30:705–728, 2003.

[Belenguer et al.(2006)] J.-M. Belenguer, E. Benavent, P. Lacomme, and C. Prins. Lower and upper bounds for the mixed capacitated arc routing problem. *Computers & Operations Research*, 33: 3363–3383, 2006.

[Fleury et al.(2004)] G. Fleury, P. Lacomme, and C. Prins. Evolutionary algorithms for stochastic arc routing problems. In *Lecture Notes in Computer Science*, 2004. Only first page.

[Golden e Wong(1981)] B. Golden and R. Wong. Capacitated arc routing problem. *Networks*, 11:305–315, 1981.

[Gouveia et al.(2010)] L. Gouveia, M. C. Mourão, and L. S. Pinto. Lower bounds for the mixed capacitated arc routing problem. *Computers & Operations Research*, 37:692–699, 2010.

[Lacomme et al.(2002)] P. Lacomme, C. Prins, and W. Ramdane-Chérif. Evolutionary algorithms for multiperiod arc routing problems. In *9th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, 2002.

[Letchford e Oukil(2009)] A. N. Letchford and A. Oukil. Exploiting sparsity in pricing routines for the capacitated arc routing problem. *Computers & Operations Research*, 36(7):2320–2327, 2009.

[Longo et al.(2006)] H. Longo, M. P. Aragão, and E. Uchoa. Solving capacitated arc routing problems a transformation to the cvrp. *Computers & Operations Research*, 33:1823–1837, 2006.

[Wøhlk(2008)] S. Wøhlk. *The Vehicle Routing Problem: Latest Advances and New Challenges*, chapter A Decade of Capacitated Arc Routing, pages 29–48. Springer Science+Business Media, 2008.

[Xing et al(2010)] L. Xing, P. Rohlfshagen, Y. Chen and X. Yao. An Evolutionary Approach to the Multidepot Capacitated Arc Routing Problem. *IEEE Transactions on Evolutionary Computation*, 14:356–374, 2010.

Energy Efficient Routing for Telecommunication Networks with Multiperiod Traffic

Dorabella Santos*, Carlos Lopes*, Amaro de Sousa*†, Filipe Alvelos‡

* *Instituto de Telecomunicações – Pólo de Aveiro, Aveiro, Portugal*
*E-mail:*dorabella@av.it.pt, clopes@av.it.pt, asou@av.it.pt

† *DETI, Universidade de Aveiro, Aveiro, Portugal*
*E-mail:*asou@ua.pt

‡ *DPS/Centro Algoritmi, Universidade de Minho, Braga, Portugal*
*E-mail:*falvelos@dps.uminho.pt

Abstract

The exponential growth of traffic demand, and supporting network infrastructures, is leading to serious energy consumption issues. With proper routing, some network links can be put on a sleep mode if demands can be routed through other links and, with multiperiod traffic, the sleeping links can change between periods. In this paper, we address the energy efficient routing problem, and we study the tradeoff between energy savings and percentage of routing paths allowed to change between periods. We present an ILP model defining the energy efficient routing problem and we propose a GRASP based heuristic to compute good routing solutions.

Keywords: energy efficient routing, traffic engineering, optimization algorithms.

1 Introduction

The continuous increase in the number of users and services, which results in exponential traffic growth and in the rapid expansion of more complex network infrastructures, raises several traffic engineering issues such as that of energy consumption. Unfortunately, most network infrastructures lack effective energy management solutions to automatically adapt their energy consumption. We address the problem of optimizing the energy consumption of a capacitated telecommunication network. We consider multiple time periods and single path routing. Single path routing is preferred by many operators because it requires a single Label Switched Path (LSP) for each traffic flow on the Multi Protocol Label Switching (MPLS) networks [Elwalid et al., 2001], keeping the total number of LSPs to a minimum and not jeopardizing the routers' performance by processing heavy routing tables.

The energy efficient routing problem consists in selecting a set of links for each time period, which minimizes the overall energy consumption of the network. The number of flows whose routing paths can change between time periods is bounded. Note that the change of routing paths between time periods requires a management system that can efficiently perform the routing changes in due time and such a system has scalability issues when the number of paths to be changed becomes too large. Moreover, the change of routing paths causes traffic instability, which can result in temporary jitter and packet loss degradation, and such instability is proportional to the number of flows whose routing paths are changed. Therefore, bounding the number of routing paths changing between time periods helps in maintaining a low network management complexity as well as the traffic stability, while still achieving energy savings.

The issue of energy efficient routing has been addressed by some recent works in different contexts. In [Hinton et al., 2011], the authors present techniques to decrease the energy consumption for the internet. In [Bolla et al., 2011], the authors present the idea of turning links into sleep mode dynamically. [Cuomo et al., 2012] and [Francois et al., 2012] also use the idea of turning links into sleep mode. Moreover, [Vasic and Kostic, 2010], [Athanasiou et al., 2011], [Xu et al, 2012] and [Amaldi et al., 2013] address the problem of energy consumption combined with load balancing. [Aldraho and Kist, 2012] addresses the problem of energy consumption and survivability. [Koster et al., 2013] addresses the problem of energy consumption using compressible flows. [Amaldi et al., 2013] addresses the problem of energy consumption for multiple time periods. All these works consider splittable routing.

The problem of energy consumption has also been addressed for single path routing in a few cases. Both works [Gelenbe and Morfopoulou, 2012] and [Gelenbe and Morfopoulou, 2012a] address the problem of energy consumption and other QoS issues using probability models. [Addis et al., 2012a] addresses the problem of energy consumption and multiple time periods, and [Addis et al., 2012a] additionally considers survivability.

The contribution of this paper is the study of the tradeoff between energy savings and the percentage of routing paths allowed to change between time periods. We present an integer linear programming (ILP) model defining the energy efficient routing problem and propose a GRASP based heuristic algorithm for obtaining approximate solutions, since the optimal integer solutions are computationally demanding to obtain. A multi-thread version of the heuristic algorithm is implemented, to further enhance its performance. We show that, in many cases, when only a small percentage of routing paths are allowed to change, we can achieve energy savings close to those where all routing paths are allowed to change. The paper is organized as follows. In Section 2, we present an ILP model defining the energy efficient routing problem. In Section 3, we describe the proposed GRASP based heuristic algorithm. In Section 4, we present computational results showing the tradeoff between energy savings and the percentage of paths that can be changed between periods. Finally, in Section 5 we draw some conclusions.

2 Problem Formulation

In this section, we present a basic formulation for the energy efficient routing problem as an ILP model and, then, some additional constraints that improve its linear programming (LP) lower bound.

2.1 Basic Formulation

Consider a network given by a graph $G = (N, A)$, where N represents the set of network nodes and A represents the set of existing network links. Each edge of A is denoted by $\{i, j\}$ and represents the link between nodes i and j. Alternatively, an element of A can be denoted by arc (i, j) which represents the edge $\{i, j\}$ directed from i to j. For each edge $\{i, j\}$, $c_{\{ij\}}$ represents its bandwidth capacity and $e_{\{ij\}}$ represents its energy consumption. Consider $V(i)$ as the set of neighboring nodes of $i \in N$, i.e., the nodes $j \in N$ such that $(i, j) \in A$.

Consider a set T of consecutive time periods identified by $t = 1, ..., |T|$. Note that T is cyclic so that, in constraints (2.4) of the formulation below, $t = |T| + 1$ actually represents $t = 1$. The parameter p_t represents the percentage of $t \in T$ duration over the total time. The network has to support a set of flows K, where each flow is routed through a single path at each time period. Each flow $k \in K$ is characterized by its origin node o_k, its destination node d_k and its bandwidth demand b_{kt} at each time period t. Let Z be the maximum number of flows whose routing paths can change between time periods.

The energy efficient routing problem consists in selecting the set of links for each time period, which minimizes the overall energy consumption. Consider the following decision variables: x_{ij}^{kt} is a binary variable which indicates if arc (i, j) is in the routing path of flow k at time period t; $\ell_{\{ij\}}^t$ is a binary variable which indicates if edge $\{i, j\}$ is being used at time period t and δ_k is a binary variable which indicates if the routing paths of flow k change between time periods. The problem can be formulated as:

$$\min \sum_{t=1,...,|T|} \sum_{\{i,j\} \in A} p_t \ell_{\{ij\}}^t e_{\{ij\}} \tag{2.1}$$

s.t.

$$\sum_{j \in V(i)} \left(x_{ij}^{kt} - x_{ji}^{kt} \right) = \begin{cases} 1, & i = o_k \\ -1, & i = d_k \\ 0, & i \neq o_k, d_k \end{cases} \qquad i \in N,\ k \in K,\ t = 1, ..., |T| \tag{2.2}$$

$$\sum_{k \in K} b_{kt} \left(x_{ij}^{kt} + x_{ji}^{kt} \right) \leq c_{\{ij\}} \ell_{\{ij\}}^t \qquad \{i, j\} \in A,\ k \in K,\ t = 1, ..., |T| \tag{2.3}$$

$$x_{ij}^{k(t+1)} - x_{ij}^{kt} \leq \delta_k \qquad (i, j) \in A,\ k \in K,\ t = 1, ..., |T| \tag{2.4}$$

$$\sum_{k \in K} \delta_k \leq Z \tag{2.5}$$

$$x_{ij}^{kt} \in \{0, 1\} \qquad (i, j) \in A,\ k \in K,\ t = 1, ..., |T| \tag{2.6}$$

$$\ell_{\{ij\}}^t \in \{0, 1\} \qquad \{i, j\} \in A,\ t = 1, ..., |T| \tag{2.7}$$

$$\delta_k \in \{0, 1\} \qquad k \in K \tag{2.8}$$

The objective function (2.1) aims at selecting the set of links for each time period, which minimizes the overall energy consumption. Constraints (2.2) are the path conservation constraints and guarantee that x_{ij}^{kt} defines a routing path for flow k at time period t. Constraints (2.3) are the capacity constraints and guarantee that $\ell_{\{ij\}}^{t}$ defines the links being used by the flows and that the bandwidth demand in each link does not exceed its capacity. Constraints (2.4) set the variable δ_k to 1 if the routing paths of flow $k \in K$ change between time periods. Constraint (2.5) bounds the number of flows whose routing paths change between time periods to parameter Z. Finally, constraints (2.6-2.8) are the variable domain constraints.

Two special cases can be defined from this general model. One is when we require the routing paths to be the same for all time periods, i.e., $Z = 0$. This case might be preferred by operators since it minimizes the network management complexity (does not require routing path reconfigurations in time period transitions) and, at the same time, elliminates traffic instability. In this case, the problem can be simplified by aggregating variables x_{ij}^{kt} into x_{ij}^{k} and discarding constraints (2.4) and (2.5). This reduces the problem to the following simplified formulation:

$$
\begin{array}{ll}
\min \displaystyle\sum_{t=1,\dots,|T|} \sum_{\{i,j\}\in A} p_t \ell_{\{ij\}} e_{\{ij\}} & \\
\text{s.t.} & \\
\displaystyle\sum_{j\in V(i)} \left(x_{ij}^{k} - x_{ji}^{k}\right) = \begin{cases} 1, & i = o_k \\ -1, & i = d_k \\ 0, & i \neq o_k, d_k \end{cases} & i \in N,\ k \in K \\
\displaystyle\sum_{k\in K} b_{kt} \left(x_{ij}^{k} + x_{ji}^{k}\right) \leq c_{\{ij\}} \ell_{\{ij\}} & \{i,j\} \in A,\ t = 1, \dots, |T| \\
x_{ij}^{k} \in \{0,1\} & (i,j) \in A,\ k \in K \\
\ell_{\{ij\}} \in \{0,1\} & \{i,j\} \in A
\end{array}
$$

The second special case is when $Z = |K|$, i.e., when we let all routing paths change between time periods. This is usually a case that operators want to avoid due to the resulting high network management complexity and network traffic instability. Nevertheless, this case is useful for comparison purposes as a measure of how much energy can be saved at best. In this case, constraints (2.4) and (2.5) become redundant and the global problem becomes a sum of T independent problems, one for each time period $t = 1, \dots, |T|$. Hence for each time period, we have:

$$
\begin{array}{ll}
\min \displaystyle\sum_{\{i,j\}\in A} \ell_{\{ij\}}^{t} e_{\{ij\}} & \\
\text{s.t.} & \\
\displaystyle\sum_{j\in V(i)} \left(x_{ij}^{kt} - x_{ji}^{kt}\right) = \begin{cases} 1, & i = o_k \\ -1, & i = d_k \\ 0, & i \neq o_k, d_k \end{cases} & i \in N,\ k \in K \\
\displaystyle\sum_{k\in K} b_{kt} \left(x_{ij}^{kt} + x_{ji}^{kt}\right) \leq c_{\{ij\}} \ell_{\{ij\}}^{t} & \{i,j\} \in A \\
x_{ij}^{kt} \in \{0,1\} & (i,j) \in A,\ k \in K \\
\ell_{\{ij\}}^{t} \in \{0,1\} & \{i,j\} \in A
\end{array}
$$

Solving the general model, or any of the two special cases, using standard solving techniques (for example, through branch-and-cut algorithms available on CPLEX [IBM, 2009]) for reasonably sized instances requires enormous computational times, as will be illustrated in Section 4.

2.2 Additional Constraints

In this section, we present two well known sets of additional constraints that, added to the previous ILP models, improve the LP lower bound of the basic formulation.

The first set of additional constraints is known as the cut-set constraints, which state that if we partition the network into two sets, the links connecting the two sets must be sufficient to support all traffic demands between them. Consider $V \subset N$ as a subset of network nodes. Consider $A(V)$ as the set of links that have one end node in V and the other end node in $N \backslash V$. Similarly, consider $K(V)$ as the set of flows that have one end node in V and the other end node in $N \backslash V$ and, additionally, $b_t(V)$ as the sum of the flows of $K(V)$ in time period t, i.e.:

$$b_t(V) = \sum_{k \in K(V)} b_{kt}$$

Consider the vector $c(V)$ of the link capacities of $A(V)$ and the vector $[c(V)]$ obtained from $c(V)$ by arranging its elements in non-increasing order. Denote the l^{th} element of $[c(V)]$ as $[c(V)]_l$. For each time period $t = 1, ..., |T|$, let $q_t(V)$ be the smallest integer such that

$$\sum_{l=1}^{q_t(V)} [c(V)]_l \geq b_t(V)$$

Then, for every set V, we can define the following additional set of constraints (one for each time period t):

$$\sum_{\{i,j\} \in A(V)} \ell^t_{\{i,j\}} \geq q_t(V) \quad t = 1, ..., |T|$$

There is an exponential number of such constraints but it is well known that the most efficient ones in improving the LP lower bounds are those involving a small number of links in $A(V)$. Therefore, we have decided to include in the models only the ones such that $|V| = 1$ or $|A(V)| \leq 3$ as long as $b_t(V) > 0$, for time period $t = 1, ..., |T|$. Then, the cut-set constraints added to the models are given by:

$$\sum_{\{i,j\} \in A(V)} \ell^t_{\{i,j\}} \geq q_t(V) \quad t = 1, ..., |T|,\, V \subset N : (|V| = 1 \vee |A(V)| \leq 3) \wedge b_t(V) > 0 \tag{2.9}$$

The second set of additional constraints state that a link $\{i, j\}$ must be in the solution if one of its arcs (i, j) or (j, i) is used by any flow. These additional constraints are formulated as:

$$x^{kt}_{ij} + x^{kt}_{ji} \leq \ell^t_{\{ij\}} \quad \{i, j\} \in A,\, k \in K,\, t = 1, ..., |T| \tag{2.10}$$

Note that the set of constraints (2.10) is very large and it is impossible to know, in advance, which ones effectively improve the LP lower bounds. Moreover, the computational tests showed that adding all these constraints significantly affects the running times. The selection of the constraints (2.10) to be added to the models was done in the following way. We first solve the linear relaxation of the basic formulation with the additional constraints (2.9) and then, iteratively, we add the constraints (2.10) which are not satisfied by the linear relaxation solution. The linear relaxation is recalculated until all constraints (2.10) are satisfied. In this way, only some constraints (2.10) are added (the ones that effectively have improved the LP bound).

3 Heuristic Algorithm

Since the optimal integer solutions are computationally expensive to obtain, we have devised a heuristic algorithm based on GRASP with Path Relinking. Greedy Randomized Adaptive Search Procedure (GRASP) is a multi-start local search method which was first introduced in [Feo and Resende, 1989]. This method consists in generating an initial solution using some greedy randomized procedure. Then, local search is applied to this solution until a local minimum solution is found. This is repeated several times (multi-start) and the best local minimum solution, found during the multiple starts, is saved as the incumbent solution (best solution found so far). Path Relinking (PR) was first proposed in [Glover, 1996] and is a technique that combines two solutions to search for better ones. The usual combination of PR with GRASP is to run PR after each GRASP iteration, combining the local minimum with a solution selected from a list of previously found elite solutions. Note that PR should not be performed with all elite solutions if the elite list is large, since it becomes time consuming and, therefore, may severely jeopardize the algorithm's performance. In our case, we have considered only the incumbent as the elite solution, i.e., PR is only performed between the GRASP local minimum and the incumbent solution.

In general, the heuristic works as follows. We generate initial solutions by starting with an empty network and ordering the demands randomly. Then, for each demand, we select the routing paths (for each time period) that minimize the incremental energy consumption. In the local search, we eliminate one link of the current solution and try to reroute the affected flows through the remaining links, hoping to obtain in this way, a better neighbor solution.

Note that, in the generation of the initial solution, the original GRASP strategy states that each path should be selected among a list of candidate paths that provide the lowest incremental energy consumption. Some preliminary tests have shown that in the energy efficient routing problem, the original strategy is less efficient since it produces worse initial solutions and the randomness provided by considering different demand orders is enough to diversify the search over the solution space.

Consider $\gamma \in [0, 1]$ to be the proportion of flows whose routing paths can change between time periods. Then $Z = \lfloor \gamma |K| \rfloor$, where $\lfloor y \rfloor$ represents the highest integer not larger than y. In order to define, in detail, the proposed heuristic, we consider three distinct cases: $\gamma = 0$, $\gamma = 1$ and $0 < \gamma < 1$.

Let us first consider the case $\gamma = 0$ (i.e., no routing path can change between time periods). Consider the additional variables $\mu^t_{\{ij\}}$ which represent the load of link $\{i, j\} \in A$, for time period $t = 1, ..., |T|$. The initial solutions are computed as follows:

1. set link loads $\mu^t_{\{ij\}} = 0$, for all $\{i, j\} \in A$ and $t = 1, ..., |T|$;
2. generate a random order for the flows $k \in K$;
3. for each flow $k \in K$ according to the previously generated order:
   (a) assign a modified cost $\omega^t_{\{ij\}}$, to each link $\{i, j\} \in A$ and each time period $t = 1, ..., |T|$, given by

$$\omega^t_{\{ij\}} = \begin{cases} 0, & \mu^t_{\{ij\}} > 0 \text{ and } \mu^t_{\{ij\}} + b_{kt} \le c_{\{ij\}} \\ p_t\, e_{\{ij\}}, & \mu^t_{\{ij\}} = 0 \text{ and } b_{kt} \le c_{\{ij\}} \\ +\infty, & \mu^t_{\{ij\}} + b_{kt} > c_{\{ij\}} \end{cases} \tag{3.1}$$

   and determine for each link $\{i, j\} \in A$, a link cost $\omega_{\{ij\}}$ given by

$$\omega_{\{ij\}} = \sum_{t=1}^{|T|} \omega^t_{\{ij\}} \tag{3.2}$$

   (b) using the costs $\omega_{\{ij\}}$, determine the shortest path for flow k (from o_k to d_k) and assign this path as the routing path of k for all time periods $t = 1, ..., |T|$;
   (c) for all $\{i, j\}$ belonging to the selected routing path, and for each $t = 1, ..., |T|$, add the bandwidth demand b_{kt} to the load $\mu^t_{\{ij\}}$.

After the initial solution has been computed, a local search procedure is applied to it. We consider a neighbor solution of the current one, as a solution where we select one of its links. The neighbor solution is computed as follows. First, we remove all flows currently routed through the link. Then, for all of these flows: (i) the shortest path is recalculated using the link costs as defined in (3.2), and (ii) the flow which has the lowest shortest path is routed through it. This is done for all remaining flows, until all of them are routed through the network. There are as many neighbor solutions as the number of links of the current solution.

Let us now consider the case $\gamma = 1$ (i.e., all routing paths can change between time periods). In this case, the initial solution is computed as follows:

For each time period $t = 1, ..., |T|$:

1. set link loads $\mu^t_{\{ij\}} = 0$, for all $\{i, j\} \in A$;
2. generate a random order for the flows $k \in K$;
3. for each flow $k \in K$ according to the previously generated order:
   (a) assign a modified cost $\omega^t_{\{ij\}}$, to each link $\{i, j\} \in A$, given by (3.1);
   (b) using the costs $\omega^t_{\{ij\}}$, determine the shortest path for flow k (from o_k to d_k) and assign this path as the routing path of k, for time period t;
   (c) for all $\{i, j\}$ belonging to the selected routing path, add the bandwidth demand b_{kt} to the load $\mu^t_{\{ij\}}$.

Note that for $\gamma = 1$, each time period is treated independently. Hence, in step 3(b), a shortest path is now determined for each flow $k \in K$ and for each time period $t = 1, ..., T$. In the same way, the local search described for $\gamma = 0$, is now applied to each time period independently.

Finally let us consider the general case $0 < \gamma < 1$ (i.e., when a percentage γ of flows can change routing paths between time periods). The aim is to find good solutions for low values of γ. So, in the general case, the algorithm for $\gamma = 0$ is a natural starting point. On each GRASP iteration, we first run both the initial solution and the local search procedures as for the case of $\gamma = 0$ and, then, we run a local search procedure, considering separately each of the time periods and discarding the neighbor solutions that violate parameter Z.

PR is included in the algorithm only for the first two cases ($\gamma = 0$ and $\gamma = 1$). In these cases, all the PR solutions (which are combinations of the starting solutions) are valid and, in fact, the computational results show that most of the incumbent solutions are provided by PR, improving significantly the algorithm's efficiency. In contrast, most of the PR solutions are invalid for $0 < \gamma < 1$ and, therefore, PR is of no value in this case since it only wastes computational resources.

4 Computational Results

To test the ILP models and the heuristic algorithms, we have defined a set of 12 instances based on the well known topology of the NSF network, as shown in Figure 1. In the first six instances (A to F), all links have a capacity of 1 Gbps (= 10^3 Mbps). In the next six instances (A' to F'), most of the links have a capacity of 1 Gbps (= 10^3 Mbps) and some links have a capacity 10 Gbps (= 10^4 Mbps). These links with higher capacity are highlighted in bold in Figure 1. In all cases, we have assumed that the energy consumption of a link is proportional to its bandwidth capacity.

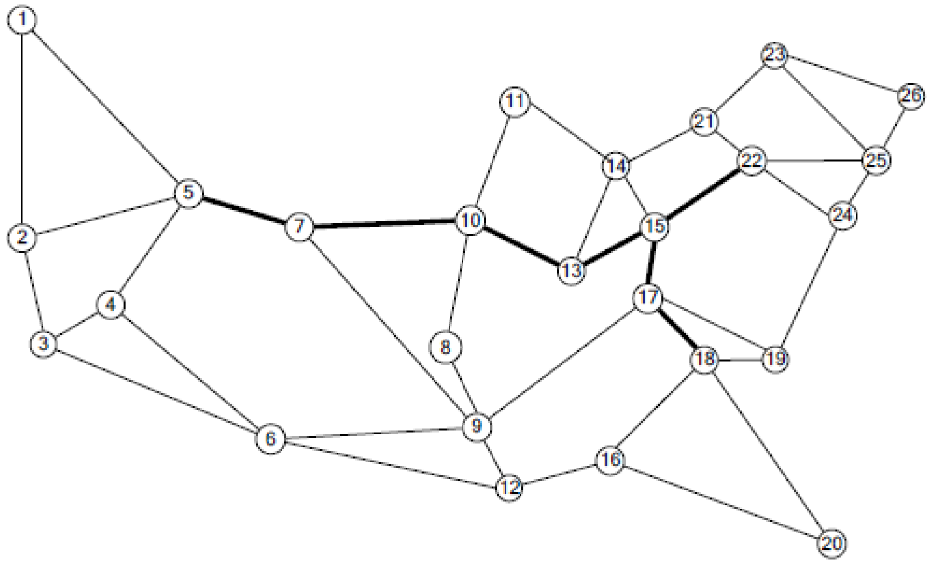


Figure 1: NSF network with 26 nodes and 42 links.

For each instance, a set $S \subset N$ of 20 traffic nodes was randomly chosen (different between instances) to model the general case where some network nodes are only transit nodes (i.e., no flow demands are originated from or destinated to them). We consider a set T of three time periods along a 24-hour period associated to the typical Morning, Afternoon and Night daily periods: $t = 1$ from 8h00 to 12h00 ($p_1 = 4/24$); $t = 2$ from 12h00 to 18h00 ($p_2 = 6/24$) and $t = 3$ from 18h00 to 8h00 ($p_3 = 14/24$).

For each time period, we have considered the existence flow demands between all pairs of traffic nodes (resulting in a total of 190 flow demands per time period) whose values were generated in the following way. A subset $W \subset S$ is randomly selected with $|W|$ nodes. The nodes belonging to W are assigned a weight of 3, while the other nodes of S are assigned a weight of 1. The flow demands are then randomly generated between a minimum value v_m and a maximum value v_M (both in Mbps), and multiplied by the weight of their end nodes (origin and destination nodes). The parameters used for each problem instance are shown in Table 1. Note that when $|W| > 0$, we assume that there exists a set of nodes that serve a larger set of clients and, therefore, the traffic demands involving these nodes are higher on average (which is accounted for by the weight of 3).

In order to assert the performance of the additional constraints on the linear relaxation of the formulations, we show in Table 2 the LP lower bounds obtained and the respective CPU times (in seconds), for instances A and A' and different values of γ, obtained using the CPLEX 12.1 software package (instances B-F behave similarly to A and instances B'-F' behave similarly to A'). Note that for the cases $\gamma = 0$ and $\gamma = 1$, we have used the simplified formulations, as described in Section 2.1. Table 2 shows the results for

Instance	v_m	v_M	$\|W\|$
A,B	0	40	0
C,D	10	30	0
E,F	0	20	6
A',B'	0	60	0
C',D'	20	40	0
E',F'	0	30	6

Table 1: Parameters for generating the flow demands.

(i) the basic formulation with no additional constraints, (ii) the formulation with constraints (2.9) and (iii) the formulation with constraints (2.9) and (2.10) iteratively added as described in Section 2.2.

		Basic		(2.9) added		(2.9)+(2.10) added	
Instance	γ	LP	time (s)	LP	time (s)	LP	time (s)
A	0.00	12.4669	27.44	15.2094	25.06	17.3939	564.30
	0.05	12.4668	23.50	15.1405	33.08	17.2174	752.80
	0.10	12.4668	22.09	15.1401	24.09	17.1734	614.95
	1.00	12.4668	1.38	15.1401	1.41	17.1610	44.85
A'	0.00	19.1737	25.73	20.7296	47.81	31.1375	1259.11
	0.05	19.1620	20.08	20.6497	49.89	30.4798	1156.98
	0.10	19.1620	17.36	20.6279	31.28	30.2654	918.09
	1.00	19.1620	1.17	20.6215	2.17	30.1356	77.42

Table 2: Linear relaxation results for instances A and A'.

These results show that constraints (2.9) improve significantly the LP lower bounds for instance A, while the improvement is not as significant for instance A'. Note that these constraints did not penalize too much the CPU times for both instances and, therefore, are always worthwhile to be used. On the other hand, constraints (2.10) strongly penalize the CPU times for both instances and the obtained LP improvement is much more significant in instance A' than instance A. Therefore, we have assumed that constraints (2.10) are only worthwhile to be used in instance A'.

Taking into account these observations, we have attempted to solve the integer problems for instances A and A' with a time limit of 24 hours (for instance A, we have only used constraints (2.9), while for instance A', we have used both constraints (2.9) and (2.10)). None of the instances ended after 24 hours. The results are shown in Table 3, where the upper bounds (UB) and lower bounds (LB) and the obtained gaps (in percentage) are presented. We note that the instances are computationally expensive to solve to optimality (even with the introduced additional constraints that improved significanty the initial LP values). The best found integer solutions are upper bounds of the optimal solutions and, as shown in Table 3, the gaps are still huge after 24 hours of running time.

Instance	γ	UB	LB	Gap (%)
A	0.00	25.000	18.335	36.35%
	0.05	24.500	18.178	34.78%
	0.01	24.250	17.994	34.77%
	1.00	22.167	18.943	17.02%
A'	0.00	55.000	50.885	8.09%
	0.05	75.663	49.025	54.34%
	0.01	56.000	48.777	14.81%
	1.00	52.333	46.369	12.86%

Table 3: Integer solution results for instances A and A' (CPU time limit of 24h).

Alternatively to obtain approximate solutions, the GRASP based algorithm was implemented in C++ programming language. All tests have been carried out on a standard PC with an Intel i7 CPU processor and 8GB of RAM. To make use of the available processing capabilities of the CPU, a simple multi-thread version of the algorithm was implemented where 8 threads are launched in parallel with the same CPU time limit and, at the end, the best out of the 8 incumbents is the final solution.

We have run our algorithm for all instances considering the values γ= 0.00, 0.05, 0.10 and 1.00. Since the algorithm is a stochastic process, different runs generally give different solutions. Therefore, we have run the algorithm 10 times for each case.

We have run the tests in several steps. First, we have run the algorithm with a time limit of 5 minutes. Then, we have identified the cases where the best solution was found in less than 8 runs (out of 10). For such cases, we have repeated the tests with a time limit of 15 minutes. Then, we have again identified the cases where the best solution was found in less than 8 runs (out of 10). Finally, for such cases, we have repeated the tests with a time limit of 30 minutes. The results are summarized in Table 4. For each case, column 'Value' shows the best solution, column '#' shows the number of runs (out of 10) that found the best solution and the column 't' shows the largest used CPU time limit (in minutes).

	$\gamma = 0.00$			$\gamma = 0.05$			$\gamma = 0.10$			$\gamma = 1.00$		
Instance	Value	#	t (m)	Value	#	t (m)	Value	#	t (m)	Value	#	t (m)
A	24.000	10	5	22.500	2	30	22.500	1	30	21.917	3	30
B	25.000	10	5	25.000	10	5	24.750	9	15	24.083	10	5
C	23.000	10	5	22.167	10	5	22.167	10	5	22.000	8	30
D	24.000	10	5	22.334	1	30	22.334	2	30	22.167	3	30
E	29.000	9	5	28.000	4	30	27.917	1	30	27.417	10	5
F	26.000	1	30	26.000	2	30	26.000	2	30	25.417	10	5
A'	55.000	10	5	54.750	1	30	54.750	7	30	52.750	3	30
B'	58.000	8	15	57.000	1	30	57.000	3	30	51.753	3	30
C'	64.000	9	5	64.000	9	5	64.000	9	5	60.250	1	30
D'	56.000	3	30	48.000	3	30	48.000	2	30	47.167	1	30
E'	78.000	1	30	77.000	1	30	76.667	1	30	76.167	5	30
F'	72.000	10	5	71.000	8	15	71.000	10	15	71.000	10	15

Table 4: Results of the GRASP based heuristic algorithm.

We observe that the case $\gamma = 0$ is slightly easier to solve, since it has more instances where the best solutions were found using only 5 minutes of CPU time. This is without surprise since this case involves the determination of a single path for all time periods, which involves a smaller number of variables. We observe also that the instances A to F are also slightly easier to solve than the instances A' to F', i.e., solving the energy efficient routing problem for networks with links of different capacities is more difficult than when all links have the same capacity.

As expected, the energy consumption values drop as γ increases. To better understand this trade-off, we present in Table 5 the energy penalties obtained by constraining the number of paths that change between periods. These penalties are based on the values of the best solutions presented in Table 4 and using the values of $\gamma = 1$ as the lowest possible energy consumption values (since, in this case, all routing paths can change between time periods).

The results shown in Table 5 exhibit all types of cases. There are cases where letting only 5% of the demands to have different routing paths between time periods ($\gamma = 0.05$) is as good (or almost as good) as the lowest energy consumption values (the cases C, D and F') while in others, letting 10% of the demands to have different routing paths between time periods ($\gamma = 0.10$) heavily penalizes the energy consumption (cases B' and C'). Moreover, there are cases where requiring no routing changes between time periods ($\gamma = 0$) does not impose too much penalty when compared to the lowest energy consumption values (the cases F, E' and F') while in others, the energy penalty is considerable (cases A, D, B' and D'). On the other hand, the results show that the energy consumption gains are higher in networks with links of different capacities.

Note that all case studies involve a total of 190 demands, where only 9 demands (when $\gamma = 0.05$) or 19 demands (when $\gamma = 0.10$) can have routing paths change between time periods. The average penalties (among the 12 problem instances) were 2.80% for $\gamma = 0.10$, 2.95% for $\gamma = 0.05$ and 6.61% for $\gamma = 0$, clearly showing that energy consumption savings are significant even with a small number of routing path changes.

As a final remark, comparing the values of the solutions of Table 4 with the lower bounds of Table 3, for instances A and A', we easily see that the gaps are still very high. Moreover, the best solutions found by the heuristic (in at most 30 minutes) is better for all cases, except one, than the integer solutions

Instance	$\gamma = 0.00$	$\gamma = 0.05$	$\gamma = 0.10$
A	9.50%	2.66%	2.66%
B	3.81%	3.81%	2.77%
C	4.55%	0.76%	0.76%
D	8.27%	0.75%	0.75%
E	5.77%	2.13%	1.82%
F	2.29%	2.29%	2.29%
A'	4.27%	3.79%	3.79%
B'	12.07%	10.14%	10.14%
C'	6.22%	6.22%	6.22%
D'	18.73%	1.77%	1.77%
E'	2.41%	1.09%	0.66%
F'	1.41%	0.00%	0.00%

Table 5: Energy penalty obtained by constraining the number of paths that change between time periods.

found by CPLEX in 24 hours (shown in Table 3). Therefore, work is still required on alternative methods to find good mathematical lower bounds that can be used to assess the quality of the solutions found by the proposed heuristic.

5 Conclusions

In this paper, we have addressed the problem of optimizing the energy consumption of a capacitated telecommunication network. We have considered different time periods and single path routing of the flows over the network. We presented an integer linear programming model for the energy efficient routing problem where the number of flows whose routing paths can change between time periods is bounded by a given parameter. Since the optimal integer solutions are computationally demanding to obtain, a GRASP based heuristic algorithm was proposed and a multi-thread version of the heuristic algorithm was implemented.

In the computational results, we have shown that this problem is very hard to be solved to optimality. Moreover, there are cases where the heuristic has computed good solutions quite efficiently but there are still other cases that require long computing times to obtain good solutions.

The possibility of bounding the number of routing paths that can change between time periods is a way to keep both network management complexity and network traffic instability low. The computational results show that the energy savings depend on each particular case but, on average, energy savings are significant even with a small number of routing path changes.

Acknowledgments

This work has been conducted under the project PTDC/EIA-EIA/100645/2008 "SearchCol: Metaheuristic Search by Column generation" (funded by FCT). Dorabella Santos was funded by Portuguese FCT under the post-doc grant SFRH/BPD/41581/2007.

References

[Addis et al., 2012] Addis, B., Capone, A., Carello, G., Gianoli, L. G., and Sanso, B. (2012). Energy aware management of resilient networks with shared protection. In *In Sustainable Internet and ICT for Sustainability (SustainIT)*, 1–9.

[Addis et al., 2012a] Addis, B., Capone, A., Carello, G., Gianoli, L. G., and Sanso, B. (2012). Energy-aware multiperiod traffic engineering with flow-based routing. In *IEEE International Conference on Communications (ICC'12)*, 5957–5961.

[Aldraho and Kist, 2012] Aldraho, A., and Kist, A. A. (2012). Enabling energy efficient and resilient networks using dynamic topologies. In *In Sustainable Internet and ICT for Sustainability (SustainIT)*, 1–8.

[Amaldi et al., 2013] Amaldi, E., Capone, A., and Gianoli, L. G. (2013). Energy-aware IP traffic engineering with shortest path routing. In *Computer Networks*, available at http://dx.doi.org/10.1016/j.comnet.2013.02.006.

[Athanasiou et al., 2011] Athanasiou, G., Tsagkaris, K., Vlacheas, P. and Demestichas, P. (2011). Introducing energy-awareness in traffic engineering for future networks. In *CNSM'11 Proceedings of the 7th Internation Conference on Network and Services Management*, 367–370.

[Bolla et al., 2011] Bolla, R., Bruschi, R., Cianfrani, A., and Listanti, M. (2011). Enabling backbone networks to sleep. In *IEEE Network Magazine, Special Issue on Energy-Efficient Networks*, 25(2):26–31.

[Cuomo et al., 2012] Cuomo, F., Cianfrani, A., Polverini, M., and Mangione, D. (2012). Network pruning for energy saving in the Internet. In *Computer Networks*, 56(10):2355–2367.

[Elwalid et al., 2001] Elwalid, A., Jin, C., Low, S., and Widjaja, I. (2001). MATE: MPLS Adaptive Traffic Engineering. *IEEE INFOCOM*, 1300–1309.

[Feo and Resende, 1989] Feo, T., and Resende, M. (1989). A probabilisitic heuristic for a computationally difficult set covering problem. *Operations Research Letters*, 8:67–71.

[Francois et al., 2012] Francois, F., Wang, N., Moessner, K., and Georgoulas, S. (2012). Optimization for time-driven link sleeping reconfigurations in ISP backbone networks. *In Network Operations and Management Symposium (NOMS)*, 221–228.

[Gelenbe and Morfopoulou, 2012] Gelenbe, E., and Morfopoulou, C. (2012). Power savings in packet networks via optimised routing. *Mobile Networks and Applications*, 17(1):152–159.

[Gelenbe and Morfopoulou, 2012a] Gelenbe, E., and Morfopoulou, C. (2012). Gradient optimisation for network power consumption. *Green Communications and Networking*, 125–134.

[Glover, 1996] Glover, F. (1996). Tabu search and adaptive memory programming - advances, applications and challenges. *Interfaces in Computer Science and Operations Research*, 1–75.

[Hinton et al., 2011] Hinton, K., Baliga, J., Feng, M., Ayre, R., and Tucker, R. S. (2011). Power consumption and energy efficiency in the internet. In *IEEE Network Magazine, Special Issue on Energy-Efficient Networks*, 25(2):6–12.

[IBM, 2009] (IBM 2009). IBM ILOG CPLEX V12.1, User's Manual for CPLEX. available at ftp://public.dhe.ibm.com/software/websphere/ilog/docs/optimization/cplex/ps_usrmancplex.pdf.

[Koster et al., 2013] Koster, A., Phan T. K., and Tieves, M. (2013). Extended cutset inequalities for the network power consumption problem. *to appear in International Network Optimization Conference (INOC 2013).*

[Vasic and Kostic, 2010] Vasic, N., and Kostic, D. (2010). Energy-aware traffic engineering. *e-Energy'10 Proceedings of the 1st International Conference on Energy-Efficient Computing anf Networking.*

[Xu et al, 2012] Xu, M., Shang, Y., Li, D., and Wang, X. (2012). Greening Data Center Networks with Throughput-guaranteed Power-aware Routing. *Computer Networks*, available at http://dx.doi.org/10.1016/j.comnet.2012.12.012.

Método Previsor Corretor Primal Dual de Pontos Interiores Aplicado a um Problema de Despacho Econômico com Restrições Ambientais

Amélia de Lorena Stanzani[1], Antonio Roberto Balbo[2]

[1] *Programa de Pós-graduação em Engenharia Produção, UFSCAR - São Carlos-SP, Brasil*
E-mail: `mel.stanzani@hotmail.com`

[2] *Depto de Matemática, FC/UNESP - Bauru-SP, Brasil*
E-mail: `arbalbo@fc.unesp.br`

Resumo

O presente trabalho apresenta o desenvolvimento de um método previsor corretor primal dual de pontos interiores com procedimento de barreira modificada e sua aplicação ao problema de despacho econômico condicionado a restrições de máxima emissão de poluentes por unidade geradora. O procedimento de barreira modificada incorporado ao método expande a região viável do problema e propicia trabalhar com pontos inicialmente infactíveis. Testa-se uma implementação computacional deste método, para a determinação de soluções aproximadas de um problema teste com 40 unidades geradoras. Os resultados obtidos demonstram a eficiência do método desenvolvido, em relação a resultados numéricos e ao tempo computacional.

Palavras chave: Método Previsor Corretor Primal Dual de Pontos Interiores, Procedimeto de Barreira Modificada, Problemas de Despacho Econômico, Problema de Despacho Ambiental, Método ε-restrito.

1 Introdução

Neste trabalho será desenvolvido um método primal-dual de pontos interiores para um problema de otimização quadrática convexa, com restrições de igualdade e desigualdade, restrições quadráticas e variáveis canalizadas, que será aplicado em problemas de despacho econômico (PDE) e ambiental (PDA). Estes são encontrados na área de sistemas de geração de energia, em engenharia elétrica, os quais analisam a geração termoelétrica baseando-se em seus aspectos econômicos (PDE), e na preocupação da redução da emissão de poluentes (PDA). O PDE é um problema de otimização não-linear que busca otimizar o processo de alocação ótima da demanda de energia elétrica entre as unidades geradoras disponíveis minimizando o custo de combustíveis empregados na geração termoelétrica, de tal forma que, as restrições operacionais sejam satisfeitas. O PDA é um problema de otimização não-linear que busca minimizar a emissão de poluentes provenientes da geração termoelétrica através da queima de combustíveis fósseis, respeitando também as restrições de demanda de energia e de limitantes operacionais dos geradores. Apresentaremos um modelo baseado no PDE e no PDA definido como um problema de otimização multiobjetivo, em que deseja-se otimizar os custos de geração de energia e concomitantemente reduzir a emissão de poluentes, os quais são objetivos conflitantes. Para a análise ε-restrito, considerando nesta abordagem a minimização do problema econômico condicionada à restrições de máxima emissão de poluentes permissível para cada unidade geradora, relativas à função ambiental. O modelo investigado de despacho econômico com restrições ambientais, obtido através do método ε-restrito, possibilita a análise de técnicas de solução para a resolução de problemas de otimização não linear, com restrições não lineares. Baseando-se em métodos previsor-corretor primal-dual de pontos interiores aplicados à resolução de problemas de otimização não-linear, apresentados em [Wu et al., 1994] e [Mehrotra e Sun, 1992], desenvolvemos um método primal-dual de pontos interiores explorando a estratégia da função barreira logarítmica modificada definida em [Polyak, 1992], a qual possibilita inicializar o problema com pontos infactíveis, e o procedimento iterativo do método opera com pontos interiores à região relaxada e exteriores à região original, enquanto busca atender a viabilidade da sequencia de pontos obtidos, de tal forma que, na otimalidade a solução é factível e interior à região original. A determinação das direções de busca do método é realizada através de um procedimento previsor-corretor, variante daquele apresentado em [Mehrotra e Sun, 1992], que considera o procedimento de centragem relativo ao parâmetro de barreira no passo previsor para o cálculo das direções de busca deste passo, utilizando-as imediatamente, numa mesma iteração, no passo corretor, quando considera-se os aproximantes de segunda ordem do sistema de direções de busca relativos às condições de complementaridade. O algoritmo do método previsor-corretor primal-dual barreira logarítmica modificada (PCPDBLM), foi implementado em linguagem computacional C++ para a determinação de soluções aproximadas do problema de despacho econômico com restrições ambientais. Esta investigação demonstra a eficiência do PCPDBLM, em relação ao tempo computacional, número de iterações e comodidade na escolha de uma solução inicial, já que esta pode ser infactível.

2 Problemas de Despacho

2.1 Problema de Despacho Econômico

O problema despacho econômico, apresentado em [Steinberg e Smith, 1943] e [Gent e Lamont, 1971], é um problema modelado em otimização não-linear, que busca otimizar o processo de alocação ótima da demanda de energia elétrica entre as unidades geradoras disponíveis, minimizando o custo de combustíveis empregados na geração termoelétrica, respeitando as restrições de demanda e operacionais das unidades geradoras. O modelo do problema despacho econômico (PDE) será apresentado em (2.1):

$$\begin{gathered}\text{Minimizar } F_e = \textstyle\sum_{i=1}^{n} (F_e)_i P_i = (a_e)_i P_i^2 + (b_e)_i P_i + (c_e)_i \\ \text{Sujeito a } \begin{cases} \sum_{i=1}^{n} (P)_i = P_d \\ P_i^{min} \leq P_i \leq P_i^{max} \end{cases}\end{gathered} \tag{2.1}$$

onde:
F_e: função custo de geração total;
$(F_e)_i P_i$: custos de cada unidade geradora i, sem considerar o efeito do ponto de válvula;
n: número total de unidades geradoras do sistema;
P_i: potência na qual cada unidade geradora i deve operar;
$(a_e)_i, (b_e)_i, (c_e)_i$: coeficientes da função custo por unidade geradora i;
P_d: potência demandada total;
P_i^{min}, P_i^{max}: limites operacionais inferiores e superiores, respectivamente, de saída de cada unidade geradora i.

2.2 Problema de Despacho Ambiental

Durante muito tempo, a geração termoelétrica considerou apenas estratégias econômicas como base de operação, contribuindo, assim, para a elevação da poluição atmosférica. Cada quilowatt de eletricidade produzida a cada hora pode estar associado a taxas de emissões através de um fator de emissão. Atualmente, as regulamentações têm concentrado esforços na redução das taxas de emissão de particulados, dióxido de enxofre ($S0_2$) e óxidos de nitrogênio (NOx), produzidos através da queima dos combustíveis fósseis nas usinas termoelétricas, os quais podem ocasionar sérios problemas à saúde devido a sua inalação pelas pessoas. O modelo geral de otimização para o problema de despacho ambiental (PDA), proposto em [El-Hawary et al., 1992] visa minimizar a função emissão F_a, sujeito às mesmas restrições impostas ao PDE, apresentado a seguir, no problema (2.2):

$$\begin{gathered}\text{Minimizar } F_a = \textstyle\sum_{i=1}^{n} (F_a)_i P_i = (a_a)_i P_i^2 + (b_a)_i P_i + (c_a)_i \\ \text{Sujeito a } \begin{cases} \sum_{i=1}^{n} (P)_i = P_d \\ P_i^{min} \leq P_i \leq P_i^{max} \end{cases}\end{gathered} \tag{2.2}$$

onde:
F_a: função emissão total;
$(F_a)_i P_i$: emissão de cada unidade geradora i;
$(a_a)_i, (b_a)_i, (c_a)_i$: coeficientes da função emissão por unidade geradora i.

2.3 Problema Multiobjetivo de Despacho Econômico e Ambiental

O Problema Multiobjetivo de Despacho Econômico e Ambiental (PMDEA) é proposto de tal forma a minimizar concomitantemente o custo de geração e a emissão de poluentes, de acordo com o problema (2.3) modelado a seguir:

$$\text{Minimizar } \{F_e, F_a\}$$

$$\text{Sujeito a } \begin{cases} \sum_{i=1}^{n}(P)_i = P_d \\ P_i^{min} \leq P_i \leq P_i^{max} \end{cases} \tag{2.3}$$

Os objetivos econômico e ambiental são conflitantes, portanto será necessário a utilização de uma estratégia que reformule o problema multiobjetivo transformando-o em um problema mono-objetivo. Para tanto, apresentaremos a seguir o método ε-restrito.

2.3.1 Problema de Despacho Econômico com Restrições Ambientais - Método ε-restrito

A resolução de problemas multiobjetivo que apresentam objetivos conflitantes, ou seja, que não podem ser minimizados (ou maximizados) ao mesmo tempo, necessitam de estratégias de resolução, que consistem em reformular o problema através de um problema mono-objetivo. A estratégia ε-restrito, proposta em [Haimes et al., 1971] sugere manter um dos objetivos do problema e considerar o outro objetivo como restrição, limitando-o superiormente por um valor constante. Através do método ε-restrito, reformula-se o PMDEA, mantendo a função econômica como objetivo de minimização e tratando a função ambiental como restrição do problema. Essa formulação mono-objetivo será apresentada a seguir em (2.4). A abordagem do presente trabalho considera as funções $(F_a)_i$ como restrições do PDE, limitadas superiormente por constantes que representam a máxima emissão de poluentes permissível por unidade geradora i:

$$\text{Minimizar } F_e = \sum_{i=1}^{n}(F_e)_i P_i = (a_e)_i P_i^2 + (b_e)_i P_i + (c_e)_i$$

$$\text{Sujeito a } \begin{cases} \sum_{i=1}^{n}(P)_i = P_d \\ P_i^{min} \leq P_i \leq P_i^{max} \\ (F_a)_i P_i \leq {F_a}^{max}{}_i \end{cases} \tag{2.4}$$

onde:

${F_a}^{max}{}_i$: máxima emissão por unidade geradora i.

O PDE clássico com restrições lineares e variáveis canalizadas, quando condicionado à restrições ambientais, passa a ter restrições quadráticas. Para resolvê-lo será desenvolvido um método previsor-corretor primal-dual de pontos interiores com estratégia de barreira modificada para facilitar a escolha de valores iniciais, que não necessariamente precisam satisfazer as restrições quadráticas.

3 Método Previsor Corretor Primal Dual de Pontos Interiores

Problemas de otimização não-linear ocorrem em muitas áreas do conhecimento tais como, matemática, engenharia, agronomia, entre outras. Devido à sua não-linearidade e quantidade de variáveis, estes problemas apresentam dificuldades à sua resolução, assim é de interesse utilizar métodos numéricos eficientes para a determinação de soluções destes, definindo critérios de convergência para tais determinações. Dentre estes métodos, podemos citar o método previsor-corretor primal-dual (PCPD) de pontos interiores, investigado neste trabalho, que utiliza a estratégia de barreira logarítmica modificada e que na prática, têm se mostrado eficientes para a resolução de problemas de programação não-linear. A estratégia de barreira modificada definida em [Polyak, 1992] consiste em aumentar a região factível do problema, possibilitando assim a utilização de pontos infactíveis no desenvolvimento do método PCPD, até que este atinja um ponto interior, factível. Definiremos a seguir um problema de minimização em sua forma geral:

$$\text{Minimizar } f(x)$$

$$\text{Sujeito a } \begin{cases} g(x) = 0 \\ h(x) \leq u \\ x^{min} \leq x \leq x^{max} \end{cases} \tag{3.1}$$
$$x \in \Re^n, g : \Re^n \to \Re^m, h : \Re^n \to \Re^r$$

O problema (3.1) com restrições de desigualdade, serão adicionadas variáveis de folga e de excesso, obtendo-se o problema equivalente 3.2:

$$\text{Minimizar } f(x)$$
$$\text{Sujeito a } \begin{cases} g(x) = 0 \\ h(x) + z_1 - u = 0 \\ x + z_2 - x^{max} = 0 \\ z_3 - x + x^{min} = 0 \\ z_1, z_2, z_3 > 0 \end{cases} \tag{3.2}$$
$$x, z_2, z_3 \in \Re^n, z_1 \in \Re^r, g : \Re^n \to \Re^m, h : \Re^n \to \Re^r$$

Uma relaxação em relação à restrição $h(x)+z_1-u=0$ do problema acima é proposta adicionando-se o parâmetro de barreira μ na restrição $z_1 > 0$, a qual é expressa por:

$$z_1 > -\mu \tag{3.3}$$

dessa maneira a região factível é expandida. Dessa forma, temos uma ampliação da região factível relativa à equação $h(x)+z_1-u=0$, o que permite ao método, nesta restrição, operar com pontos iniciais infactíveis. A restrição $z_1 > -\mu$ será representada pela restrição $z_1^{\diamondsuit} > 0$, onde:

$$(z_1^{\diamondsuit})_i = 1 + \mu^{-1}(z_1)_i \tag{3.4}$$

Redefinimos o problema de tal forma a garantir a positividade das variáveis $z_1^{\diamondsuit}, z_2, z_3$, através da função barreira logarítmica. A inclusão da restrição $z_1^{\diamondsuit} > 0$ implica, de acordo com [Polyak, 1992], na utilização da função barreira modificada que possibilita operar o método com pontos inicialmente infactíveis.

$$\text{Minimizar } f(x) - \mu \textstyle\sum_{j=1}^{r} \delta_j ln(z_1^{\diamondsuit})_j - \mu \sum_{j=1}^{n} ln(z_2)_j - \mu \sum_{j=1}^{n} ln(z_3)_j$$
$$\text{Sujeito a } \begin{cases} g(x) = 0 \\ h(x) + z_1 - u = 0 \\ x + z_2 - x^{max} = 0 \\ z_3 - x + x^{min} = 0 \end{cases} \tag{3.5}$$
$$x, z_2, z_3 \in \Re^n, z_1^{\diamondsuit} \in \Re^r, g : \Re^n \to \Re^m, h : \Re^n \to \Re^r$$

O método PCPDBLM será aplicado a um problema equivalente e irrestrito, expresso pela função lagrangiana barreira logarítmica modificada definida da seguinte forma:

$$\begin{gathered} \text{Minimizar } L_{LM} = f(x) - \mu \textstyle\sum_{j=1}^{r} \delta_j ln(z_1^{\diamondsuit})_j + \\ -\mu \textstyle\sum_{j=1}^{n} ln(z_2)_j - \mu \sum_{j=1}^{n} ln(z_3)_j + \\ + \textstyle\sum_{j=1}^{m} (\lambda_0)_j (g(x))_j + \sum_{j=1}^{r} (\lambda_1)_j (h(x) + z_1 - u)_j + \\ + \textstyle\sum_{j=1}^{n} (\lambda_2)_j (x + z_2 - x^{max})_j + \sum_{j=1}^{n} (\lambda_3)_j (z_3 - x + x^{min})_j \end{gathered} \tag{3.6}$$

A partir do problema irrestrito será desenvolvido o algoritmo do método primal-dual de pontos interiores, com procedimento previsor-corretor e com barreira logarítmica modificada. Aplicando as condições necessárias de otimalidade de KKT,$\nabla L_{LM}(x, z_1, z_2, z_3, \lambda_0, \lambda_1, \lambda_2, \lambda_3)$, temos o seguinte sistema de equações:

$$\begin{gathered} \nabla L_{LM(x)} = \nabla f(x) + \nabla g(x)^t \lambda_0 + \nabla h(x)^t \lambda_1 + \lambda_2 - \lambda_3 = 0 \\ \nabla L_{LM(\lambda_0)} = g(x) = 0 \\ \nabla L_{LM(\lambda_1)} = h(x) + z_1 - u = 0 \\ \nabla L_{LM(\lambda_2)} = x + z_2 - x^{max} = 0 \\ \nabla L_{LM(\lambda_3)} = z_3 - x + x^{min} = 0 \\ \nabla L_{LM(z_1)} = \mu (Z_1^{\diamondsuit})^{-1} e_1 - I\lambda_1 \\ \nabla L_{LM(z_2)} = \mu (Z_2)^{-1} e_2 - I\lambda_2 \\ \nabla L_{LM(z_3)} = \mu (Z_3)^{-1} e_2 - I\lambda_3 \end{gathered} \tag{3.7}$$

Se o ponto $(x^k, {z_1}^k, {z_2}^k, {z_3}^k, {\lambda_0}^k, {\lambda_1}^k, {\lambda_2}^k, {\lambda_3}^k)$ de uma iteração corrente k não satisfaz todas as equações dadas em (3.7), ou seja, $\nabla L_{LM}(x^k, {z_1}^k, {z_2}^k, {z_3}^k, {\lambda_0}^k, {\lambda_1}^k, {\lambda_2}^k, {\lambda_3}^k) \neq 0$, então são gerados os resíduos $(m^k, {t_0}^k, {t_1}^k, {t_2}^k, {t_3}^k, {\pi_1}^k, {\pi_2}^k, {\pi_3}^k)$. Esses resíduos serão utilizados no critério de parada do método PCPDBLM. A definição de um novo ponto depende diretamente das direções de movimento e comprimento de passo nesta direção, sendo definido através de:

$$
\begin{aligned}
x^{k+1} &= x^k + \alpha_p {d_x}^k \\
{z_1}^{k+1} &= {z_1}^k + \alpha_p {d_{z_1}}^k \\
{z_2}^{k+1} &= {z_2}^k + \alpha_p {d_{z_2}}^k \\
{z_3}^{k+1} &= {z_3}^k + \alpha_p {d_{z_3}}^k \\
{\lambda_0}^{k+1} &= {\lambda_0}^k + \alpha_d {d_{\lambda_0}}^k \\
{\lambda_1}^{k+1} &= {\lambda_1}^k + \alpha_d {d_{\lambda_1}}^k \\
{\lambda_2}^{k+1} &= {\lambda_2}^k + \alpha_d {d_{\lambda_2}}^k \\
{\lambda_3}^{k+1} &= {\lambda_3}^k + \alpha_d {d_{\lambda_3}}^k
\end{aligned}
\tag{3.8}
$$

As direções de busca do passo previsor $({d_x}^k, {d_{z_1}}^k, {d_{z_2}}^k, {d_{z_3}}^k, {d_{\lambda_0}}^k, {d_{\lambda_1}}^k, {d_{\lambda_2}}^k, {d_{\lambda_3}}^k)$ são determinadas considerando uma aproximação de primeira ordem do sistema sobre o novo ponto, enquanto as direções de busca do passo corretor $(\widetilde{d_x}^k, \widetilde{d_{z_1}}^k, \widetilde{d_{z_2}}^k, \widetilde{d_{z_3}}^k, \widetilde{d_{\lambda_0}}^k, \widetilde{d_{\lambda_1}}^k, \widetilde{d_{\lambda_2}}^k, \widetilde{d_{\lambda_3}}^k)$, consideram aproximações de segunda ordem.O parâmetro de barreira μ será pré-multiplicado pelo estimador δ, atualizado por:

$$
{\delta_i}^{k+1} = \frac{{\mu_i}^k {\delta_i}^k}{{\mu_i}^k + {z_i}^k} \tag{3.9}
$$

com sua respectiva direção d_δ, determinada por:

$$
\begin{aligned}
{\delta_i}^{k+1} = {\delta_i}^k + {(d_\delta)_i}^k = \frac{{\mu_i}^k {\delta_i}^k}{{\mu_i}^k + {z_i}^k} \Leftrightarrow {(d_\delta)_i}^k = \frac{{\mu_i}^k {\delta_i}^k}{{\mu_i}^k + {z_i}^k} - {\delta_i}^k \Leftrightarrow \\
\Leftrightarrow {(d_\delta)_i}^k = \frac{{\mu_i}^k {\delta_i}^k - ({\mu_i}^k + {z_i}^k){\delta_i}^k}{{\mu_i}^k + {z_i}^k} \Leftrightarrow {(d_\delta)_i}^k = -\frac{{\mu_i}^k {\delta_i}^k}{{z_i}^k + {z_i}^k}
\end{aligned}
\tag{3.10}
$$

O método PCPDBLM desenvolvido foi implementado em linguagem de programação C++, e aplicado a um problema de despacho econômico com restrições ambientais com 40 unidades geradoras, cujos resultados serão apresentados a seguir.

4 Resultados

O algoritmo do método PCPDBLM, foi implementado em Linguagem de Programação C++ e aplicado a um problema de despacho econômico com restrições ambientais, para 40 unidades geradoras, cujos coeficientes da função objetivo econômica, limites operacionais, podem ser encontrados em [Arantes et al., 2006] e [Coelho e Mariani, 2006]. Os coeficientes da função ambiental vistos na tabela 1, que não foram apresentados em [Arantes et al., 2006] e [Coelho e Mariani, 2006], foram gerados randomicamente através de um aplicativo do software C++, em subintervalos definidos através de valores máximos e mínimos para os coeficientes a_{ai}, b_{ai}, c_{ai} , determinados a partir da relação de grandeza entre os valores das funções econômica e ambiental, explorando-se a inter-relação destes vistas no problema com 6 unidades geradoras, cujos dados podem ser encontrados em [Souza, 2010].

Foram considerados 2 vetores de limitantes ambientais ${F_a}^{max} = ({F_a}^{max}{}_1, {F_a}^{max}{}_2, ..., {F_a}^{max}{}_{40})$ (casos I e II) que representam a máxima emissão de poluentes para cada uma das 40 unidades geradoras. Nas tabelas 2 e 3 estão discriminados os valores de máxima emissão ${F_a}^{max}{}_i$, os valores da emissão final $(F_a)_i$, atingida após a aplicação do método PCPDBLM, para cada unidade geradora i. O procedimento de barreira modificada permitiu inicializar o método com valores infactíveis na restrição quadrática, ou seja, mesmo que a restrição estivesse violada inicialmente, o método buscou, nas iterações subsequentes, pontos interiores à região viável.

Analisando as tabelas 2 e 3 constatamos que todas as emissões $(F_a)_i$ foram menores ou iguais ao máximo ${F_a}^{max}{}_i$ estipulado, atendendo aos limites superiores impostos.

4.1 Análise dos Resultados

A tabela 4 mostra os resultados da função econômica F_e e ambiental F_a para os dois casos testados neste trabalho em comparação aos resultados dos algoritmos encontrados em [Arantes et al., 2006], que resolveram um problema de despacho econômico com pontos de válvula através de uma investigação híbrida entre uma estratégia evolutiva e o método quase-Newton para busca local (EEQN), e [Coelho e Mariani, 2006], na qual o mesmo problema foi resolvido através da estratégia da evolução diferencial (ED). Com os valores das potências ótimas geradas relativas ao Caso I do método PCPDBLM, Caso II do método PCPDBLM, EEQN e ED respectivamente, foram determinados os valores da potência gerada total, $\sum_{i=1}^{40} P_i$, que deve

Tabela 1: Coeficientes da função ambiental, 40 unidades geradoras.

i	1	2	3	4	5	6	7	8	9	10
a_{ai}	0,0057	0,0046	0,0025	0,0028	0,0058	0,0053	0,0052	0,0056	0,0057	0,0052
b_{ai}	0,033	0,0458	0,0469	-0,0446	0,0008	0,0481	0,0167	0,0478	0,0499	0,0411
c_{ai}	7,248	19,834	18,317	19,22	10,18	14,774	6,007	17,934	14,468	17,984
i	11	12	13	14	15	16	17	18	19	20
a_{ai}	0,0033	0,0059	0,0047	0,0047	0,004	0,0056	0,0059	0,0043	0,0051	0,0049
b_{ai}	-0,0553	0,0281	0,01	-0,0319	0,0498	0,046	-0,0208	-0,0417	-0,0034	0,0463
c_{ai}	11,002	21,727	16,742	5,492	17,754	19,684	13,608	6,374	17,277	6,81
i	21	22	23	24	25	26	27	28	29	30
a_{ai}	0,0024	0,004	0,005	0,0036	0,0027	0,0038	0,0056	0,006	0,0025	0,0024
b_{ai}	0,0092	0,0387	0,0479	0,0462	0,0497	0,0356	0,0054	0,0088	0,0472	-0,0435
c_{ai}	20,634	11,574	9,36	19,848	12,101	18,162	21,305	18,734	19,399	14,765
i	31	32	33	34	35	36	37	38	39	40
a_{ai}	0,0029	0,0049	0,0051	0,0042	0,005	0,006	0,0058	0,0022	0,0056	0,0026
b_{ai}	0,0491	-0,0328	0,0311	-0,0313	0,0069	-0,0009	0,03	0,0423	0,0327	-0,0408
c_{ai}	5,914	7,28	7,546	20,767	22	9,143	7,102	11,21	11,206	6,195

Tabela 2: Vetor de Limitantes - Caso I

i	1	2	3	4	5	6	7	8	9	10
${F_a}^{max}_i$	86	85	60	112	65	126	480	472	460	500
$(F_a)_i$	85,09	84,84	59,95	111,83	64,83	125,39	479,02	472	460	498,31
i	11	12	13	14	15	16	17	18	19	20
${F_a}^{max}_i$	455	458	571	574	652	474	463	631	532	526
$(F_a)_i$	454,33	458	571	574	652	474	463	631	532	523
i	21	22	23	24	25	26	27	28	29	30
${F_a}^{max}_i$	752	651	518	726	856	692	149	156	83	35
$(F_a)_i$	751,69	651	518	725	856	692	146,95	153,80	82,18	33,18
i	31	32	33	34	35	36	37	38	39	40
${F_a}^{max}_i$	120	178	198	185	225	250	85	45	85	780
$(F_a)_i$	119,93	177,94	197,57	182,51	223,38	248,96	80,58	42,48	82,56	770,26

Tabela 3: Vetor de Limitantes - Caso II

i	1	2	3	4	5	6	7	8	9	10
${F_a}^{max}_i$	86	85	60	112	65	126	480	472	460	500
$(F_a)_i$	85,09	84,84	59,95	111,83	64,83	125,39	479,02	472	460	498,31
i	11	12	13	14	15	16	17	18	19	20
${F_a}^{max}_i$	455	458	571	574	652	474	463	631	532	526
$(F_a)_i$	454,33	458	571	574	652	474	463	631	532	523
i	21	22	23	24	25	26	27	28	29	30
${F_a}^{max}_i$	752	651	518	726	856	692	149	156	83	35
$(F_a)_i$	751,69	651	518	725	856	692	146,95	153,80	82,18	33,13
i	31	32	33	34	35	36	37	38	39	40
${F_a}^{max}_i$	120	178	198	185	225	250	85	45	85	780
$(F_a)_i$	119,93	177,94	197,57	182,51	223,38	248,96	80,58	42,48	82,56	770,255

Tabela 4: Comparação de Resultados

	PCPDBLM Caso I	PCPDBLM Caso II	EEQN	ED
$\sum_{i=1}^{40} P_i$	10500	10500	10500	10296,34
F_e	**118280,1**	155409,8	120364,7	130934,3
F_a	18012,24	**14506,05**	17877,9	16814,6

ser igual a potência demandada de 10500MW, da função econômica, F_e, utilizando os coeficientes encontrados em [Arantes et al., 2006] e [Coelho e Mariani, 2006], e da função ambiental, F_a, utilizando-se os dados gerados randomicamente apresentados na Tabela 1.

Os resultados apresentados demonstram a eficiência do método PCPDBLM. Este operou inicialmente com algumas restrições de emissão máxima por unidade geradora violadas, factibilizando-as no decorrer do processo, obtendo a solução ótima para o problema investigado, relativos aos Casos I e II, as quais são melhores do que as soluções obtidas pelos métodos EEQN e ED, vistos em [Arantes et al., 2006] e [Coelho e Mariani, 2006], de acordo com os valores das funções objetivos econômica e ambiental apresentados a Tabela 4.

5 Conclusões

No presente trabalho de desenvolvemos um método primal-dual de pontos interiores com procedimento previsor-corretor, baseado na função barreira logarítmica modificada, os quais foram aplicados à determinação de soluções eficientes de problemas multiobjetivo de despacho econômico e ambiental. Para a aplicação do método desenvolvido, foi considerado um problema de despacho econômico com restrições ambientais para cada unidade geradora, obtido através do método ε-restrito. Os resultados obtidos demonstraram a eficiência do método PCPDBLM, tanto computacionalmente, quanto nos resultados numéricos, em uma aplicação para um problema teste de 40 unidades geradoras, em comparação com resultados encontrados na literatura, principalmente em [Arantes et al., 2006] e [Coelho e Mariani, 2006]. Esses autores resolveram o caso mono-objetivo de despacho econômico considerando ponto de válvula, enquanto que neste trabalho, após gerarmos randomicamente os coeficientes da função ambiental, aplicamos o método PCPDBLM que possibilitou operar com valores infactíveis nas iterações iniciais e que na prática facilitou a escolha de pontos iniciais para a inicialização do método.

Referências

[Arantes et al., 2006] Arantes, M. B., Oliveira, G. T. S. and Saramago, S. F. P. (2006). Evolução Diferencial Aplicada à Solução de Alguns Problemas de Engenharia de Produção. *FAMAT em Revista*, n. 06.

[Coelho e Mariani, 2006] Coelho, L. S. and Mariani, V. C. (2006). Otimização de Despacho Econômico com Ponto de Válvula Usando Estratégia Evolutiva e Método Quase-Newton. *Learning Nonlinear Models, Revista da Sociedade Brasileira de Redes Neurais (SBRN)*, vol. 04, n. 01, 1–12.

[El-Hawary et al., 1992] EL-Hawary, M. E., EL-Hawary, F., Mbamalu, G. A. N. (1992). NOx emission performance models in electric power system. *Canadian Conference on Electrical and Computer Engineering*, vol. 02.

[Fang e Puthenpura, 1993] Fang, S. C., Puthenpura, S.(1993). Linear Optimization and Extensions: Theory and Algorithms. *Englewood Cliffs, New Jersey.*

[Fiacco e McCormick, 1968] Fiacco A. V., McCormick G. P. (1968). Nonlinear programming: sequential unconstrained minimization techniques. *John Wiley e Sons, New York.*

[Frisch, 1955] Frisch, K. R. (1955). The Logarithmic Potential Method of Convex Programming. *University Institute of Economics (manuscript). Oslo, Norway.*

[Gent e Lamont, 1971] Gent, M. R., Lamont J. W. (1971). Miminum-emission dispatch. *IEEE Transactions on Power Apparatus and Systems*, vol. 03, n. 06, 2650–2660.

[Haimes et al., 1971] Haimes, Y. Y., Lasdon, L., Wismer, D. A. (1971). On a Bicriterion Formulation of the Problems of Integrated System Identification and System Optimization. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-1(3).

[Karmarkar, 1984] Karmarkar, N. (1984). A new polynomial time algorithm for linear programming. *Combinatória 4*, 373–395.

[Kojima et al., 1989] Kojima, M., Mizuno, S., Yoshise, A. (1989). A primal dual - interior point method for linear programming, it Progress in Mathematical Programming: Interior-Point and Related Methods. *Springer-Verlag, New York*, 29–48.

[Mehrotra e Sun, 1992] Mehrotra, S., Sun, J. (1992). An algorithm for convex quadratic programming that requires 0 $(n^{3.5}L)$ arithmetic operations. *Mathematics of Operations Research*, vol. 15 342–363.

[Miettinen, 1999] Miettinen, K. (1999). Nonlinear Multiobjective Optimization. *Kluwer, Boston.*

[Monteiro et al., 1990] Monteiro, R. C., Adler, I., Resende, M. C. (1990). A polynomial-time primal-dual affine scaling algorithm for linear and convex quadratic programming and its power séries extension. *Mathematics of Operations Research*, vol. 15, 191–214.

[Polyak, 1992] Polyak, R. A. (1992). Modified barrier functions. *Mathematical Programming*, vol. 54, n. 02, 177–222.

[Souza, 2010] Souza, M. A. S. (2010). Investigação e Aplicação de Métodos Primal Dual de Pontos Interiores em Problemas de Despacho Econômico e Ambiental. *Dissertação de Mestrado, Universidade Estadual Paulista - Unesp - Bauru-SP.*

[Steinberg e Smith, 1943] Steinberg, M. J. C., Smith, T. H. (1943). Economic Loading of Power Plants and Electric Systems. *MacGraw-Hill.*

[Wright, 1997] Wright, S. J. (1997). Primal-Dual Interior Point Methods. *SIAM Journal*, 289–304.

[Wu et al., 1994] Wu, Y. C., Debs, A. S., Marsten, R. E.(1994). A direct nonlinear Predictor-Corrector Primal-Dual Interior Point Algorithm for Optimal Power Flows. *IEEE Transactions on Power Systems*, vol.09, n. 02, 876–883.

Cell-free Layer Measurements in Bifurcating Microchannels: a global approach

B. Taboada*, D. Bento*, D. Pinho*, A.I. Pereira‡, R. Lima*

*Instituto Politécnico de Bragança, Portugal & CEFT, Faculty of Engineering at University of Porto, Portugal.
E-mail:{brunataboada,davidbento,diana,ruimec}@ipb.pt

‡ Instituto Politécnico de Bragança, Portugal & Algoritmi, Universidade do Minho, Portugal.
E-mail:apereira@ipb.pt

Abstract

In the present work, *in vitro* blood flowing through bifurcating microchannels was studied, with the aim of characterizing the cell-free layer (CFL). The original images were obtained by means of a high-speed video microscopy system and then processed in MatLab using the Image Processing Toolbox. The numerical data was obtained automatically and analyzed by optimization techniques using the genetic algorithm approach. The results suggest that the CFL were formed in a similar way at the upper and lower regions in all bifurcations, and the measurements can be approximated through a sum of trigonometric functions.

Keywords: Cell-free Layer. Image Processing. Nonlinear Optimization. Global Optimization.

1 Introduction

Blood flow in microcirculation shows several hemodynamic phenomena that happen at both, *in vivo* and *in vitro* blood flowing. Hence, over the years *in vitro* blood studies in microchannels have been extensively performed in order to obtain an understanding of blood rheology and its flow dynamics [Goldsmith et al., 1989, Pires et al., 1992, Lima et al., 2012]. A well known hemodynamic phenomenon observed in both studies, *in vivo* and *in vitro*, states that for narrow microtubes ($< 300\ \mu m$) both hematocrit (Hct) and apparent blood viscosity decreases as the tube diameter is reduced [Goldsmith et al., 1989, Pires et al., 1992, Lima et al., 2012]. The physical reasons for these phenomena and for the formation of the cell-free layer (CFL), is known as Fahraeus-Lindqvist effect that is the tendency of red blood cells (RBCs) to migrate toward the centre of the microtube resulting in a marginal CFL at regions adjacent to the wall [Caro et al., 1978]. Recently several studies showed strong evidence that the formation of the CFL is affected by the geometry of the microchannel [Pinho, 2011, Lima et al., 2008a, Lima et al., 2008b, Leble et al., 2011] and the physiological conditions of the working fluid, such as the hematocrit [Garcia et al., 2012, Fujiwara et al., 2009].

Although there have been several studies on the measurement of CFL thickness [Namgung et al., 2010, Kim et al., 2006] according to our knowledge there have been very few studies on the determination of CFL measurements [Bento et al., 2012, Bento et al., 2013]. For these studies image analysis is extremely important to obtain crucial information [Pinho et al., 2012].

The main purpose of the present work is to measure the CFL in a microchannels with bifurcations by means of an automatic method developed in MatLab [Bento et al., 2012, Bento et al., 2013], and characterize the CFL along this microchannel using global optimization techniques.

The paper is organized as follows. The Section 2 presents the materials used in this work and the methods that were applied in this study. The Section 3 presents a brief description of the genetic algorithm. The Section 4 presents the numerical results and some discussion. The last section presents some conclusions and future work.

2 Materials and Methods

2.1 Experimental set-up

The high-speed video microscopy system used in this study consists of an inverted microscope (IX71; Olympus) combined with a high-speed camera (i-SPEED LT). The microchannel was placed on the stage of the inverted microscope and by using a syringe pump (PHD ULTRA) a pressure-driven flow was kept constant (cf. Figure 1).

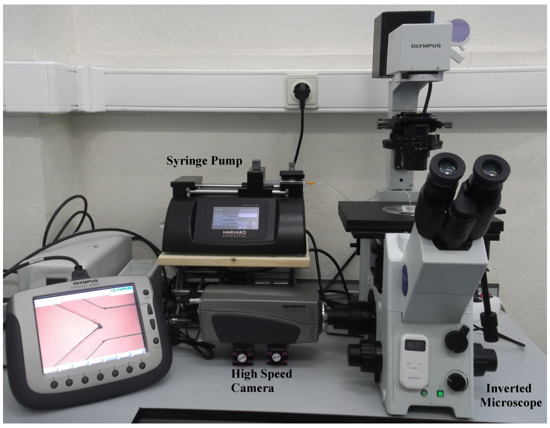


Figure 1: High-speed video microscopy system used in this staudy.

The series of microscope images were captured with a resolution of 600×800 pixels. All the images were recorded at the center plane of the microchannels at a rate of 200 frames/second and were transferred to the computer and then evaluated using image analysis techniques.

2.2 Working fluids and microchannel geometry

The blood samples used were collected from a healthy adult sheep, and ethylenediaminetetraacetic acid (EDTA) was added to prevent coagulation. The red blood cells (RBCs) were separated from the blood by centrifugation and washed twice with physiological saline (PS). The washed RBCs were suspended in Dextran 40 (Dx 40) to make up the required RBCs concentration by volume - hematocrit. In this study the Hct of 10% was used. All blood samples were stored hermetically at $4°C$ until the experiment was performed.

The microchannels fabricated for the proposed study have been produced in PDMS by a soft-lithography technique. The geometry used is a network of microchannels, containing several bifurcations and confluences. Figure 2 illustrates the configuration of the network and the regions where the CFL was measured.

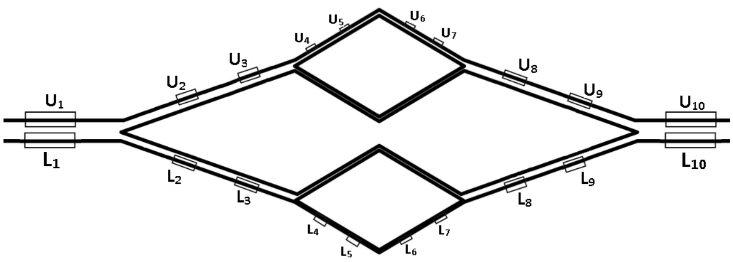


Figure 2: The geometry of the network and the regions where the CFL was measured.

2.3 Image Analysis

All image sequences, like Figure 3, were processed using Image Processing toolbox, available in MatLab [Matlab, 2002], and an automatic method was developed and tested [Bento et al., 2013]. The method consists in the combination of binarization of all images sequence and evaluation of all pixels using the maximum intensity.

Firstly a median filter with a 3×3 pixel mask was applied to each frame to reduce the noise, and then, the intensity of each pixel in the frame sequence was evaluated to obtain an image with the maximum intensity, as is possible to observe in Figure 4.

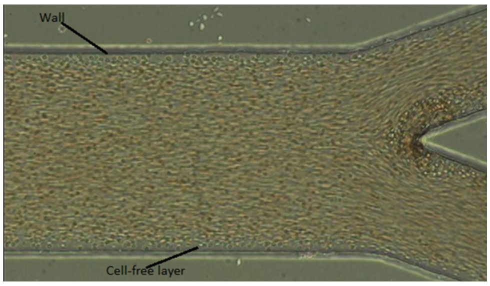


Figure 3: The original image of blood sample flowing in a bifurcating microchannel.

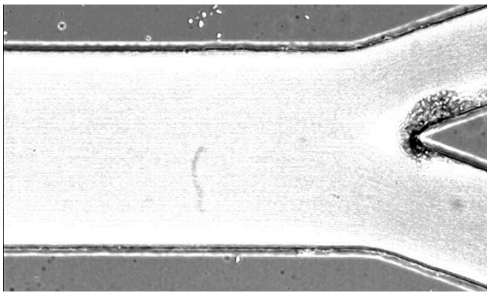


Figure 4: The image after the evaluation with the maximum intensity.

With this last step it is possible to identify the region of the concentration of blood cells (RBC core) and the region where blood cells do not exist, the cell-free layer. As the final step the image (Figure 4) was converted into a binary image. Finally, the region of interest is selected and the upper and lower CFLs are automatically measured. Figure 5 represents the image after binarization.

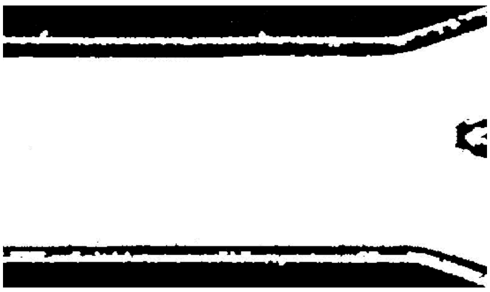


Figure 5: Binary image which shows clearly the boundary of CFL and RBC core.

3 Global optimization method: Genetic Algorithm

In the present work the genetic algorithm (GA) was used. The GA is an optimization technique based on the principles of evolution, this method allows to find a global minimum in a large search space. Given an individual, that is defined by the genes, this method creates new generations through analysis of the individuals genes from previous generations and selecting the best ones by applying the fitness function.

The GA uses crossover process, where the genes of the best individuals are crossed with genes from other individuals which also have good performance. The GA also applies the concept of mutation, improving the optimization process by introduction values that were not present in the previous generations.

This technique does not guarantee that the global minimum is found, however it is accepted that the final solution is close to the global minimum provided that successive generations cannot produce better individuals [Catlin et al., 2011].

4 Results and Discussion

The captured videos were analyzed and the numerical data was taken in the regions already defined in Figure 2. The flow rate used was 500 nl/min and one fluid was studied with 10% of Hct.

To fit the numerical data the nonlinear least squares theory was used. In each region U_i, for $i = 1, ..., 10$, and L_i, for $i = 1, ..., 10$, we applied the nonlinear optimization problem defined as follows

$$\begin{array}{ll} \min & f(y) \equiv \sum_{k=1}^{N_R} (M_k - g_h(y, x_k))^2 \\ s.t & g_h(y, x_k) \geq 0 \ \forall k = 1, ..., N_R \end{array} \tag{4.1}$$

where (x_k, M_k), for $k = 1, ..., N_R$ are the CFL measurements of the region R (defined as U_i and L_i, for $i = 1, ..., 10$). The functions g_h, for $h = 1, ..., 3$, are defined as follows

$$\begin{array}{l} g_1(y, x) = y_1 x^2 + y_2 x + y_3, \\ g_2(y, x) = y_1 x + y_2, \\ g_3(y, x) = sin(y_1 x) + cos(y_2 x) + y_3. \end{array} \tag{4.2}$$

The following table presents the obtained numerical results using the genetic algorithm to solve the optimization problem (4.1). Since the genetic algorithm is a stochastic method, each problem was solved 30 times. Table 1 presents the regions where the problem (4.1) was applied, the average of the optimum value and the minimum value obtained in the all 30 runs.

Table 1: Numerical results obtained using genetic algorithm.

	Upper cell-free layer				Lower cell-free layer		
Region	Function	Average	Minimun	**Region**	Function	Average	Minimun
$\mathbf{U_1}$	g_1	2.71×10^7	8.14×10^3	$\mathbf{L_1}$	g_1	1.13×10^7	1.05×10^5
	g_2	2.17×10^3	1.91×10^3		g_2	1.41×10^3	1.34×10^3
	$\mathbf{g_3}$	$\mathbf{1.90 \times 10^3}$	$\mathbf{1.50 \times 10^3}$		$\mathbf{g_3}$	$\mathbf{1.19 \times 10^3}$	$\mathbf{1.00 \times 10^3}$
$\mathbf{U_2}$	g_1	4.29×10^7	2.86×10^3	$\mathbf{L_2}$	g_1	4.63×10^6	2.36×10^3
	g_2	1.27×10^3	4.02×10^2		g_2	2.32×10^3	2.07×10^3
	$\mathbf{g_3}$	$\mathbf{7.68 \times 10^2}$	$\mathbf{5.38 \times 10^2}$		$\mathbf{g_3}$	$\mathbf{2.31 \times 10^3}$	$\mathbf{1.85 \times 10^3}$
$\mathbf{U_3}$	g_1	1.32×10^8	5.99×10^3	$\mathbf{L_3}$	g_1	9.30×10^6	3.44×10^3
	g_2	4.76×10^3	1.67×10^3		g_2	3.34×10^3	3.06×10^3
	$\mathbf{g_3}$	$\mathbf{1.92 \times 10^3}$	$\mathbf{1.60 \times 10^3}$		$\mathbf{g_3}$	$\mathbf{2.81 \times 10^3}$	$\mathbf{2.36 \times 10^3}$
$\mathbf{U_4}$	g_1	2.88×10^6	1.14×10^4	$\mathbf{L_4}$	g_1	4.61×10^7	9.63×10^4
	g_2	1.27×10^3	1.21×10^3		g_2	5.97×10^3	5.18×10^3
	$\mathbf{g_3}$	$\mathbf{9.94 \times 10^2}$	$\mathbf{7.75 \times 10^2}$		$\mathbf{g_3}$	$\mathbf{4.61 \times 10^3}$	$\mathbf{4.07 \times 10^3}$
$\mathbf{U_5}$	g_1	1.30×10^7	1.68×10^3	$\mathbf{L_5}$	g_1	1.28×10^7	2.26×10^3
	g_2	1.77×10^3	1.62×10^3		$\mathbf{g_2}$	$\mathbf{2.38 \times 10^3}$	$\mathbf{2.04 \times 10^3}$
	$\mathbf{g_3}$	$\mathbf{1.43 \times 10^3}$	$\mathbf{1.20 \times 10^3}$		g_3	3.74×10^3	3.35×10^3
$\mathbf{U_6}$	g_1	2.61×10^6	2.66×10^3	$\mathbf{L_6}$	g_1	8.35×10^6	5.49×10^3
	g_2	6.07×10^2	5.02×10^2		g_2	9.12×10^3	5.46×10^3
	$\mathbf{g_3}$	$\mathbf{3.70 \times 10^2}$	$\mathbf{3.31 \times 10^2}$		$\mathbf{g_3}$	$\mathbf{1.31 \times 10^3}$	$\mathbf{9.90 \times 10^2}$
$\mathbf{U_7}$	g_1	1.18×10^7	3.73×10^4	$\mathbf{L_7}$	g_1	6.90×10^5	3.53×10^3
	g_2	8.65×10^2	6.53×10^2		g_2	4.91×10^3	2.71×10^3
	$\mathbf{g_3}$	$\mathbf{6.45 \times 10^2}$	$\mathbf{5.28 \times 10^2}$		$\mathbf{g_3}$	$\mathbf{3.30 \times 10^3}$	$\mathbf{2.55 \times 10^3}$
$\mathbf{U_8}$	g_1	1.53×10^7	1.20×10^4	$\mathbf{L_8}$	g_1	3.32×10^7	8.49×10^4
	g_2	2.72×10^3	1.62×10^3		$\mathbf{g_2}$	$\mathbf{2.96 \times 10^3}$	$\mathbf{1.31 \times 10^3}$
	$\mathbf{g_3}$	$\mathbf{1.82 \times 10^3}$	$\mathbf{1.34 \times 10^3}$		g_3	3.90×10^3	2.76×10^3
$\mathbf{U_9}$	g_1	1.58×10^8	2.73×10^3	$\mathbf{L_9}$	g_1	2.10×10^8	7.36×10^4
	g_2	1.24×10^3	9.09×10^2		$\mathbf{g_2}$	$\mathbf{1.52 \times 10^3}$	$\mathbf{9.08 \times 10^2}$
	$\mathbf{g_3}$	$\mathbf{1.10 \times 10^3}$	$\mathbf{8.30 \times 10^2}$		g_3	1.59×10^3	1.09×10^3
$\mathbf{U_{10}}$	g_1	1.21×10^7	1.64×10^4	$\mathbf{L_{10}}$	g_1	2.29×10^8	3.08×10^3
	g_2	2.03×10^3	1.74×10^3		g_2	3.04×10^3	2.41×10^3
	$\mathbf{g_3}$	$\mathbf{1.77 \times 10^3}$	$\mathbf{1.38 \times 10^3}$		$\mathbf{g_3}$	$\mathbf{2.19 \times 10^3}$	$\mathbf{1.81 \times 10^3}$

The numerical results suggested that the function g_3 fits better the CFL boundary, since the minimum value was obtained for this function. Other important fact is that the boundary has the same behavior in all bifurcations. The only exception was verified in the regions L_5, L_8 and L_9, in these regions the best fit was obtained with the function g_2.

5 Conclusions and Future Work

In this study, we present a method to obtained automatically the CFL measurements. The CFL boundary was fitted using three different functions. The genetic algorithm was used to solve the constrained optimization problem. The best fit was obtained using the function g_3, i.e. a sum of trigonometric functions. As future work, we will test more fluids with different properties and different functions to fit the CFL measurements.

Acknowledgments

The authors acknowledge the financial support provided by: PTDC/SAU-BEB/108728/2008, PTDC/SAU-BEB/105650/2008, PTDC/EME-MFE/099109/2008 and PTDC/SAU-ENB/116929/2010 from FCT (Science and Technology Foundation), COMPETE, QREN and European Union (FEDER).

References

[Bento et al., 2013] Bento, D., Pinho, D., Pinto, E., Tomoko, Y., Correia, T., Lima, J., Pereira, A.I., Lima, R. (2013). Cell-free layer measurements in a bifurcation microchannel: comparison between a manual and automatic methods. 5th Portuguese Congress on Biomechanics: 359-362.

[Bento et al., 2012] Bento, D., Pinho, D., Pereira, A., Lima, R. (2012) Cell-Free Layer (CFL) Analysis in a Glass Capillary: Comparison Between a Manual and Automatic Method. AIP Conference Proceedings 1479: 786-789.

[Caro et al., 1978] Caro, C., Pedley, T., Schroter, R., Seed, W. (1978). The mechanics of the circulation. *Oxford University Press.*

[Catlin et al., 2011] Catlin, G., Advani, S., Prasad, A. (2011). Optimization of polymer electrolyte membrane fuel cell channels using a genetic algorithm. *Journal of Power Sources*, 196:9407–9418.

[Fujiwara et al., 2009] Fujiwara, H., Ishikawa, T., Lima, R., Matsuki, N., Imai, Y., Kaji, H., Nishizawa, M., Yamaguchi, T. (2009). Red blood cells motions in high-hematocrit blood flowing through a stenosed microchannel. *Journal of Biomechanics*, 42:838–843.

[Garcia et al., 2012] Garcia, V., Dias, R., Lima R. (2012). In Vitro Blood Flow Behaviour in Microchannels with Simple and Complex Geometries. *Journal of Biomechanics Applied Biological Engineering - Principles and Practice, Ganesh R. Naik (ed.), InTechcs*, 17:394–416.

[Goldsmith et al., 1989] Goldsmith, H., Cokelet, G., Gaehtgens, P. (1989). Robin Fahraeus: evolution of his concepts in cardiovascular physiology. *American Journal of Physiology*, 257:H1005–H10015.

[Kim et al., 2006] Kim, S., Kong, R. L., Popel, A. S., Intaglietta, M., Jonhson J. C.(2006). A computer-based method for determination of the cell-free layer width in microcirculation. *Microcirculation*, 13:199-207.

[Leble et al., 2011] Leble, V., Lima, R., Dias, R. P., Fernandes, C. S., Ishikawa, T., Imai, Y., Yamaguchi, T. (2011). Asymmetry of red blood cell motions in a microchannel with a diverging and converging bifurcation. *Biomicrofluidics*, 5:044120.

[Lima et al., 2008a] Lima, R., Wada, S., Tanak, S., Takeda, M. (2008). In vitro blood flow in a rectangular PDMS microchannel: experimental observations using a confocal micro-PIV system. *Biomedical Microdevices*, 10(2):153–167.

[Lima et al., 2008b] Lima, R., Ishikawa, T., Imai, Y. (2008). Radial dispersion of red blood cells in blood flowing through glass capillaries: role of hematocrit and geometry. *Journal of Biomechanics*, 41:2188-2197.

[Lima et al., 2012] Lima, R., Ishikawa, T., Imai, Y., Yamaguchi, T. (2012) Single and two-phase flows on Chemical and Biomedical Engineering. Bentham Science, Dias et al., 513-547.

[Matlab, 2002] Eddins, S. L., Gonzalez, R. C., Woods, R. E. (2002). Digital Image Processing Using Matlab. *MathWorks.*

[Namgung et al., 2010] Namgung, B., Ong, P. K., Wong, Y. H., Lim, D., Chun, K. C., Kim, S. (2010). A comparative study of histogram-based thresholding methods for the determination of cell-free layer width in small blood vessels. *Physiological Measurement*, 31:N61–N70.

[Pinho, 2011] Pinho, D. (2011). Determination and characterization of red blood cells trajectories: a semi-automatic method. *Master in Biomedical Technology, Polytechnic Institute of Bragança, Portugal, 2011 (in portuguese).*

[Pinho et al., 2012] Pinho, D., Lima, R., Pereira, A. I., Gayubo, F. (2012). Automatic tracking of labeled red blood cells in microchannels. *International Journal for Numerical Methods in Biomedical Engineering*, DOI: 10.1002/cnm.

[Pires et al., 1992] Pires, A. A., Neuhaus, D., Gaehtgens P. (1992). Blood viscosity in tube flow: dependence on diameter and hematocrit. *American Journal of Physiology*, 263:H1770–H1778.

OR/MS EDUCATION: an overview of the 2003-2012 decade

Ana Paula Teixeira*, João Miranda†

* *Department of Mathematics, Sciences and Technology School, University of Trás-os-Montes e Alto Douro, Vila Real, Portugal; CIO - Operations Research Center, FCUL,*
E-mail:`ateixeir@utad.pt`

† *College of Technology and Management, Portalegre Polytechnics Institute, Lugar da Abadessa, Apt. 148, 7301-999 Portalegre, Portugal; CPQ-Centre for Chemical Processes (IST/UTL),*
E-mail:`jlmiranda@estgp.pt`

Abstract

The main characteristics of talks on OR/MS education are described and analysed, not only in order to determine the interest of this topic and its evolution over the last decade, but also to improve our understanding on this issue. A database with information about the talks on OR/MS education is built, considering relevant conferences in OR/MS and Education at national or international level, within the period 2003-2012. Beyond the good practices in OR/MS education subjects, such as classroom approaches, courses design, and assessment, the application of OR/MS tools on Education is also analysed, namely: Data Envelopment Analysis (DEA), Modeling, Scheduling/Timetables, Decision Support Systems (DSS), and Routing for school transportation. Additionally, the developments on OR/MS education in the last decade are evaluated, observing what has been done, and describing the related applications and procedures. The result is a review of the good practices that have been developed in OR/MS education and of the OR/MS tools applied to education, providing a new insight on the recent enhancements of OR/MS education and also outlining other pathways for the near future.

Keywords: OR/MS education; OR/MS tools; overview; good practices; conferences.

1 Introduction

In the last few decades, there has been increasing interest within the Operations Research/Management Science (OR/MS) academic community into how to shape methodologies and approaches for enhancing education; for instance, as described by [Teixeira et al. 2011] for learners at the pre-university levels. Recently, [Miranda and Teixeira 2011b] also focused the Portuguese efforts involved on the Mathematics Action Program (*Plano de Acção da Matemática*) and considered its achievements and results in face of the Engineering Education needs.

Additionally, [Miranda 2011], [Amaro et al. 2012] and [Nagy and Miranda 2012] addressed the quality and enhancements of OR/MS education from the point of view of the international cooperation and by sharing good practices. However, not much effort has been made to examine the good practices and their attributes that are crucial to successful OR/MS education. A seminal work on these attributes and good practices has been done by [Miranda and Teixeira 2011a], but they consider that further analysis on this subject is still needed, in order to provide better guidance to OR/MS teaching staff.

The purpose of this paper is to present some of the work that has been developed concerning the theme of OR/MS education. An overview both on the good practices in "OR/MS education" and "OR/MS tools applied to education" is outlined, in order to gain insight on the future directions of this field and to draw some improvements from the most recent developments. Here, the focus goes essentially to the works presented in OR/MS and Education conferences in the last decade (2003-2012), both at national and international level. This study is only a small part of the research work that the authors continue to develop on this theme.

By their personal perception, the authors characterized the OR/MS education topic and identified the related key attributes, as follows:

- selecting the practices directly related to the education of OR/MS, like classroom practices, design and review of OR/MS courses and learning assessments;

- targeting Education issues that required the application of several OR/MS tools, for instance, DEA, Modeling, Scheduling/Timetables, DSS in education, and Routing for school transportation.

Although a significant number of OR/MS education enhancements have been presented in conferences during the last decade, further developments and applications are still in need. The option to focus conferences and the related documentation is based on the rapidity of dissemination and the sense of utility by the peers, as referred by [Ranyard, 2011], among others, on the IFORS2011 stream dedicated to the global survey of OR practice.

We have been building a database of the sessions/workshops and talks/posters in OR/MS education presented in relevant OR/MS and Education conferences during the last decade. The topic, the methodology and techniques analysis, as well as other important aspects of the talks were explored. The aim of this study is to analyze the main features of a series of conferences on the concept of OR/MS education that have been disseminated in some of the most prestigious OR/MS and Education events in the last decade, 2003-2012.

This literature analysis is quite large and it must be delimited, so that the quantity and diversity of data do not prevent us of achieving our purposes. We should point out that we are only interested in those events that can be considered essential references in the OR/MS education field. Therefore, we have selected a set of relevant national and international series of conferences.

Furthermore, it is necessary to analyze works of good quality, so that our conclusions will be based on education, research and innovation developments that are observing good quality criteria. Thus, we have selected a group of conferences of renowned prestige among the national and international academic community, completing the selection with other events that are highly regarded in the area of OR/MS education. We also observe that the authors use to attend to most of the analysed conference series, thus having a better perception of the main areas and most focused themes, as well as the prior knowledge of the key subjects to search in.

The structure of the paper considers: in Section 2, the methodology of the study is presented, targeting good practices in OR/MS education and the application of OR/MS tools on Education; in Section 3, the key subjects of this overview are described; and in Section 4, the main conclusions are presented.

2 Methodology of the Study

The sources of information required for the present research consist of a group of specialized conferences considered likely to include works in the OR/MS education field. We have selected sessions, workshops, talks, and posters, under the assumption that quality works tend to be presented in this type of conferences

The next step was to select the scientific conferences that would be considered in this study. We opted to select the most prestigious conferences in the area of OR/MS and Education, both at international (global and European) and national (Portuguese and Spanish) level. We chose the topics 'Education', 'Good practices' and 'Operations Research and Management Science', which we felt that would be three of the most likely to include works on our focus upon OR/MS education.

As we pointed out earlier, the idea was to analyze those talks concerning:

- aspects directly related to 'Good practices' in OR/MS education, in particular on the: i) OR/MS practices in the classroom; ii) design, implementation and review of OR/MS courses, and c) students' performance assessment in OR/MS subjects;
- application of OR/MS tools on Education, namely, DEA, Modeling, Scheduling/Timetables, DSS in education, and Routing for school transportation.

With this in mind, we carried out a conference search using the terms 'good practices', 'education' and 'operations research'/'management science' in the title, abstract or keywords of the talks appearing in the selected conferences during the period of analysis. We also conducted a manual survey of all the abstract books/proceedings because we noticed that some talks on the topics under study were included in streams not associated with these themes. We recall that the authors attended to most of the analysed conferences series, having actively participated in several OR Education sessions; this allowed us to have a better perception of the key areas and issues covered. During these participations, we analyzed not only the presentations with titles of interest, but also the abstracts/papers published.

Our database reports to the following six series of conferences:

- *European Conference on Operational Research-EURO;*
- *Conference of the International Federation of Operational Research Societies-IFORS;*
- *International Conference on Engineering Education-ICEE;*
- *Congresso da Associação Portuguesa de Investigação Operacional-IO;*
- *Congreso Nacional de Estadística e Investigación Operativa-SEIO;*

- *Congreso da Sociedade Galega para a Promoción da Estatística e Investigación de Operacións-SGAPEIO.*

As it can be observed in Table 1, we consider thirty six conferences involving: i) fifty eight sessions/workshops and one hundred and nineteen talks/posters devoted to the theme of Education of OR/MS; and ii) fifty seven sessions/workshops and one hundred and two talks/posters devoted to the theme of OR/MS tools applied to Education. This information was obtained from the respective conference Book of Proceedings/Abstracts. In Tables 1 to 7 the results appear in the following form: "data relative to Education of OR/MS" at the slash (/) left side; and in its right side, the "data relative to OR/MS tools applied to Education".

Table 1: List of conferences

Conference Series	**Number of Conferences**	**Number of Sessions**	**Number of Talks**
EURO	6	23/39	49/75
IFORS	3	19/8	42/14
ICEE	10	3/0	3/0
IO	5	4/7	12/10
SEIO	7	5/1	6/1
SGAPEIO	5	4/2	7/2
Total	**36**	**58/57**	**119/102**

We note that ICEE conferences are annual, while the others have different cadences. As the posters and workshops appear in a very small number, we decided to include them in talks and sessions, respectively. In particular, IFORS 2008 and EURO 2006 had two workshops each; in SEIO 2010 there were one workshop and one poster. Finally, in IO 2006 there were three posters and in SGAPEIO 2003, SGAPEIO 2009 and SEIO 2012 there was one poster each.

On a brief analysis of the aggregated results on Table 1, we consider:

- the high interest of the OR/MS education subjects both in the EURO and IFORS conferences;
- although the OR/MS education topic is addressed in ICEE conferences, it seems to be too specific for this conference series; in fact, ICEE considers various subjects of Engineering Education;
- a direct comparison between the Portuguese (IO) and Spanish (SEIO, SGAPEIO) conferences in OR/MS seems to indicate a more significant interest in the OR/MS education subjects in the IO conference series.

3 Results of the Study

The six series of conferences here considered differ somewhat from each other in terms of main focus and dimension:

- one of them is generally targeting Education subjects (ICEE series), while all the others are specific OR/MS conference series; and
- some are national conferences (IO, SEIO, and SGAPEIO series) and the others are international (EURO, ICEE, and IFORS series).

For these motives we decided to analyze the themes 'Education of OR/MS' and 'OR/MS tools applied to Education' separately for each series of conferences.

We observe that OR/MS education has experienced increasing interest in the period of analysis considered, a fact that is mainly reflected in the significant number of talks presented in the analyzed conferences (Tables 2 to 7). Analyzing the talks by research topic, we verify that during the period at hand (2003-2012):

- Six EURO conferences were held (Table 2). The theme 'Education of OR/MS' included twenty three sessions and forty nine talks; three on the student's performance assessment concerning OR/MS subjects, thirty four of them were about OR/MS practices in the classroom, other ten concerned the design, implementation and review of OR/MS courses, and two mentioned both these last two topics. On 'OR/MS tools applied to Education' occurred seventy five talks within thirty nine sessions; thirty six talks were about DEA, nine on Modeling, nineteen on Scheduling/Timetables, eight on DSS, two on Routing for school transportation and one in Routing for school transportation and Scheduling.

Table 2: EURO conferences

Conferences	Number of Sessions	Number of Talks
EURO 2004	1/3	1/7
EURO 2006	8/1	17/2
EURO 2007	2/7	5/18
EURO 2009	2/8	4/14
EURO 2010	6/10	16/20
EURO 2012	4/10	6/14
Total	**23/39**	**49/75**

- Three IFORS conferences were organized with nineteen sessions on 'Education of OR/MS' (Table 3); three on the student's performance assessment concerning OR/MS subjects, thirty one aimed OR/MS practices in the classroom, seven focused the design, implementation and review of OR/MS courses, and one mentioning both these last two topics. On 'OR/MS tools applied to Education' occurred eight sessions with twelve talks related to DEA and two in DSS.

Table 3: IFORS conferences

Conferences	Number of Sessions	Number of Talks
IFORS 2005	7/4	15/4
IFORS 2008	4/3	8/6
IFORS 2011	8/1	19/4
Total	**19/8**	**42/14**

- Ten ICEE conferences took place (Table 4), including three sessions and three talks on OR/MS practices in the classroom.

Table 4: ICEE conferences

Conferences	Number of Sessions	Number of Talks
ICEE 2003	1/0	1/0
ICEE 2004		
ICEE 2005		
ICEE 2006		
ICEE 2007	1/0	1/0
ICEE 2008	1/0	1/0
ICEE 2009		
ICEE 2010		
ICEE 2011		
ICEE 2012		
Total	**3/0**	**3/0**

- Five IO conferences have been organized (Table 5). This series included four 'Education of OR/MS' sessions, eight talks on OR/MS practices in the classroom, one on design, implementation and review of OR/MS courses, and three talks on students' performance assessment concerning OR/MS subjects. The theme 'OR/MS tools applied to Education' involved seven sessions with five talks in DEA, four in Scheduling/Timetables and one in Routing for school transportation.
- Seven SEIO conferences were also held (Table 6); including one session on 'OR/MS tools applied to Education' with one talk in Routing for school transportation, and five sessions on 'Education of OR/MS' with two talks on the students' performance assessment in OR/MS subjects and four talks aiming OR/MS practices in the classroom.

Table 5: IO conferences

Conferences	**Number of Sessions**	**Number of Talks**
IO 2004		
IO 2006	1/3	3/6
IO 2008	1/1	3/1
IO 2009	1/2	3/2
IO 2011	1/1	3/1
Total	**4/7**	**12/10**

Table 6: SEIO conferences

Conferences	**Number of Sessions**	**Number of Talks**
SEIO 2003		
SEIO 2004		
SEIO 2006	0/1	0/1
SEIO 2007		
SEIO 2009	1/0	1/0
SEIO 2010	2/0	3/0
SEIO 2012	2/0	2/0
Total	**5/1**	**6/1**

- Five SGAPEIO conferences were organized (Table 7), with four sessions on 'Education of OR/MS' and two sessions on 'OR/MS tools applied to Education'. The first session included six talks on OR/MS practices in the classroom and one on the students' performance assessment concerning OR/MS subjects; the second session had one talk on DEA and another on Routing for school transportation.

Table 7: SGAPEIO conferences

Conferences	**Number of Sessions**	**Number of Talks**
SGAPEIO 2003	1/0	1/0
SGAPEIO 2005	1/1	3/1
SGAPEIO 2007	0/1	0/1
SGAPEIO 2009	1/0	2/0
SGAPEIO 2011	1/0	1/0
Total	**4/2**	**7/2**

The EURO series (Table 2) presents approximately twice (50%) more sessions (talks) on 'OR/MS tools applied to Education' than on 'Education of OR/MS'. Concerning the IFORS series (Table 3), the sessions on 'Education of OR/MS' are almost three times more than on 'OR/MS tools applied to Education', presenting the number of talks similar proportion. The theme 'OR/MS tools applied to Education' doesn't appear on ICEE conferences (Table 4) held on the last decade and they only include three sessions and three talks on 'Education of OR/MS'. Although the five considered IO conferences (Table 5) present almost twice more sessions on 'OR/MS tools applied to Education' than on 'Education of OR/MS', the number of talks on both themes is similar. The SEIO series (Table 6) only include one session and one talk on 'OR/MS tools applied to Education', while the theme 'Education of OR/MS' has five sessions and six talks. Finally, the SGAPEIO series (Table 7) only include four (two) sessions in 'Education of OR/MS' ('OR/MS tools applied to Education'), being seven (two) the number of talks.

We observe that EURO is the conference series with the highest number of participants, over than two thousand; ICEE usually gathers between five and seven hundreds of researchers and the other conferences present a number of participants quite inferior to five hundred, being SGAPEIO the smallest one. This difference in the dimension of the six analyzed conferences can, somehow, explain the difference in the number of sessions/talks that exist on both studied themes. This number is much higher in EURO and IFORS than in the other conferences. Another reason for the lack of 'OR/MS tools applied to Education' sessions in ICEE may be explained, as already mentioned, by the fact that this is a conference on engineering education, a very broader area, being the topics 'Education of OR/MS' and 'OR/MS tools applied to Education' too specific.

4 Conclusions

The research on OR/MS education over the period 2003-2012 is analysed, based on the scientific talks appearing in relevant national and international conferences in the fields of education and OR/MS. It is important to note that throughout the study period, the number of sessions/talks on both considered topics is increasing, which shows the growing interest of teachers and researchers in both subjects.

During the last decade and for the six series of the analised conferences, the number of sessions on both studied themes is quite similar, but the number of talks on 'Education of OR/MS' is 15% higher than on 'OR/MS tools applied to Education'. But this is not a homogeneous behavior for all the six series of studied conferences. In fact:

- EURO and IO conferences focused more on the theme 'OR/MS tools applied to Education' than on 'Education of OR/MS', while
- the reverse situation occurred for the remaining conferences; no 'OR/MS tools applied to Education' talks appeared on ICEE series.

Concerning the ICEE series, the fact of this being an Engineering Education international conference series and the specificity of OR MS themes, may help to explain the reduced number of sessions and talks on OR/MS, compared with the other focused themes in this series of conferences.

We also observe the IO series of conferences addresses the theme of OR/MS Education since the 2006 conference. In IO-2006, large attention was devoted to this subject and a plenary session dedicated to OR/MS Education was held. Thereafter, the interest on education subjects has been increasing among Portuguese OR/MS teachers and researchers, and there have been parallel sessions on this topic in every IO conference.

On the contrary to what happens in Portugal, the areas of Statistics and of Operations Research are together in the same societies in Spain,, namely in SEIO and SGAPEIO. The analysis of the information obtained in this study seems to corroborate our perception, which result of the successive participation in SEIO and SGAPEIO conferences: the Statistics area has more emphasis in these conferences than the area of Operations Research and although "Education" is a topic of interest for many researchers and teachers participating these conferences, this subject is mainly addressed and explored by the Statistics participants.

Regarding EURO and IFORS, both are important series of international OR conferences, with a large number of participants and of general scope. The large number of sessions / talks on the theme of OR/MS Education shows the interest of the OR/MS community onto the education subjects and, somehow, corroborates the importance of this theme and the attention that it deserves.

We should stress that our intention was to analyze the main characteristics of the research in OR/MS education, an ambitious task interacting with many other fields of research within engineering fields, sciences of enterprise and management, and exact sciences. This work is pursuing our research on the theme 'OR/MS Education'; we intend to continue the study of this issue, for example, increasing both the time horizon under consideration and the topics to be addressed. Additionally, beyond EURO series, the conferences of other IFORS regional societies are to be addressed in future works, namely INFORMS, ALIO, and APORS. Moreover, some conferences of OR societies at national level may also be focused, for instance, the annual OR Society conference series.

We should stress that our intention was to analyze the main characteristics of the research in OR/MS education, an ambitious task interacting with many other fields of research within engineering fields, sciences of enterprise and management, and exact sciences. This work is pursuing the authors research on the theme 'OR/MS Education'; they intend to continue the study of this issue, for example, increasing both the time horizon under consideration and the topics to be addressed.

Acknowledgements

The authors thank the University of Trás-os-Montes and Alto Douro and, the Polytechnics Institute of Portalegre. They thank the CIO-Operations Research Center, and the Centre for the Chemical Processes (CPQ) that are supporting this work using funds from FCT-Fundação para a Ciência e a Tecnologia, under the projects PEst-OE/MAT/UI0152 and PEst-OE/EQB/UI0088. We also thank our Colleagues that have provided us some of the conferences proceedings.

References

[Amaro et al. 2012] Amaro, A., Bruno, G., Casquilho, M., Corominas, A., García-López, J.M., Genovese, A., Lusa, A., Magnusson, J., Matos, H., Miranda, J., Rubio, S. Odss.(2012). 4SC: A summer school in Optimization and Decision Support Systems for Supply Chains. In *Book of Abstracts of EUROXXV - 25th European Conference on Operational Research (Vilnius, Lithuania, July-2012).*

[Andersen et al. 2004] Andersen, T., Hoit, M., (2004). *Abstract Book of the ICEE 2004 Conference, Florida, USA.*

[Arthur and Labbé 2005] Arthur, J., Labbé, M. (2005). *Proceedings of the IFORS 2005 Conference, Honolulu, Hawaii.*

[Ayuso 2012] Ayuso, A. (2012). *Livro de Resumos da SEIO2012 - XXXIII Congreso Nacional de Estadística e Investigación Operativa. Cádiz, Spain.*

[Ballester and Ruiz 2003] Ballester, E., Ruiz, L (2003). *Abstract Book of the ICEE 2003 Conference, Valencia, Spain.*

[Cardoso and Clímaco 2011] Cardoso, D., Clímaco, J. (2011). *Livro de Resumos do IO2011-15 Congresso da APDIO. Universidade de Coimbra, Coimbra, Portugal.*

[Costa and Barbosa-Póvoa 2009] Costa, R., Barbosa-Póvoa, A. (2009). *Livro de Resumos do IO2009-14 Congresso da APDIO. Universidade Nova de Lisboa, Campus de Caparica, Portugal.*

[Guimarães and Júdice 2011] Guimarães, R., Júdice, J. (2011). *Livro de Resumos do IO2004-11 Congresso da APDIO. Universidade do Porto, Porto, Portugal.*

[Hinojosa 2004] Hinojosa, J. (2004). *Livro de Resumos da SEIO2004 - XXVIII Congreso Nacional de Estadística e Investigación Operativa. Cádiz, Espanha.*

[Ittmann and Bartholdi 2008] Ittmann, H., Bartholdi III, J. (2008). *Proceedings of the IFORS 2008 Conference, Sandton Convention Centre, Sandton, Gauteng Province, South Africa.*

[Kim and Kim 2009] Kim, C.Y., Kim., H.S. (2009). *Abstract Book of the ICEE2009 Conference. University of Seoul, Seoul, South Korea.*

[Kinnunen and Jablonsky 2007] Kinnunen, T., Jablonsky, J. (2007) *Book of Abstracts of the 22nd Conference on Operational Research. University of Economics, Prague, Czech Republic.*

[Mecsi 2008] Mecsi, J., (2008) *Abstract Book of the ICEE2008 Conference. University of Pécs and Budapest University of Technology and Economics, Hungary.*

[Mertanen and Roslöf 2012] Mertanen, O., Roslöf, J., (2012) *Abstract Book of the ICEE2012 Conference. Turku University of Applied Sciences, Turku, Finland.*

[Miranda 2011] Miranda, J.L.(2011) A study in cooperation and good practices in OR/MS teaching. In *Proceedings of IFORS2011 - 19th Conference for the International Federation of Operational Research Societies (Melbourne, Australia, July-2011) 86.*

[Miranda and Teixeira 2011a] Miranda, J.L., Teixeira, A.P. (2011) A First Overview of Good Practices in OR/MS Education. In *Proceedings of X Congreso Galego de Estatística e Investigación de Operacións*, pp 6.

[Miranda and Teixeira 2011b] Miranda, J.L., Teixeira, A.P. (2011) Portuguese Mathematics Action Program: are Engineering Education needs fulfilled?. In *Proceedings of ICEE2011-International Conference on Engineering Education (University of Ulster, Belfast, UK, August-2011)*, 10pp.

[Moscinski and Rutkowski 2010] Moscinski, J., Rutkowski, J. (2010) *Abstract Book of the ICEE2010 Conference. Silesian University of Technology, Gliwice, Poland.*

[Moscinski and Zielinki 2005] Moscinski, J., Zielinki, W. (2005) *Abstract Book of the ICEE2005 Conference, Gliwice, Poland.*

[Nagy and Miranda 2012] Nagy, M., Miranda, J.L., (2012) *AComputer Application For Interactive Teaching Of Decision Making Methods.* In *WCIT2012 - 3rd World Conference on Information Technology (University of Barcelona, Barcelona, Spain, November-2012)*, 6pp.

[Ólafsson and Kinnunen 2006] Ólafsson, S., Kinnunen, T. (2006) *Book of Abstracts of the 21st Conference on Operational Research.. University of Iceland, Reykjavik, Iceland.*

[Paixão and Martello 2010] Paixão, J.P., Martello, S. (2010) *Book of Abstracts of the 24th Conference on Operational Research. University of Lisbon, Lisbon, Portugal.*

[Pato and Oliveira 2006] Pato, M.V., Oliveira, P. (2006) *Livro de Resumos do IO2006-12. Congresso da APDIO. Universidade Técnica de Lisboa, Lisboa, Portugal.*

[Pelegrín and Torregrosa 2009] Pelegrín, B., Torregrosa, F. (2009) *Livro de Resumos da SEIO2009 - XXXI Congreso Nacional de Estadística e Investigación Operativa. Departamento de Estatística e Investigación Operativa da Universidad de Murcia, Murcia, Espanha.*

[Pesch and Woeginger 2009] Pesch, E., Woeginger, G.J. (2009) *Book of Abstracts of the 23rd Conference on Operational Research. University of Siegen, Bonn, Germany.*

[Ramos 2007] Ramos, A., (2007) *Livro de Resumos da SEIO2007 - XXX Congreso Nacional de Estadística e Investigación Operativa. Departamento de Estatística e Investigación Operativa da Universidade de Valladolid, Valladolid, Espanha.*

[Rasteiro and Furtado 2007] Rasteiro, M.G., Furtado, C.S. (2007) *Abstract Book of the ICEE2007 Conference. University of Coimbra, Coimbra, Portugal.*

[Rodríguez 2006] Rodríguez, J. (2006) *Livro de Resumos da SEIO2006 - XXIX Congreso Nacional de Estadística e Investigación Operativa. Departamento de Estadística, Investigación Operativa y Computación da Universidad de La Laguna, Tenerife, Espanha.*

[Ranyard 2011] Ranyard, J. (2011) Some Overall Results from the Global OR Practice Survey. In *IFORS2011. Melbourne, Victoria, Australia.*

[Salgueiro et al.2009] Salgueiro, C., Yánez, T., Cid, G. (2009) *Actas da SGAPEIO2009 - IX Congreso Galego de Estatística e Investigación de Operacións. Departamento de Estatística e Investigación Operativa da Universidade de Vigo, Ourense, Espanha.*

[Sakalauskas and Christiansen 2012] Sakalauskas, L., Christiansen, M. (2012) *Book of Abstracts of the 25th Conference on Operational Research. Vilnius, Lithuania.*

[Sellero and Jurado 2007] Sellero, C., Jurado, I. (2007) *Actas da SGAPEIO2007 - VIII Congreso Galego de Estatística e Investigación de Operacións. Facultade de Matemáticas da Universidade de Santiago de Compostela, Santiago de Compostela, Espanha.*

[Siskos and Bisdorff 2004] Siskos, Y., Bisdorff, R. (2004) *Book of Abstracts of the 20th Conference on Operational Research. Rhodes, Greece.*

[Sousa and Fontes 2005] Sousa, B., Fontes, F., (2005) *Actas da SGAPEIO2005 - VII Congreso Galego de Estatística e Investigación de Operacións. Escola de Ciências da Universidade do Minho, Guimarães, Portugal.*

[Uña 2003] Uña, J. (2003) *Actas da SGAPEIO2003 - VI Congreso Galego de Estatística e Investigación de Operacións. Vigo, Espanha.*

[Teixeira and Captivo 2008] Teixeira, A.P.T., Captivo, M.E. (2008) *Livro de Resumos do IO2008-13 Congresso da APDIO. Universidade de Trás-os-Montes e Alto Douro, Vila Real, Portugal.*

[Teixeira et al. 2011] Teixeira, A.P., Nascimento, M.M., Monteiro, H. (2011) *Programação Linear: uma experiência na Escola de Verão de Matemática.* In *Livro de Resumos do IO2011-15 Congresso da APDIO (Coimbra, Portugal, 2011)*, 1-12.

[Tobin and Leung 2011] Tobin, P., Leung, J. (2011) *Proceedings of the IFORS 2011 Conference, Melbourne Convention and Exhibition Centre, Melbourne, Australia..*

[Uhomoibhi 2011] Uhomoibhi, J. (2011) *Abstract Book of the ICEE2011 Conference. University of Ulster, Belfast, Northern Ireland, United Kingdom.*

[Vásquez-Espinosa and Vélez-Arocho 2006] Vásquez-Espinosa, R.E., Vélez-Arocho, J.I. (2006) *Abstract Book of the ICEE2006 Conference. University of Puerto Rico, San Juan, Puerto Rico, USA.*

[Vilar Fernández 2010] Vilar Fernández, J. M., (2010) *Livro de Resumos da SEIO2010 - XXXII Congreso Nacional de Estadística e Investigación Operativa. Departamento de Matemáticas da Universidade de Coruña, Coruña, Espanha.*

Framework for performance assessment of wind farms

Clara Bento Vaz*, Ângela Paula Ferreira†

* *Escola Superior de Tecnologia e Gestão, Instituto Politécnico de Bragança, Portugal*
E-mail:`clvaz@ipb.pt`

† *Escola Superior de Tecnologia e Gestão, Instituto Politécnico de Bragança, Portugal,*
E-mail:`apf@ipb.pt`

Abstract

This study develops a framework to provide insights regarding the performance of the farms of an energy player in the Portuguese wind sector. Firstly, the Data Envelopment Analysis is used to measure the efficiency of wind farms in producing electrical energy from the resources available and exogenous variables, during operating stage. This analysis enables the identification of the best practices of the efficient farms which can be emulated by inefficient ones. Secondly, Malmquist index is used to evaluate the changes in wind farms productivity. Bootstrap procedures are applied to obtain statistical inference on the efficiency estimates.

Keywords: Data Envelopment Analysis; Wind farm; Benchmarking; Bootstrap.

1 Introduction

This study proposes a framework based on Data Envelopment Analysis (DEA) to provide insights regarding the performance of the farms of an energy player in the Portuguese wind sector. DEA has been accepted as an important approach for performance assessment and benchmarking in several sectors. Zhou *et al.* [2008] presented a survey of DEA in energy and environmental modeling, from which benchmarking of electricity utilities accounts for the large number of studies although it does not include any application in wind energy sector.

In a Policies Scenario, taking into account both existing policies and declared intentions by countries, world primary energy demand is projected to increase by 1.2% per year, on average, between the current year and 2035. Electricity demand is projected to grow by a higher rate, 2.2% per year, given that it is expected that applications, formerly based on chemical energy, will be based on electrical energy in the following decades [IEA, 2010]. In this context, the share of world electricity generation from renewable sources is projected to tripling in the same period. Policies are being implemented to enhance the transition toward low-carbon technologies in the power sector, in which wind energy plays an essential role. According to Global Wind Energy Council, installed wind capacity has grown to accumulative worldwide installation level of 200 GW from which 38 GW had been installed in 2010. Europe is still the largest wind energy generator, despite the fact that other markets (*e.g.* USA, India, China) have also launched in recent years. Portugal accounts for about five percent of the wind energy installed capacity of the European Union, with approximately 4 GW of accumulated installed capacity in 2011 which is capable to generate about 15% of electricity consumed.

Several factors contributed to the development of the wind sector in Portugal. Since 2002, the implementation of a legal stable framework by the Portuguese government and several support schemes implemented by the European Commission have promoted the penetration of electricity produced from renewable energy sources [Martins *et al.*, 2011]. Despite technology's potential and investments in the clean energy market, the progress is too slow on attending outlined targets. The main reasons for the slow progress are related with a low share of energy-related investment in R&D activities, high investments when compared with thermal based electricity, uncertain time for the return of the capital invested, technical transmission system limitations and environmental impact.

There are ten main wind farm promoters acting in the wind energy sector in Portugal, with farms connected to the transmission or distribution grid system. Each promoter is concerned, among appropriate financial management, to ensure the maximum energy generation, with the highest availability rates and cost-effectiveness in terms of operation and maintenance. In this context, the development of performance assessment methodologies in the portfolio of a given promoter allows the identification of the best practices in operating stage in order to be emulated by inefficient farms. The use of DEA can contribute to enhance those methods through assessment of the potential for efficiency improvements and exploring their productivity change over time, as the emergent interest on productivity growth in electrical utilities. This is explored by using Malmquist index which can be decomposed in efficiency change and technological change. The efficiency change can be associated to internal operating practices observed in each farm, while the technological change can be related to specific conditions in which farms have to operate, for instance, the level of wind availability in each year.

In literature, the studies which use frontier methods to assess the performance of wind farms are scarce. Pestana Barros and Sequeira Antunes [2011] use Stochastic Frontier Analysis (SFA) to assess efficiency of Portuguese wind farms from different promoters. Outputs are measured by produced energy and capacity utilization, and the inputs are price for labor and capital invested proxied by the book value of physical assets. Findings of this study are that Portuguese wind farms' operational activity is affected by heterogeneous factors, farm size, managerial practices and ownership by energy companies, which have an impact on the efficiency. Iglesias *et al.* [2010] use DEA and SFA methodologies to measure the efficiency of a group of wind farms located in Spain. Models are output oriented concerning the produced energy, based on a relationship between capital, labor and fuel, similar to a conventional electricity generation technology. Capital factor is evaluated by the installed capacity in each farm and labor factor considers the number of fulltime employers responsible for operation, control and maintenance of the farms. Concerning fuel input, it is estimated based on the wind power incident per unit time on the interposed surface of the wind turbines of the farm and the annual average wind speed at each site. These two studies focus on the efficiency assessment of wind farms and argue the importance to model the non-discretionary factors regarding the wind in each farm.

This paper intends to improve the existing methodologies in performance assessment of wind farms from a given promoter to provide insights concerning their efficiency and productivity growth over time. Findings of this study can support the decision maker in benchmarking wind farms during operating stage and in repowering or overpowering processes. In a first stage, DEA is used to measure the operating efficiency of the wind farms and establishing benchmarks, followed by a second stage, were changes in wind farms productivity are investigated using the Malmquist index over two years, in which the wind energy sector suffered a considerable decrease in electric energy produced. The robustness of the scores achieved by DEA models can be tested by using bootstrapping frameworks [Simar and Wilson, 1998, Simar and Wilson, 1999]. Finally, the proposed framework is applied to a case study, giving insights of the performance assessment of wind farms from Iberwind which has a market share of 18% on the Portuguese wind energy sector.

This study is organized as follows: next section presents the methodology for performance assessment of wind farms, section 3 presents context setting in the wind energy sector and applies the methodology to the case study, and section 4 presents main conclusions of this study.

2 Performance assessment methodology

The methodology proposed in this study intends to explore the productivity and the efficiency of wind farms. In a first stage, DEA is used to assess the farms efficiency by taking into account the resources and the non-discretionary variable, the wind, available in each farm to generate electric energy. This approach enables benchmarking among farms. The robustness of efficiency scores is tested by using bootstrap framework [Simar and Wilson, 1998]. In a second stage, we use panel data to assess the overall productivity change over time of the farms by using the Malmquist index and its components [Färe *et al.*, 1994], efficiency change and technological change. The efficiency change measures if the farm is moving closer or farther from the frontier while the technological change measures shifts in the frontier that can be characterized by progression, regression, or both. Finally, the robustness of these indexes is tested by using bootstrapping [Simar and Wilson, 1999] which allows the identification of significant aspects that may explain the performance of each farm over time. Following sections present the proposed methodology in detail.

2.1 DEA Model

DEA is a linear programming based technique to assess the relative efficiency of an homogeneous set of Decision Making Units (DMUs) in producing multiple outputs from multiple inputs. This allows to identify the "best practices DMUs" and their linear combination defines the frontier technology. By reference to this frontier, a single summary measure of efficiency is calculated for each DMU. In the original DEA model proposed by Charnes *et al.* [1978], the efficiency score of each DMU is estimated by using the frontier technology characterized by Constant Returns to Scale (CRS). For an output oriented analysis, we consider a technology involving n production units defined by j $(j = 1, \ldots, n)$, which use the inputs x_{ij} $(x_{1j}, \ldots, x_{mj}) \in R_{+}^{m}$, to obtain the outputs y_{rj} $(y_{1j}, ..., y_{sj}) \in R_{+}^{s}$, *i.e.*, the production possibility set. In this model, the efficiency of each DMU j_o is given by the reciprocal of the factor (θ) by which the outputs of the DMU j_o can be expanded, according to the following linear model:

$$\max\Big\{h_{j_o} = \theta |\ x_{ij_o} \geq \sum_{j=1}^{n} \lambda_j\ x_{ij}, \quad i = 1, \ldots, m \qquad (2.1)$$

$$\theta\ y_{rj_o} \leq \sum_{j=1}^{n} \lambda_j\ y_{rj}, \quad r = 1, \ldots, s$$

$$\lambda_j,\ \geq 0, \quad \forall_{j,i,r}\Big\}$$

Model (2.1) assesses the relative efficiency of DMUs in the achievement of the output levels given the resources used. The measure of efficiency, given by $1/\theta^*$, equals to 100% when the unit under assessment is efficient, whereas lower scores indicate the existence of inefficiencies. For the inefficient units there is evidence that it is possible to obtain higher levels of outputs with the same or lower levels of the inputs currently used. For these units, it is also possible to obtain, as by-products of the DEA efficiency assessment, a set of targets for becoming efficient. The input and output targets for a DMU j_o under assessment are obtained as follows:

$$x_{ij_o}^{o} = x_{ij_o} - s_i^* = \sum_{j=1}^{n} \lambda_j^*\ x_{ij}$$

$$y_{rj_o}^{o} = \theta_o^*\ y_{rj_o} + s_r^* = \sum_{j=1}^{n} \lambda_j^*\ y_{rj} \qquad (2.2)$$

where the variables s_i^* and s_r^* are the slacks corresponding to the input i and output r constraints, respectively, obtained at the optimal solution of model (2.1). The benchmarks for the inefficient DMUs j_o are the units with values of $\lambda_j^* > 0$ in the optimal solution of model (2.1). These are the Pareto-efficient DMUs.

The model (2.1) enables to assess the Technical Efficiency (TE) for each DMU which can be due to the ineffective operation of the production process in transforming inputs to outputs and also due to the divergence of the entity from the Most Productive Scale Size (MPSS). Banker *et al.* [1984] propose the DEA model that assesses the Pure Technical Efficiency (PTE) for each DMU by using the frontier technology characterized by Variable Returns to Scale (VRS). This model is achieved by including the constraint $\sum_{j=1}^{n} \lambda_j = 1$ in model (2.1). The divergence of the DMU from the MPSS is given by scale efficiency, which is determined by the ratio $\frac{TE}{PTE}$.

For Pareto-efficient DMUs it is possible to identify the local Returns to Scale (RTS) which enables to know if there are advantages in changing the scale of DMUs. In the study case under analysis, this information is very useful in repowering processes of wind farms. If increasing returns to scale hold at a Pareto-efficient DMU, then increasing its input levels by a given percentage will lead to expansion of its output levels by a larger percentage. This indicates that the DMU should increase its scale size. Similarly, a DMU operating at a point where decreasing returns to scale hold, it should decrease its scale. If a DMU operates at constant returns to scale point, its size is considered optimal. The Färe *et al.* [1985] approach is used to characterize the RTS of Pareto-efficient wind farms.

The DEA models enable to perform the benchmarking analysis of wind farms which can be used to support the decisions makers in management of the units observed. Although, these results should be explored with carefulness, since DEA assumes that the distance from each unit and the projected point in the frontier corresponds to inefficiency, which makes this method sensible to random noise in data. For this, we propose an integrative analysis in which the robustness of θ derived from model (2.1) is tested by using bootstrapping method proposed by Simar and Wilson [1998]. Thus, we derive for each DMU a confidence interval for θ, the bias and the bias-corrected score $\hat{\hat{\theta}}$. These scores are used to assess the wind farms performance.

2.2 Malmquist index on evaluation of overall productivity

In energy sectors, it is of great interest the investigation of the productivity change over time. The Malmquist productivity index was introduced by Caves *et al.* [1982] and developed further in the context of performance assessments by Färe *et al.* [1994] for conducting performance comparisons of DMUs over time. The high popularity of this method is related with several factors. Firstly, the index is applied to the measurement of productivity change over time, and can be decomposed into an efficiency change index and a technological change index. Secondly, it is not necessary to use price data, assumptions of cost minimization or revenue maximization. Thirdly, it can be used in case of oriented and non-oriented analysis. Fourthly, it enables the determination of the total factor productivity in the generic case where production technology uses multiple inputs to produce multiple outputs by deriving efficiency scores in DEA models.

The Malmquist index, as proposed by Färe *et al.* [1994], is used to derive the overall productivity of each DMU. It is based on radial measures which are defined by distance functions. In output-oriented analysis, the output distance function is equal to the efficiency score estimated by model (2.1), given by $1/\theta^*$ for each DMU for a given period. Considering n DMUs in period t, which use the inputs $x^t \in R_+^m$ to obtain the outputs $y^t \in R_+^s$, and the same n DMUs in period $t+1$, which use the inputs $x^{t+1} \in R_+^m$ to obtain the outputs $y^{t+1} \in R_+^s$. To simplify the notation, the efficiency score estimated for each DMU_{j_o} in period t is given by $E_o^t(t)$ while the efficiency score estimated for each DMU in period $t+1$ is given by $E_o^{t+1}(t+1)$. Thus, the score in parenthesis represents the period in each DMU is assessed while the superscript denotes the frontier technology used as reference. The Malmquist index derived for each DMU_{j_o} is calculated as follows:

$$I_o^{t+1,t} = \left(\frac{E_o^t(t+1)}{E_o^t(t)} \frac{E_o^{t+1}(t+1)}{E_o^{t+1}(t)} \right)^{\frac{1}{2}} \tag{2.3}$$

In terms of interpretation, a score of $I_o^{t+1,t}$ greater than one indicates better performance in period $t+1$ than in period t.

The mixed-period distance functions, $E_o^t(t+1)$ and $E_o^{t+1}(t)$, can be greater, equal or lower than 1. For example, the distance function derived to the period $t+1$ for a DMU observed in period t can be lower or equal to 1 if the input-output vector of this DMU belongs to the Production Possibility Set (PPS) of period $t+1$. This occurs for $E_o^t(t)$ and $E_o^{t+1}(t+1)$ cases. In opposite, the distance function derived to the period $t+1$ for a DMU observed in period t is higher than 1, if the input-output vector of this DMU is outside the PPS of the period $t+1$.

According to Färe *et al.* [1994], this index can be decomposed in two components: $IE_o^{t+1,t}$ and $IF_o^{t+1,t}$. The sub-index $IE_o^{t+1,t}$ corresponds to efficiency change and compares the efficiency spread between the periods observed for each DMU_{j_o}. The sub-index $IF_o^{t+1,t}$ corresponds to technological change and compares the relative position of the frontiers associated to periods t and $t+1$ for the input-output mix of each DMU_{j_o} observed. This decomposition implies that the sources of better performance can be associated with two factors: less dispersion in the efficiency score of DMU in each period and/or better productivity associated to the period frontier.

The efficiency change derived for each DMU_{j_o} is calculated according to:

$$IE_o^{t+1,t} = \frac{E_o^{t+1}(t+1)}{E_o^t(t)} \tag{2.4}$$

A value of index $IE_o^{t+1,t}$ greater than 1 means that efficiency spread is smaller in DMU_{j_o} observed in period $t+1$ than that observed in period t, measuring how much the DMU_{j_o} is getting closer (catch up) or farther from the frontier.

Concerning the technological change derived for each DMU_{j_o}, it is given by:

$$IF_o^{t+1,t} = \left(\frac{E_o^t(t)}{E_o^{t+1}(t)} \frac{E_o^t(t+1)}{E_o^{t+1}(t+1)} \right)^{\frac{1}{2}} \tag{2.5}$$

When $IF_o^{t+1,t}$ is higher than 1, this means that the productivity of frontier $t+1$ is better than the productivity of frontier t, which implies that the frontier has progressed. This index can be seen as an average aggregated change in technology of a DMU_{j_o} since it is obtained as the geometric mean of two components. The first component $(\frac{E_o^t(t)}{E_o^{t+1}(t)})$ corresponds to the distances between the frontiers t and $t+1$ when assessed for the DMU_{j_o} observed in period t. The second component $(\frac{E_o^t(t+1)}{E_o^{t+1}(t+1)})$ is calculated in a similar way for the same DMU observed in period $t+1$.

It is possible to analyze globally the relative position of the two frontiers, which enables to identify if the frontiers have regressed, progressed or crossed over. To do so, it is necessary to analyze each component of $IF_o^{t+1,t}$ for all DMUs observed in the periods under analysis. Some typical situations may occur: for instance, if the component is always higher than 1, this means that there has been a progression in the technology; on the other hand, if the component is always lower than 1, this means that there has been a regression in the technology; in the case when there are at least one component higher than 1 and one component lower than 1, this indicates that frontiers are crossed over, signifying that for some input-output mix the frontier progressed and for others, the frontier regressed.

The bootstrapping framework proposed by Simar and Wilson [1999] is used to evaluate the robustness of the estimates of $I_o^{t+1,t}$, $IE_o^{t+1,t}$ and $IF_o^{t+1,t}$ (hereafter *I*, *IE* and *IF*, respectively) obtained for each DMU_{j_o}, which allows the computation of confidences intervals for each index. If the interval contains the value 1, we cannot infer that significant changes occurred in the DMU_{j_o}. On the other hand, if the lower and upper bounds are smaller (or higher) than 1, this implies that there was a decline (or progress) in the DMU_{j_o}. This approach is currently used in several studies [Gilbert and Wilson, 1998, Tortosaausina *et al.* 2008, Odeck, 2009, Horta *et al.*, 2012]. This analysis is extended to the components of IF for all DMUs observed to find out the relative position of the frontiers.

3 Performance assessment of wind farms

This section applies the methodology proposed in previous section to evaluate the performance of wind farms owned by Iberwind in Portugal. This study focus in the wind farms efficiency analysis during operating stage, for a given distribution of wind speed in the location of the farms, installed capacity and number of wind turbines, oriented to the maximization of the output electric energy produced.

3.1 Contextual setting

Relevant decisions and factors that affect the productivity of wind farms are prior to start-up, including for instance wind farm location and engineering design process such as the installed capacity, type of generator and turbine layout. The focus of wind farms performance assessment is during operating phase, when wind farms perform the energy conversion and it is delivered to the grid. Even though, the performance of a wind farm is closely linked to prior start-up phase, the operating stage is relevant throughout estimated lifetime of the assets from the point of view of maximizing the energy production, ensuring the highest availability rates and cost-effective operation and maintenance.

Wind is a variable source of power: output rises and falls as wind strength fluctuates in a hourly or 10 minutes time scale although it is consistent from year to year. This variability poses a challenge when integrating wind power into grids, especially when wind becomes a major component of the total system. Wind speeds suitable for electricity generation range from 5 m/s (cut in) to 25 m/s (cut out). The frequency of wind speeds usually fits a Weibull distribution function and an average value does not relate to the amount of energy a wind farm can produce. Installed capacity and the number of wind turbines in a farm, along with the variability of wind, relate to the capacity factor of a wind farm, i.e., the ratio of actual productivity in a year to its theoretical maximum. The rated power of a unit of the wind farm (given by the ratio of the installed capacity and the number of wind turbines) if small, could lead to an higher capacity factor but the farm may not be able to produce energy at higher wind speeds, which translates in less profit. On the other hand, if the rated power of units is high, the turbine may stall at low wind speeds and the extra power at high wind speeds may not compensate the higher costs of the turbine. Therefore, these resources are important to assess efficiency and productivity analysis during operating phase and may provide useful information in repowering or overpowering processes. Concerning the output, it should be point out that electric energy produced from wind is not constrained by load demand or other market players, as currently regulated.

3.2 DEA model

The DMU is formed by a group of wind turbines connected to the transmission or distribution grid utility. The number of DMUs under analysis is 31, spread out in North and Center of Portugal. The final data set considers 30 wind farms since one of them was eliminated due to a repowering process that began in 2010. Total capacity installed ascends to 683.75 MW through 319 wind turbines, from 15 different models, provided by five manufacturers (Vestas, Nordex, Enercon, GE and Winwind). We consider a panel data set regarding 2010 and 2011 years collected from Annual Reports and Accounts1[1]. The 30 farms under analysis are located in six wind typical locations in Portugal (Bragança, Vila Real, Viseu, Coimbra, Leiria, Lisboa).

The wind farms can be considered homogenous as they result from similar set-up stages and use a similar generation process. The strong disposability of inputs and output is adopted and the convexity of the frontier is valid, as any input combination inside the PPS determined by the wind farms sample is feasible. The output-oriented perspective is used, as the objective of the farms is to produce maximum electric energy, taking into account the non-discretionary variable, the wind, and the resources available in each farm, during the period under analysis. We use the constant returns to scale to assess the technical efficiency of wind farms observed. In order to model the farm activity, the input-output set should cover the full range of resources used and the outputs that are relevant for the objectives of the analysis [Dyson *et al.*, 2001]. The output corresponds to the amount of electric energy delivered to the grid. Concerning the inputs, we consider installed power, number of turbines and wind availability. The descriptive measures concerning inputs and output under analysis are summarized in Table 1. Installed power capacity of the farms is determined by the number of wind turbines multiplied by the rated power of each one. The number of turbines relates with the area occupied by the farm. To capture the effect of the wind variability into the model, we consider the number of hours per year that wind speed is within the range defined by cut in and cut out speeds (hereafter named wind hours). For each wind farm location, the wind data is collected from a meteorological data base throughout identification of the station which represents its wind profile, defined by the nearest meteorological station. The inclusion of this non-controllable input assures that a farm with unfavorable conditions regarding wind resource is not penalized in the performance assessment. This is an internal non-discretionary input because it is used for the definition of the PPS, according to Camanho *et al.* [2009]. Data concerning maintenance and operation costs are confidential and, for this reason, they are not included in the model.

Table 1: Mean and standard deviation values for inputs and output of wind farms

	2010		2011	
	Mean	Stand. Deviat.	Mean	Stand. Deviat.
Inputs				
Installed power (MW)	22.4	30.4	22.4	30.4
No of wind turbines	10.4	10.7	10.4	10.7
Wind hours	3773.6	1082.4	3280.5	980.4
Output				
Electric Energy (GWh)	56.5	81.5	51.1	75.8

The standard deviation of observed variables is quite high relative to the mean, indicating a considerable amount of diversity in the wind farms.

The summary of the technical efficiency estimates, using the formulation in model (2.1) is presented in Table 2. The robustness of these estimates is tested by calculating the bias corrected efficiency scores (as the inverse of $\hat{\hat{\theta}}$) [Simar and Wilson, 1998] which summary results are presented in Table 2.

Table 2: Summary results of original and bootstrapped efficiency scores

	Year 2010			Year 2011		
	Bias corrected eff.	Bias	Efficiency estimated	Bias corrected eff.	Bias	Efficiency estimated
Mean	72.92%	-8.00%	77.73%	66.33%	-10.54%	73.97%
Stand. Deviat.	10.13%	4.30%	12.33%	10.86%	4.92%	13.44%

The efficiency scores estimated are relative, since the farms in a given year are only compared with all farms in the sample operating in the same year. We may observe that the wind farms analyzed are more homogenous in 2010 than in 2011 which is confirmed by bootstrapping analysis. The absolute value of bias is slightly higher in 2011 due to the differences between bias-corrected efficiency scores, and efficiency estimates are higher in 2011 year. Globally, this indicates that farms moved farther from the frontier. This effect is captured by the analysis of efficiency change index (IE) for each farm, which is explored in productivity analysis. Figure 1 presents the average of bias corrected efficiency of the farms located in the same region for both years. We confirm that the level of efficiency spread increased in 2011 for all regions but we observe some differences among them.

[1]The constraint of the panel data is limited to 2010 and 2011, because there is missing wind data in former years from meteorological stations.

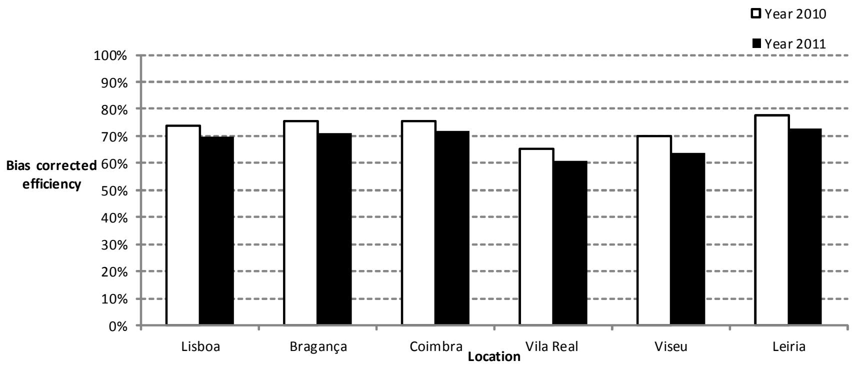


Figure 1: Comparison of the average of bias corrected efficiency of the farms located in the same region for 2010 and 2011 years

3.3 Benchmarking analysis

The benchmarks farms and their best practices should be identified in order to be emulated by inefficient units. These practices may be related to the use of more efficient wind turbines, enhanced wind farm design and layout, better operation and maintenance schemes, which may be used to support the inefficient farms to achieve the appropriate targets. It is also important to identify the nature of returns to scale of the Pareto-efficient farms to explore changes in their size.

From the sample used, there are only 3 efficient farms: Achada, Candeeiros and Pampilhosa. These farms maintain the efficiency status in both years. In 2011, Achada and Candeeiros are the benchmarks which are used as reference 27 and 22 times, respectively. There are no units which are compared with Pampilhosa, since this farm is the largest unit in terms of number of wind turbines and installed power. In the following, we explore the profile of the benchmarks in terms of location and type of wind turbines used.

Benchmarks are located in areas with high wind potential (Lisboa, Leiria and Coimbra) and their energy conversion system is based on asynchronous generators. The wind turbines of Achada are produced by Nordex while the wind turbines of Candeeiros and Pampilhosa are produced by Vestas. These farms are the largest ones while Achada is a smaller farm. Figure 2 compares the age, inputs and output of benchmarks with those observed in inefficient farms, in 2011 (the some profile occurs in 2010). In this graph the scores were normalized by the average scores observed in benchmarks to simplify the comparison. The installed power of inefficient units is, in average, 80% less of that observed in benchmarks and the electric energy produced follows a decrement of the same magnitude. The inefficient units have, in average, 67% less number of wind turbines of those observed in benchmarks. Given that wind hours in geographical areas where inefficient units are located have a small reduction (about 14%), this suggests inefficient farms are prone to a repowering process, in order to increase their electric energy, as they are not exploring all wind energy potential. The fact that inefficient farms are, in average, 15% older than benchmarks can explain some inefficiency in some farms.

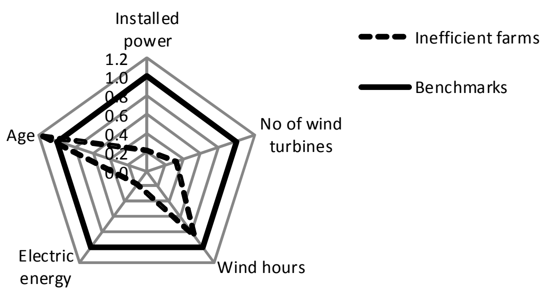


Figure 2: Comparison between benchmarks and inefficient farms in 2011

In both years, the most inefficient unit is the same farm: the Lomba Seixa I with scores equal to 56.7%, in 2010, and 49.4%, in 2011. The lowest score can be due to age of technology of the conversion system installed in this farm, as it is 11 years old. This farm has 11 wind turbines which were produced by Nordex.

The results also indicate that all inefficient farms have slack in the constraint relative to number of wind turbines. Conversely, there is no farm with slack in installed power. Thus, the inefficient farms would increase the energy produced by using a lower number of wind turbines and the actual installed power. This upshot is important in repowering processes. Although, it is important to use a relevant number of turbines to catch the wind potential in a given location, these results suggest that wind farm design could be enhanced and used to decrease the environment impact of future wind farms projects. There are only four farms which have slack relative to available wind hours in both years (Borninhos, Jarmeleira, Rabaçal and Serra Escusa). This may be related with inefficient scheduling of maintenance operations, in times with high wind potential.

These findings should be explored and discussed with the promoter, in order to enhance performance of the wind farms. For inefficient units, it is possible to specify appropriate targets based on internal benchmarking, as proposed in the next section.

3.3.1 Target setting

For each inefficient farm, we can define targets for performance improvement. These targets are determined by linear combination of the benchmarks for each inefficient unit. For example, the technical efficiency of Farm 17 (Lousã I) is about 67%. The scores for 2011 period, regarding inputs and output of farm 17, DEA targets (determined by (2.2)) and peers, are presented in Table 3.

Table 3: Target setting of Farm 17

			Benchmarks	
			Farm 1	**Farm 7**
	Observed	Target	$\lambda = 0.213$	$\lambda = 0.302$
Installed Power (MW)	35	35	6.9	111
No of wind turbines	14	11.8	3	37
Wind hours	2598	2598	5397	4787
Electric Energy produced (GWh)	71.7	107.1	22.4	338.9

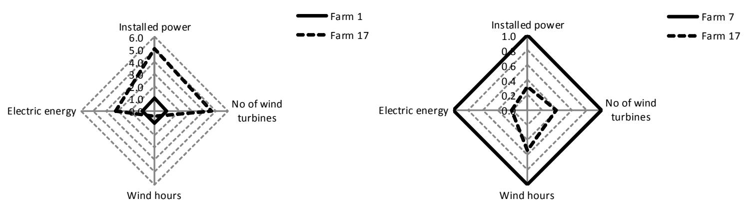


Figure 3: Comparison between actual values of Farm 17 with benchmarks 1 and 7

The target for a given variable (input or output) of farm 17 is defined by the linear combination of 21.3% of the score observed in farm 1 (Achada) and 30.2% of the score observed in farm 7 (Candeeiros). Farm 17 is larger than farm 1 and smaller than farm 7. Targets indicate that it is possible to increase the electric energy to 107.1 GWh by using the actual installed power with the same wind hours available in 2011, with a slack roughly equal to 2 turbines. In theory, the ratio between the total installed power and the number of turbines should be increased, via increment of the rated power of each turbine. We can compare the actual inputs and output observed in farm 17 with each benchmark by using the radar graphs in Figure 3, where the scores were normalized by those observed in benchmarks to simplify the comparison. Farm 17 has an installed power and a number of wind turbines which are almost 5 times higher, but 48% less wind hours than those observed in benchmark 1. Taking into consideration the exogenous characteristic of the input wind hours, thus non-controllable, it is not evident a possible increase in electric energy production. On the other hand, farm 17 has a installed power and a number of wind turbines about 70% and 60%, respectively, less than those are observed in benchmark 7 while the reduction in the wind hours is about 50%. We can conclude that inputs of farm 17 are, in average, 60% lower than those observed in benchmark 7 and a similar percentage of reduction in electric energy would be expected and not a decrease of 80%, as observed. From the comparison with peers, namely wind farm 7, the farm 17 could produce higher level of electric energy from the resources observed. Hence, it is necessary to identify the best practices observed in benchmarks 1 and 7 which should be emulated by inefficient farm 17.

As the production technology of wind farms is characterized by constant returns to scale, the farm efficiency score, $1/\theta^*$, includes sources related with resources under-utilization and scale size. Next, we explore the scale of the farms based on internal benchmarking.

3.3.2 Exploring changes in wind farms size

As the scale size affects the productivity of a DMU, it is important to calculate the scale efficiency to measure the distance between CRS and VRS frontier technologies at the scale size of the assessed unit. So, the larger the difference between TE and PTE efficiency scores, the lower the value of scale efficiency is, and the adverse impact of scale size on productivity is more significative. The average scale efficiency score is, in average, 95.38%, and 93.31%, in 2010 and 2011, respectively. This means that scale size only affects the productivity of a small proportion of units observed (Jarmeleira, Borninhos, Rabaçal, Chiqueiro, Malhadizes, Degracias, Lousã I, Lousã II, Malhadas), where the scale efficiency has the lowest scores, with a range between 73.4% and 88.9%. This strengthens using constant returns to scale frontier technology to assess the wind farms efficiency.

The analysis of local returns to scale according to Färe *et al.* [1985] shows that Achada, Candeeiros and Pampilhosa are characterized by an optimum size. Jarmeleira, Lousã II, Malhadas and Rabaçal have increasing returns to scale, which indicates that the size of these units could be increased with a repowering process which enables increasing their productivity. There is no unit where we observe decreasing returns to scale, which indicate that there is no one with higher size than the required, taken into account the level of electric energy produced.

3.4 Productivity analysis

In a second stage, we investigate the productivity of wind farms by disentangling the efficiency change and technological change effects observed in wind farms in 2010 and 2011 years. An aggregate analysis is performed by identifying the global effects which had occurred in the period under analysis. Changes in efficiency (IE), technology (IF) and productivity (I) indexes of farms are explored through identification of scores higher, lower or equal to 1 which correspond to improvement, deterioration or stagnation, respectively. This analysis is complemented with bootstrapping framework, as proposed by Simar and Wilson [1999], to identify if those changes, for each farm, are significant. Table 4 aggregates the results in terms of number of wind farms which improve, decline or maintain the performance for each index.

Table 4: Significant scores for I, IE and IF

	I	IE	IF
Improvement	1 (in 1)	1 (in 3)	-
Deterioration	27 (in 29)	19 (in 24)	17 (in 30)
Stagnation	2	10(3)	13

We observed that 27 farms decreased overall productivity levels in year 2011, as indicated by significant scores of I index. This effect is mainly due to deterioration in the productivity levels of the frontier for some inputs-output mix and decreasing efficiency levels in some farms. Only Serra Escusa improves overall productivity level due to improvement on its efficiency in 2011. Pampilhosa and Candeeiros maintain overall productivity levels in 2011.

There are 19 farms that moved farther from the frontier in 2011, as indicated by significant scores of IE index. These farms had the worst performance in 2011, so the reasons for that should be investigated. Only Serra Escusa moved closer to the best practices. It is recommended to identify how this farm carried out its operations and maintenance services in order to be emulated by the inefficient farms. The remaining farms maintained the efficiency spread levels observed in 2010.

Globally, the productivity of the best-practices frontier decreased considerably in 2011 for the input-output combinations of 17 farms, although for the remaining input-output mix, the productivity of the frontier maintained the level of productivity observed in 2010. This is connected with the decrease of produced electric energy observed in wind energy sector in 2011. Next, we explore the relative position of frontiers for the farms observed in each period. To do so, we analyze if the ratios of IF ($\frac{D^{2010}(2011)}{D^{2011}(2011)}$, $\frac{D^{2010}(2010)}{D^{2011}(2010)}$) are statistical significant throughout bootstrap framework [Simar and Wilson, 1999]. Table 5 aggregates the results in terms of number of wind farms which improve, decline or maintain the performance for each ratio.

Table 5: Significant scores for ratios of IF

	$E^{2010}(2010)/E^{2011}(2010)$	$E^{2010}(2011)/E^{2011}(2011)$
Improvement	0	0 (in 2)
Deterioration	14 (in 30)	18 (in 28)
Stagnation	16	12

The input-output combinations of 14 farms observed in 2010 are located in areas of the production possibility set where the productivity of the frontier declined. The remaining farms are located in areas of the PPS where the frontier maintained the productivity. During 2011, there are 18 input-output combinations of wind farms located in areas of the PPS where the frontier regressed, while the other remaining farms are located in areas where the frontier maintained the productivity. There is no statistical evidence of crossed frontiers for all input-output combinations.

4 Conclusions

This study proposes a methodology to assess the efficiency and productivity change of wind farms, which can support decision makers during operating phase of wind farms, and also in location and project design phases of new farms and in repowering processes. In first stage, the efficiency assessment of wind farms enables to identify benchmark profiles, set targets for inefficient units and explore the scale size of existing farms. These results can be useful in project design and layout of new farms and also in repowering processes. The second stage explores the efficiency and productivity over time of wind farms by identifying the global effects which occurred in terms of changes in internal practices observed and productivity of the frontier, during the established time frame.

Regarding the operating stage of the farms analyzed, 3 farms are the benchmarks, whose best practices can be related with well-performing operations and maintenance programs. Between 2010 and 2011, different profiles of wind farms were identified in terms of overall productivity change, efficiency change and technological change. Almost all farms decreased overall productivity levels, mainly due to the decline in the productivity levels of the frontier, which is in accordance with decrease in wind availability, measured in wind hours, observed in 2011. The productivity of the frontier declined for some input-output combinations observed in 2011 and for the other combinations, the frontier maintained its productivity. In this case, there is one farm that improved its overall productivity due to the improvement of its efficiency in 2011 and two farms which maintained overall productivity as they kept the efficiency levels. We observed also that 19 farms had the worst performance in 2011 which requires further investigation to reveal the reasons.

Further research should be conducted using a large panel data set in order to analyze the impact of wind availability on the productivity of wind farms. The inclusion of variables concerning the operation and maintenance schemes should also be explored in future performance assessments of wind farms.

References

[Banker *et al.*, 1984] Banker, R. D., Charnes, A. and Cooper, W. W. (1984). Some models for estimating technical and scale inefficiencies in Data Envelopment Analysis. *Management Science*, 30(9):1078–1092.

[Camanho *et al.*, 2009] Camanho, A. S., Portela, M. C. and Vaz, C. B. (2009). Efficiency analysis accounting for internal and external non-discretionary factors. *Computers & Operations Research*, 36(5):1591– 1601.

[Caves *et al.*, 1982] Caves, D. W., Christensen, L. R. and Diewert, W. E. (1982). The economic theory of index numbers and the measurement of input, output and productivity. *Econometrica*, 50:1393–1414.

[Charnes *et al.*, 1978] Charnes, A., Cooper, W. W. and Rhodes, E. (1978). Measuring efficiency of decision-making units. *European Journal of Operational Research*, 2(6):429–444.

[Dyson *et al.*, 2001] Dyson, R. G., Allen, R., Camanho, A. S., Podinovski, V. V., Sarrico, C. S. and Shale, E. A. (2001). Pitfalls and protocols in DEA. *European Journal of Operational Research*, 132:245–259.

[Färe *et al.*, 1985] Färe, R., Grosskopf, S. and Lovell, C. A. K. (1985). The measurement of efficiency of production. Kluwer Nijhoff Publishing, Boston.

[Färe *et al.*, 1994] Färe, R., Grosskopf, S., Lindgren, B. and Roos, P. (1994). Productivity developments in swedish hospitals: a Malmquist output index approach. In *Charnes, A., Cooper, W. W., Lewin, A., and Seiford, L., editors, Data envelopment analysis: theory, methodology and applications*: 253–272. Kluwer Academic Publishers, Boston.

[Gilbert and Wilson, 1998] Gilbert, R. A. and Wilson, P. W. (1998). Effects of deregulation on the productivity of Korean banks. In *Journal of Economics and Business*, 50(2):133–155.

[Horta *et al.*, 2012] Horta, I. M., Camanho, A. S., Johnes, J. and Johnes, G. (2012). Performance trends in the construction industry worldwide: an overview of the turn of the century. Journal of Productivity Analysis, 39(1):89–99.

[IEA, 2010] International Energy Agency, IEA (2010). World Energy Outlook 2010, IEA Publications, Paris, France.

[Iglesias *et al.*, 2010] Iglesias, G., Castellanos, P. and Seijas, A. (2010). Measurement of productive efficiency with frontier methods: A case study for wind farms. *Energy Economics*, 32(5):1199–1208.

[Martins *et al.*, 2011] Martins, A. C., Marques, R. C. and Cruz, C. O. (2011). Public-private partnerships for wind power generation: The Portuguese case. *Energy Policy*, 39:94–104.

[Odeck, 2009] Odeck, J. (2009). Statistical precision of DEA and malmquist indices: A bootstrap application to norwegian grain producers. In *Omega*, 37(5):1007–1017.

[Pestana Barros and Sequeira Antunes, 2011] Pestana Barros, C. and Sequeira Antunes, O. (2011). Performance assessment of portuguese wind farms: Ownership and managerial efficiency. *Energy Policy*, 39(6):3055–3063.

[Simar and Wilson, 1998] Simar, L. and Wilson, P. W. (1998). Sensitivity analysis of efficiency scores: How to bootstrap in nonparametric frontier models. *Management Science*, 44(1):49–61.

[Simar and Wilson, 1999] Simar, L. and Wilson, P. W. (1999). Estimating and bootstrapping Malmquist indices. *European Journal of Operational Research*, 115:459–471.

[Tortosaausina *et al.* 2008] Tortosaausina, E., Grifelltatje, E., Armero, C. and Conesa, D. (2008). Sensitivity analysis of efficiency and malmquist productivity indices: An application to spanish savings banks. *European Journal of Operational Research*, 184(3):1062–1084.

[Zhou *et al.* 2008] Zhou, P., Ang, B. W. and Poh, K. L. (2008). A survey of data envelopment analysis in energy and environmental studies. *European Journal of Operational Research*, 189(1):1–18.

Multi-period and multi-product inventory management model with lateral transshipments

Joaquim Jorge Vicente*, Susana Relvas*, Ana Barbosa-Póvoa*

* *CEG-IST-Centre for Management Studies, Instituto Superior Técnico, Technical University of Lisbon, Avenida Rovisco Pais 1049-001 Lisbon, Portugal*

E-mail:`joaquim.jorge.vicente@gmail.com`

E-mail:`susana.relvas@ist.utl.pt`

E-mail:`apovoa@ist.utl.pt`

Abstract

Inventory management plays an important role in supply chains. Through a correct inventory management policy, supply chains can close the gap created by the imbalance between supply and demand, eliminating costly supply chains. Such imbalance is often caused by differences on the supply and demand timings, quantities or locations of materials. Thus, inventory should be efficiently managed across the supply chain so as to guarantee that materials are available in the right location and quantities at the right time.

This paper aims to contribute to this goal and presents an Inventory Management (IM) policy implemented on a Mixed Integer Linear Programming (MILP) model that optimizes, in a dynamic setting, the flow of products through a multi-period and multi-product supply chain. Demands are received at the retailers who replenish their stock from the regional warehouses, which, in turn, are supplied by a central warehouse. Each retailer has a demand which may be represented by a normal distribution. Lateral transshipment is allowed among regional warehouses and among retailers.

In order to validate and compare the proposed policy against commonly used policies, the Continuous Review and the Periodic Review policies are modeled using the same approach and acting over the same system. The comparison of inventory management policies shows that the IM policy outperforms the classical policies in terms of material availability leading to an overall reduction of operational costs.

Keywords: Supply chain management; inventory management; mixed integer linear programming; continuous review; periodic review.

1 Introduction

Supply chain management deals with the organization of the flows of products and information throughout the supply chain structure so as to ensure the requirements of customers, while minimizing operating costs. To attain such goal, inventory management is, within supply chains, a key activity since it ensures the continuity and balance between supply and demand. In this context inventory management appears as an important and challenging problem, since decisions made by a chain member may affect, with different impacts, the remaining supply chain entities. Kautter (1999) presented new perspectives for corporative management on the inventory theory area and stated that inventories appear as the result of mismanagement of the different supply chain processes. The need to fully control the processes was identified by the author as a way of optimizing inventories.

Abdul-Jalbar et al. (2006) dealt with the classic deterministic one-warehouse multi-retailer inventory/ distribution system where customer demand rates were assumed to be known and constant and there was no backlog or lost sales. The retailers placed orders to satisfy customer demands generating demands at the warehouse. Order lead times were assumed to be instantaneous, so no lead time was considered. Costs at each facility consisted of a fixed charge per order and of a holding cost.

Ozdemir et al. (2006) studied the multi-location transshipment problem with capacitated transportation. They used a simulation approach that incorporate transportation capacity such that transshipment quantities between stocking locations are bounded due to transportation media or the location's transshipment policy.

Hsiao Yu-Cheng (2008) investigated the classic deterministic one-warehouse multi-retailer inventory/ distribution system. In this study the customer demand rates were assumed to be known and constant. Shortages were not permitted and lead times were assumed negligible. A method that reached the optimal solution in most of the instances studied was proposed. In the same year, Monthatipkul and Yenradee (2008) also developed an inventory control system for a one-warehouse multiple-retailers supply chain. They considered only one product. A mixed integer linear model was proposed to determine the optimal inventory and distribution plan that minimized total related costs. The efficiency of the inventory control system was compared to a periodic review policy.

More recently Paterson et al. (2011) researched and reviewed inventory models with lateral transshipments. Models of many different systems have been considered. This paper provides a literature review which categorizes the research to date on lateral transshipments, so that these differences can be understood and gaps within the literature can be identified.

Most of these studies dealt with a supply chain management structure formed by a single warehouse that supplies multiple retailers with a single product item. A single period of analysis also characterizes most of the models proposed. Based on this analysis it is clear that some space still exists to generalize the inventory management models proposed. In particular, new models should be developed to deal with more generic supply chain structures. Aspects that will allow a closer description of the real problems should be explored namely: i) generic structure with links between the different entities present (e.g. transshipment); ii) inclusion of supply lead times; iii) safety stock considerations and finally iv) inclusion of lost sales at all retailers.

The present work follows this need and develops an Inventory Management (IM) policy that may support in an optimized way, by minimizing total operational costs, the definition of the product flows through a multi-warehouse/multi-retailer/multi-product and multi-period supply chain. Costs include ordering, holding in stock and in transit, transportation, transshipping and lost sales. The system in study and the associated IM policy are modeled through a MILP model. The classical Continuous Review (CR) and the Periodic Review (PR) policies are also modeled by means of mathematical programming models that act in the same system, so as to compare the results of the three studied policies.

The remainder of this paper is organized as follows. The problem definition is given in section 2. Section 3 describes the IM mathematical model and the CR and PR mathematical models are presented respectively in sections 4 and 5. An inventory management case study is presented in section 6. Finally the conclusions are drawn in section 7.

2 Problem definition

A generic supply chain is considered in this study. It comprises multiple regional warehouses and multiple retailers as depicted in figure 1, where multiple products are distributed over a given time horizon.

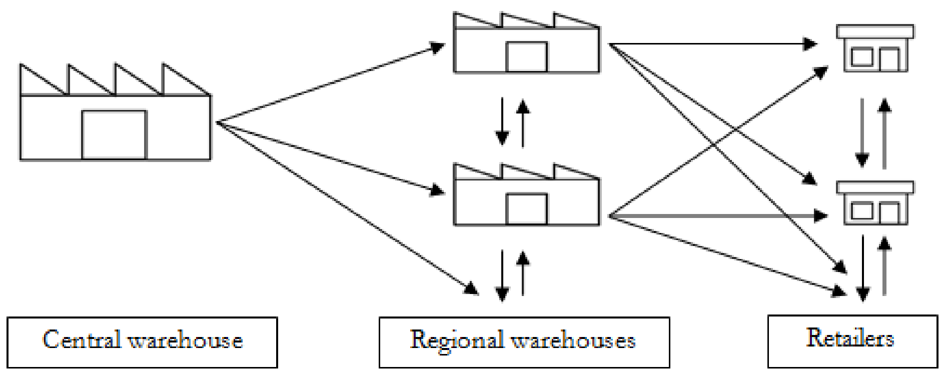


Figure 1: Supply chain structure.

The structure considered assumes that retailers replenish their inventories from the regional warehouses, these replenish their inventories from a central warehouse and customer demand is observed at the retailers. Each retailer has a normally distributed demand with a mean and a standard deviation in each unit of time, which is independent of demands of the other retailers. It is also assumed that all storage and transportation capacities are limited and that transportation occurs after orders have been placed. Lateral transshipment between regional warehouses or between retailers is allowed. If the demand in a given time period and at a given retailer is not satisfied, this is considered as a lost sale. Different cost types are included. These are related to the ordering process, holding in storage and holding in-transit, transportation, transshipping and lost sale. Fixed ordering costs occur each time a regional warehouse or a retailer places an order and are related to the ordering activity, being independent of the quantity ordered. Holding costs are defined for both storage and in-transit inventory. The first ones are defined per unit stored and per time period on each regional warehouse or retailer. The second ones are defined per unit of product transported and are dependent of the lead times. Transportation costs are considered per unit of material transported between the different stages of the supply chain. Related to these are the transshipment costs that represent the costs of transporting items between two locations belonging to the same echelon. Finally, lost sale costs are associated to the demand that cannot be satisfied and are defined per unit of product. Thus, it is important to effectively represent and optimize the flows of products through the entire supply chain so as to minimize costs. These aspects are included in the problem under study and the relevant decision that needs to be modeled is then to determine the shipping quantity to be sent from the regional warehouses to each retailer in each period so as to minimize the total system costs. The problem in study can be generally defined as follows:

Given:

- The planning time horizon and a discrete time scale;
- The number of regional warehouses and retailers;
- The number of products;
- Initial inventory level of each product at each regional warehouse and retailer;
- Retailer demands for each product in all time periods;
- Storage capacities for each product in each regional warehouse and retailer per time period;
- Transportation capacities for each product between entities;
- Safety stock by product in each regional warehouse and retailer;
- Transportation lead times between entities;
- Ordering costs per order of each product at each regional warehouse and retailer (independent of order quantity);
- Unitary holding cost per time period per product at each regional warehouse or retailer;
- Unitary holding cost per time period per product in transit (dependent of lead time);
- Unitary transportation and transshipping cost per product;
- Unitary lost sale cost per product and per each time period;

Determine:

- The inventory profiles for each product throughout the time horizon in each regional warehouse and retailer;
- The flows of products across the supply chain for each time period. These involve shipping quantities between entities on different supply chain levels and transshipment quantities between entities on the same supply chain level;
- Lost sale quantities of each product at each retailer at each time period;

So as to minimize the total costs over the time horizon considered. A mathematical programming formulation is proposed for the problem, which will be presented in the subsequent section.

3 Inventory Management model (IM model)

The supply chain inventory management problem presented is formulated as a MILP model. As referred, it aims to minimize the total costs during the time horizon in study. The MILP model considers time through a discretized time scale, where the time intervals have equal durations. The indices, sets, parameters and variables (binary and continuous) used in the model formulation are defined using the following notation:

Indices

i product

j, k, l, m entity node

t time period

Sets

$i \in P = \{1, 2, 3, \ldots, S\}$ products

$j, k, l, m \in I = \{0, 1, 2, \ldots, NW, NW + 1, NW + 2, \ldots, NW + NR\}$ supply chain nodes

$t \in T = \{1, 2, 3, \ldots, TN\}$ time periods

$W = \{1, 2, 3, \ldots, NW\}, W \subset I$ regional warehouses

$R = \{NW + 1, NW + 2, \ldots, NW + NR\}, R \subset I$ retailers

$W_o = \{0\}, W_o \subset I$ central warehouse

$DN = \{1, 2, 3, \ldots, NW, NW + 1, \ldots, NW + NR\}, DN \subset I$ demand nodes (regional warehouses and retailers)

$SN = \{0, 1, 2, 3, \ldots, NW\}, SN \subset I$ supply nodes (central warehouse and regional warehouses)

Note that $W_o \cup W \cup R = I$

Parameters

BGM a large positive number

CD_{ijt} customer demand of the product i at node j in time period t (note that customer demand occurs only at the retailers, but not at the warehouses)

HOC_{ij} unitary holding cost of the product i at node j per time unit

$HTCR_{ijk}$ unitary holding cost of the product i in transit from a regional warehouse $j \in W$ to a retailer $k \in R$ per time unit

$HTCW_{ijk}$ unitary holding cost of the product i in transit from a central warehouse $j \in W_o$ to a regional warehouse $k \in W$ by time unit

Ito_{ij} initial inventory level of the product i at node j

LSC_{ijt} unitary lost sale cost of the product i at node j in time period t

LTT_{ijk} transportation lead time of the product i from node j to node k

OC_{ij} ordering cost of the product i at node j (note that ordering cost is independent of quantity of product i)

SS_{ij} safety stock factor of the product i at node j

STC_{ijt} storage capacity of the product i at node j in time period t

$TRACMAX_{ijk}$ maximum transportation capacity of the product i from node j to node k

$TRACMIN_{ijk}$ minimum transportation capacity of the product i from node j to node k

$TRCR_{ijk}$ transportation cost of one unit of product i from a regional warehouse $j \in W$ to a retailer $k \in R$

$TRCW_{ijk}$ transportation cost of one unit of product i from a central warehouse $j \in W_o$ to a regional warehouse $k \in R$

$TSCW_{ijl}$ transshipment cost of one unit of product i from a regional warehouse $j \in W$ to another regional warehouse $l \in W$

$TSCR_{ikm}$ transshipment cost of one unit of product i from a retailer $k \in R$ to another retailer $m \in R$

Continuous variables

FI_{ijt} inventory of product i at node j at the end of time period t

LS_{ijt} lost sales quantity of product i at node j during time period t (note that lost sales only occur at the retailers)

SQ_{ijkt} shipping quantity of product i from node j to node k during time period t

Binary variable

$BV1_{ijt}$ equal to 1 if an order of product i is placed by node j in time period t; 0 otherwise

Objective function

The objective function consists of the minimization of the total cost is given as follows:

$$\begin{aligned}
&\textit{minimize total cost} = \sum_{i\in P}\sum_{j\in DN}\sum_{t\in T} OC_{ij} \times BV1_{ijt} + \sum_{i\in P}\sum_{j\in DN}\sum_{t\in T} HOC_{ij} \times FI_{ijt} \\
&+\sum_{i\in P}\sum_{j\in W_o}\sum_{k\in W}\sum_{t\in T} HTCW_{ijk} \times LTT_{ijk} \times SQ_{ijkt} + \sum_{i\in P}\sum_{j\in W}\sum_{k\in R}\sum_{t\in T} HTCR_{ijk} \times LTT_{ijk} \times SQ_{ijkt} \\
&+\sum_{i\in P}\sum_{j\in W_o}\sum_{k\in W}\sum_{t\in T} TRCW_{ijk} \times SQ_{ijkt} + \sum_{i\in P}\sum_{j\in W}\sum_{k\in R}\sum_{t\in T} TRCR_{ijk} \times SQ_{ijkt} \\
&+\sum_{i\in P}\sum_{j\in W}\sum_{l\in W}\sum_{t\in T} TSCW_{ijl} \times SQ_{ijlt} + \sum_{i\in P}\sum_{k\in R}\sum_{m\in R}\sum_{t\in T} TSCR_{ikm} \times SQ_{ikmt} \\
&+\sum_{i\in P}\sum_{j\in R}\sum_{t\in T} LSC_{ijt} \times LS_{ijt} \qquad (3.1)
\end{aligned}$$

The first term of objective function (1) is the ordering cost. The second term expresses the holding costs at both stages of the supply chain, regional warehouses and retailers. The third and the fourth terms are the holding cost in transit at both supply chain stages. The holding cost in transit is lead time dependent. The fifth and the sixth terms are the transportation costs at both supply chain stages. The transshipping cost, between regional warehouses and between retailers, is given by the seventh and eighth terms and, finally, the last term represents the lost sale cost.

Constraints

The model developed consists of different types of constraints. These are grouped into: inventory constraints; shipping constraints; storage capacities; transportation capacities; safety stock policy and non-negativity and binary conditions.

Inventory constraints

Inventory constraints have to be defined for both the regional warehouses and retailers, taking into account all inputs and outputs at each time period.

Regional warehouses:

The total incoming quantity at each regional warehouse is equal to the shipping quantity from the central warehouse to the regional warehouse, plus the transshipping quantity from the others regional warehouses, considering the transportation lead time through the introduction of a time lag.

The total outgoing quantity at each regional warehouse is equal to the sum of shipping quantity from the regional warehouse to the retailers plus the transshipping quantity to the others regional warehouses, at time unit t. For $t = 1$ the inventory of product i at the end of this time period at regional warehouses is given by constraints (2), which takes into account the initial inventory level at each regional warehouse Ito_{im}.

$$\begin{aligned}
FI_{imt} = Ito_{im} + SQ_{i,0,m,t-LTT_{i0m}|LTT_{i0m}=0} - \sum_{k\in R} SQ_{imkt} - \sum_{l\in W \wedge l\neq m} SQ_{imlt} \\
+ \sum_{l\in W \wedge l\neq m} SQ_{i,l,m,t-LTT_{ilm}|LTT_{ilm}=0}, \quad i\in P, m\in W, t=1 \qquad (3.2)
\end{aligned}$$

For the remaining time periods the inventory at the end of these time periods at regional warehouses is given by constraints (3).

$$FI_{imt} = FI_{i,m,t-1} + SQ_{i,0,m,t-LTT_{i0m}|LTT_{i0m}<t} - \sum_{k\in R} SQ_{imkt} - \sum_{l\in W \wedge l\neq m} SQ_{imlt}$$

$$+\sum_{l\in W\wedge l\neq m} SQ_{i,l,m,t-LTT_{ilm}|LTT_{ilm}<t}, \quad i\in P, m\in W, t\in T\backslash\{1\} \tag{3.3}$$

Retailers:

At each retailer, the incoming quantity is equal to the sum of the shipping quantity from the regional warehouses to that retailer, plus the transshipping quantity from the others retailers, at time unit t, considering the transportation lead time.

At each retailer the outgoing quantity is equal to the customer demand minus the lost sale of that retailer plus the transshipping quantity to the others retailers, at time unit t.

For $t = 1$ the inventory of product i at the end of this time period at the retailers is given by constraints (4), which accounts for the initial inventory of product i at retailer k (Ito_{ik}) whereas constraints (5) is applicable for the remaining time periods.

$$FI_{ikt} = Ito_{ik} + \sum_{j\in W} SQ_{i,j,k,t-LTT_{ijk}|LTT_{ijk}=0} - (CD_{ikt} - LS_{ikt}) - \sum_{m\in R\wedge m\neq k} SQ_{ikmt}$$

$$+\sum_{m\in R\wedge m\neq k} SQ_{i,m,k,t-LTT_{imk}|LTT_{imk}=0}, \quad i\in P, k\in R, t=1 \tag{3.4}$$

$$FI_{ikt} = FI_{i,k,t-1} + \sum_{j\in W} SQ_{i,j,k,t-LTT_{ijk}|LTT_{ijk}<t} - (CD_{ikt} - LS_{ikt}) - \sum_{m\in R\wedge m\neq k} SQ_{ikmt}$$

$$+\sum_{m\in R\wedge m\neq k} SQ_{i,m,k,t-LTT_{imk}|LTT_{imk}<t}, \quad i\in P, k\in R, t\in T\backslash\{1\} \tag{3.5}$$

Shipping constraints

Since transportation occurs after an order has been placed from a destination to its source, it is assumed that the fixed ordering cost is always incurred when the transportation occurs. Hence, if the transportation amount is not zero the binary variable $BV1_{ikt}$ equals 1, as implied in constraints (6). The left hand side of this constraint represents the quantity received by a regional warehouse, which can come from the central warehouse (first term) or any other regional warehouse (second term).

$$SQ_{i0kt} + \sum_{l\in W\wedge l\neq k} SQ_{ilkt} \leq BGM \times BV1_{ikt}, \quad i\in P, k\in W, t\in T \tag{3.6}$$

Equivalents constraints are defined for retailers, constraints (7). The BGM value will have a value that is valid as an upper bound of any quantity that can be ordered by a regional warehouse or retailer.

$$\sum_{j\in W} SQ_{ijkt} + \sum_{l\in R\wedge l\neq k} SQ_{ilkt} \leq BGM \times BV1_{ikt}, \quad i\in P, k\in R, t\in T \tag{3.7}$$

Storage capacities

At any time period t, the inventory level of each product i must respect the storage capacity in each demand node j, which is enforced by constraints (8). In this formulation we consider storage capacities time dependent to illustrate that, depending on the time period, capacities may vary since we are dealing on an operational perspective.

$$FI_{ijt} \leq STC_{ijt}, \quad i\in P, j\in DN, t\in T \tag{3.8}$$

Transportation capacities

At any time period t, the shipping quantity of each product i must respect the transportation lower and maximum limits between each two nodes j and k, as stated in constraints (9) and (10). Note that the quantities in transit were not considered for the usage of transportation capacity, but for the total transportation capacity. Also the transportation capacity from the central warehouse is unlimited.

$$SQ_{ijkt} \leq TRACMAX_{ijk}, \quad i\in P, j\in DN, k\in DN, t\in T \tag{3.9}$$

$$TRACMIN_{ijk} \leq SQ_{ijkt}, \quad i\in P, j\in DN, k\in DN, t\in T \tag{3.10}$$

Safety stock policy

Constraints (11) ensure that the inventory of each product i at each node j at each time period t must be higher than the required safety stock level for that product in that node.

$$SS_{ij} \leq FI_{ijt}, \quad i\in P, j\in DN, t\in T \tag{3.11}$$

Non-negativity and binary conditions

As defined above, the model uses both positive continuous variables (12) and binary variables (13).

$$SQ_{ijkt}, FI_{ijt}, LS_{ijt} \geq 0, \quad i \in P, j \in I, k \in I, t \in T \tag{3.12}$$

$$BV1_{ijt} \in \{0,1\}, \quad i \in P, j \in I, t \in T \tag{3.13}$$

The above model is formed by constraints (2) to (13), using the objective function (1) that describes the proposed IM model. In order to compare the performance and adequacy of this model with classical inventory policies, two other models are developed: the continuous review inventory model (CR, section 4) and the periodic review inventory model (PR, section 5).

4 Continuous review inventory system model (CR model)

In the CR policy, the inventory level is continuously reviewed over the time horizon. Thus, whenever the inventory level is at or below a given reorder point level s, an order is placed that has a fixed quantity. Since inventory is tracked and the order is made when necessary, this inventory management policy is characterized by a fixed order quantity requested at variable time intervals. In order to represent this policy through a MILP model, two more parameters must be added to the list proposed in section 3:

RSC_{ik} reference stock level of product i at node k (used for regional warehouses and retailers)

s_{ik} reorder point level of product i at node k (used for regional warehouses and retailers)

The fixed order quantity is then given by $(RSC_{ik} - s_{ik})$ since at the moment that an order is placed the inventory position (inventory level plus inventory in transit) must reach the reference stock level. In terms of model representation, the reference stock value can replace the BGM value used in constraints (6) and (7), since that value also works as an upper bound of any SQ variable. The MILP model that represents the continuous review policy consists of constrains (2) to (13), with the change in constraints (6) and (7), while the objective function (1) remains equal.

5 Periodic review inventory system model (PR model)

In this policy, the inventory level is reviewed at fixed time points, determined by a fixed review period. If at that time the inventory level is below a reference stock RS, an order is placed. The order quantity is determined by the difference between the reference stock and the current inventory level, to bring the inventory position up to level RS; otherwise, nothing is done until the next review point. A periodic inventory system uses variable order sizes at fixed time intervals. We add a subset of T that includes all the time periods for which there will be orders.

$ST \subset T$ all the time periods for which there will be orders due to the periodic review policy.

As in section 4 also in here is necessary to add an extra parameter to the initial list of section 3.

RSP_{ik} reference stock level of product i at node k (used for regional warehouses and retailers)

The original constraints (6) and (7) are now replaced by constraints (14) and (15).

$$SQ_{i0kt} + \sum_{l \in W \wedge l \neq k} SQ_{ilkt} \leq (RSP_{ik} - FI_{ikt}) \times BV1_{ikt}, \quad i \in P, k \in W, t \in ST \tag{5.1}$$

$$\sum_{j \in W} SQ_{ijkt} + \sum_{l \in R \wedge l \neq k} SQ_{ilkt} \leq (RSP_{ik} - FI_{ikt}) \times BV1_{ikt}, \quad i \in P, k \in R, t \in ST \tag{5.2}$$

In order to solve the non-linearity on the right end-side of constraints (14) and (15), we define the auxiliary variable Y_{ikt} (positive continuous variable) that represents the inventory of product i at node k at the end of time period t. Thus, the non-linear term $FI_{ikt} \times BV1_{ikt}$ is replaced by the continuous variable Y_{ikt} in constraints (14) and (15). The value of this auxiliary variable is given by equation (16):

$$Y_{ikt} = FI_{ikt} \times BV1_{ikt}, \quad i \in P, k \in I, t \in T \tag{5.3}$$

Using the definition of variable Y_{ikt} it is possible to impose the logical conditions (17) and (18). If the binary variable $BV1_{ikt}$ is 0, then the auxiliary variable Y_{ikt} is also 0 (condition (17)). If, on the other hand, the binary variable $BV1_{ikt}$ is equal to 1, we want to ensure that the new auxiliary variable takes the value of the inventory in the current time interval (FI_{ikt}), as expressed in condition (18).

$$BV1_{ikt} = 0 \implies Y_{ikt} = 0, \quad i \in P, k \in I, t \in T \tag{5.4}$$

$$BV1_{ikt} = 1 \implies Y_{ikt} = FI_{ikt}, \quad i \in P, k \in I, t \in T \tag{5.5}$$

To translate these logical conditions into the MILP model representation of the periodic review policy, we need to add the extra constraints (19), (20) and (21).

$$Y_{ikt} - RS_{ik} \times BV1_{ikt} \leq 0, \quad i \in P, k \in I, t \in T \tag{5.6}$$

$$-FI_{ikt} + Y_{ikt} \leq 0, \quad i \in P, k \in I, t \in T \tag{5.7}$$

$$FI_{ikt} - Y_{ikt} + RS_{ik} \times BV1_{ikt} \leq RS_{ik}, \quad i \in P, k \in I, t \in T \tag{5.8}$$

where RS_{ik} is the upper bound for FI_{ikt} (and hence also for Y_{ikt}). Constraints (19) and (20) ensure that the auxiliary variable takes the value of 0 if the binary variable is equal to 0 (constraints (21)). If this variable is equal to 1, then the auxiliary variable takes, at most, the value of FI_{ikt} (constraints (20)). In order to ensure that in this situation the auxiliary variable takes exactly the value of FI_{ikt}, we add constraints (21). Note that this equation only becomes active whenever $BV1_{ikt} = 1$. Finally, we add constraints (22) and (23) that are applied for time periods where no orders are to be placed. They impose that no transportation activity occurs at these time periods.

$$SQ_{i0kt} + \sum_{l \in W \wedge l \neq k} SQ_{ilkt} = 0, \quad i \in P, k \in W, t \notin ST \tag{5.9}$$

$$\sum_{j \in W} SQ_{ijkt} + \sum_{l \in R \wedge l \neq k} SQ_{ilkt} = 0, \quad i \in P, k \in R, t \notin ST \tag{5.10}$$

The periodic review policy MILP model consists of constraints (2) to (5), (8) to (13), (16) and constraints (19) to (23), using objective function (1).

6 Inventory management case study

In this section we present a case study based on a Portuguese Company that we use to compare the three inventory management policies modeled, and to test the proposed IM policy. Due to confidentiality reasons the data provided has been changed but still describes the real operation. Please note that the products amounts involved in the present case study although referred as units they correspond to euro-pallets. The models were implemented in GAMS 23.5 modeling language and solved using the CPLEX 12.2.0.0 solver in an Intel CORE i5 CPU 2.27GHz and 4GHz RAM. The stopping criterion was either a computational time limit of 3600 seconds or the determination of the optimal solution.

6.1 Data and parameters

The supply chain considered involves one central warehouse, three regional warehouses and four retailers. Three main aggregated type of products are considered. The storage capacity of all warehouses and retailers is of 500 units and the transportation capacity is between 0-500 units for each product. We consider that demand is represented by a normal distribution, with parameters presented in table 4. A 7-period planning horizon was assumed to test the three different inventory management policies. The first scenario considers the IM policy (modeled in section 3), which uses a variable order quantity at variable time intervals. The second scenario explores the CR policy (section 4) that uses a fixed order quantity at variable time intervals. Finally, the third scenario explores the PR policy (section 5) where a variable order quantity at fixed time intervals is considered. Tables from 1 to 5 present all the parameters.

Table 6 illustrates the reference stock level for the CR and PR policies for the warehouses and retailers. The reorder point value is assumed equal to the safety stock. In the PR policy that implies a review period taking place every three time periods, the first time period with revision is time period 3. The initial inventory level, the customer demand, reference stock level for warehouses and retailers and the supply chain structure are the same for the three policies.

Table 1: Models parameters (euro).

parameters	values
OC for warehouses and retailers	20
HOC for warehouses	0.2
HOC for retailers	0.6
HTCW from central warehouse to regional warehouses	0.3
HTCR from regional warehouses to retailers	0.9
LS for retailers	25

Table 2: Unit transportation cost for all products (euro).

		warehouse1	warehouse2	warehouse3	
TRCW	warehouse0	0.55	0.22	0.70	
TSCW	warehouse1	0	0.35	0.75	
	warehouse2	0.35	0	0.40	
	warehouse3	0.75	0.40	0	
		retailer1	retailer2	retailer3	retailer4
TRCR	warehouse1	0.22	0.20	0.32	0.38
	warehouse2	0.68	0.52	0.34	0.10
	warehouse3	0.95	0.70	0.40	0.25
TSCR	retailer1	0	0.10	0.40	0.65
	retailer2	0.10	0	0.15	0.50
	retailer3	0.40	0.15	0	0.18
	retailer4	0.65	0.50	0.18	0

Table 3: Initial inventory level (Ito)/Safety stock (SS) on warehouses and retailers (unit).

		warehouse1	warehouse2	warehouse3	retailer1	retailer2	retailer3	retailer4
Ito	product1	45	30	40	24	22	20	18
	product2	15	11	12	16	14	12	10
	product3	11	9	10	8	4	6	9
		warehouse1	warehouse2	warehouse3	retailer1	retailer2	retailer3	retailer4
SS	product1	14	14	14	7	3	3	3
	product2	11	11	11	2	2	2	2
	product3	8	8	8	2	2	1	1

Table 4: Customer demand (CD) parameters for product1/product2/product3 (unit).

		mean demand	standard deviation
CD	retailer1	12 / 8 / 4	4 / 4 / 2
	retailer2	11 / 7 / 4	4 / 4 / 3
	retailer3	10 / 6 / 4	6 / 3 /1
	retailer4	9 / 5 / 5	3 / 3 / 1

Table 5: Transportation lead time (LTT) for all products (time period).

		warehouse1	warehouse2	warehouse3	retailer1	retailer2	retailer3	retailer4
LTT	warehouse0	2	1	3	0	0	0	0
	warehouse1	0	1	1	1	1	1	1
	warehouse2	1	0	1	2	2	1	0
	warehouse3	1	1	0	3	2	1	1
	retailer1	0	0	0	0	1	1	1
	retailer2	0	0	0	1	0	1	1
	retailer3	0	0	0	1	1	0	1
	retailer4	0	0	0	1	1	1	0

Table 6: Reference stock level for CR policy (RSC)/ Reference stock level for PR policy (RSP) on warehouses and retailers (unit).

		warehouse1	warehouse2	warehouse3	retailer1	retailer2	retailer3	retailer4
RSC	product1	280	280	280	72	66	60	54
	product2	220	220	220	48	42	36	30
RSP	product3	160	160	160	36	30	28	20

6.2 Experimental results

The costs per nature of the objective function and the computational statistics for all three policies are shown in tables 7 and 8 respectively. For the IM policy the holding cost, of 557.60 euro, is the highest cost term. In terms of computational statistics, it is solved to optimality within less than 3 seconds and the objective function reaches a value of 1955.46 euro. For the CR policy the objective function has a higher value than the observed for the IM policy, being the cost with highest contribution the lost sales cost. Holding costs also increase when compared to the IM policy. The model size is similar to the IM policy and the optimal solution was obtained in less than 1 second. The PR policy has the lowest ordering costs. On the other hand, since inventory levels are not replenished frequently, the highest cost becomes associated with the lost sales. The objective function value is 10218.68 euro, being the highest one. In conclusion the IM policy appears as more flexible than the CR or PR policies, which is then confirmed by the lower operational costs. This is explained by the policy flexibility in terms of managing the flows allowing for an order occurrence strictly when necessary. This leads to a less costly system.

Table 7: Costs per nature for all three policies (euro).

	IM policy		CR policy		PR policy	
	value	percentage	value	percentage	value	percentage
Ordering	520.00	26.59	380.00	10.26	220.00	2.15
Holding	557.60	28.52	1026.00	27.71	704.20	6.89
Holding in transit	168.00	8.59	559.50	15.11	521.40	5.10
Transportation	135.48	6.93	286.58	7.74	238.84	2.34
Transshipping	124.38	6.36	100.91	2.73	34.24	0.34
Lost-sales	450.00	23.01	1350.00	36.45	8500.00	83.18
Total	1955.46	100.00	3702.99	100.00	10218.68	100.00

Table 8: Computational statistics for all three policies.

	IM policy	CR policy	PR policy
MIP solution	1955.46	3702.99	10218.68
Single equations	475	475	601
Single variables	1126	1126	1168
Discrete variables	168	168	168
Computational time (second)	2.65	0.53	0.06

7 Conclusions

This paper proposes a generic IM policy applied to a multi-period and multi-product supply chain, which provides an inventory and distribution plan that minimizes the total costs. Based on the experimental results, it could be concluded that the proposed IM policy proved to be more rentable since the total costs are lower when compared to the classical policies for the same network and under the same conditions.

The lost sales comprise a high cost that reflects somewhat in the total costs for our test company. Therefore, for generality, other model driving force should be tested, namely profit or even to balance costs with service level. As future work we also will focus on optimizing the safety stock and improve model formulation. Our model can evolve to have shared product constraints, such as capacity, which is, in reality, shared, either in warehouses or transports.

The results obtained and consequent conclusions are related with our test company. In future works, we should provide more instances so as to study the potential generalization of our conclusions.

References

[Abdul-Jalbar et al., 2006] Abdul-Jalbar, B., Gutiérrez, J., and Sicilia, J. (2006). Single cycle policies for the one-warehouse N-retailer inventory/distribution system. *OMEGA the International Journal of Management Science*, 34:196–208.

[Hsiao Yu-Cheng, 2008] Hsiao Yu-Cheng (2008). Optimal single-cycle policies for the one-warehouse multi-retailer inventory/distribution system. *International Journal of Production Economics*, 114:219–229.

[Kautter, 1999] Kautter, J. (1999). Inventory theory: new perspectives for cooperative management. In *International Journal of Production Economics*, 59:129–134.

[Monthatipkul and Yenradee, 2008] Monthatipkul, C., and Yenradee, P. (2008). Inventory/distribution control system in a one-warehouse/multi-retailer supply chain. *International Journal of Production Economics*, 114:119–133.

[Ozdemir et al., 2006] Ozdemir, D., Yucesan, E., and Herer, Y. (2006). Multi-location transshipment problem with capacitated transportation. *European Journal of Operation Research*, 175:602–621.

[Paterson et al., 2011] Paterson, C., Kiesmuller, G., Teunter, R., and Glazebrook, K. (2011). Inventory models with lateral transshipments: A review. *European Journal of Operation Research*, 210:125–136.

Scheduling batch processes using the RTN discrete time formulation: a case study

Miguel Vieira*, Tânia Pinto-Varela*, Ana Paula Barbosa-Póvoa*

*Center for Management Studies (CEG-IST), Instituto Superior Técnico, Technical University of Lisbon, Portugal
migueljvieira@ist.utl.pt
tania.pinto.varela@ist.utl.pt
apovoa@ist.utl.pt

Abstract

In order to guarantee the correct allocation of resources, an efficient and uniform methodology is required to address the wide diversity of operational problems. Scheduling has emerged as a critical aspect in industrial management due to the condition of high process flexibility. In this work, a case study in the chemical industry is presented, where the complexity of resource allocation and schedule optimization is addressed through the use of the Resource Task Network methodology. A Mixed Integer Linear Programming model is implemented for the production maximization. The results are discussed based in plant resources allocation and availability of required workforce.

Keywords: Scheduling optimization, RTN, multipurpose plant

1 Introduction

Process control and monitoring has become increasingly important in process industries as a consequence of global competition and fast changing economic conditions. With current manufacturing processes requiring the coordination of multiproduct, recipes and equipment, an efficient methodology is needed to address an optimal allocation of resources and performance. For that purpose, the development of computer-based decision-support tools to assist daily operations is becoming especially important in plants with high process flexibility. Scheduling of operations is crucial for improving production performance, as long as the operational constrains are guaranteed. Moreover, the production flexibility is gathering importance in current facilities management, requiring the plant redesign to adjust production.

Nowadays, an extensive literature displays the great research interest in this area. With focus in scheduling problems, it addresses the challenge not only from the mathematical point of view, but also its relevance in solving real cases studies [Méndez et al., 2006]. Developing scheduling systems requires a model formulation grounded in an adequate methodology to characterize the problem without ambiguities. The problem characterization can be done using two main methodologies: one state that design and scheduling problems are very diverse and where specific models are required; the other assume similarity between the problems and a uniform representations have been proposed [Pinto et al., 2008]. The latter was explored by Kondili (1988), proposing the first generic representation for scheduling batch processes called State-Task-Network (STN). Later on, Pantelides (1994) proposed the Resource-Task Network (RTN), a more uniform methodology able to address the optimal allocation of resources to tasks. More recently, some adaptations and detailed aspects were taken into account by other authors, namely Pinto et al. (2008).

The research upon the mathematical representation of complex networks have provided the development of unified frameworks, integrating diverse operational events (e.g. variable/fixed batch size, storage/transfer policies, energy integration, changeovers, continuous/discrete time representation) or multiple objective criteria (e.g. makespan, earliness, or cost minimization). Several reviews of state-of-the-art optimization methods for short-term scheduling and design of batch processes can be found e.g. Méndez et al. (2006) or Barbosa-Póvoa (2007). Mainly mathematical programming is used in multiproduct/multipurpose cases, where both Mixed Integer Linear and Non-Linear Programming (MILP and MINLP) formulations are developed. Heuristics and stochastic methods have also been explored to overcome the high computational burden: new approaches are emerging such as problem oriented heuristics, evolutionary algorithms (e.g. Genetic Algorithms) and meta-heuristics (Simulated Annealing, Tabu Search or Ant Colonies) [Chibeles-Martins et al., 2010, Simaria et al., 2011].

The study here presented takes a case of a chemical process in paint industry, addressing the complexity of resource allocation in accordance with the scheduling optimization, based on the RTN representation. The aim of this work is, not only, to determine the optimal scheduling while considering the production maximization, but also, analyze the workforce availability. With this analysis, it is aimed to improve the utilization rate of teams required to execute the demanded production plan. The paper is structured as follows: the second section presents the problem characterization, followed by the modeling framework; the case study results follows and the paper finalizes with the main conclusions and future remarks.

2 Problem definition

Process scheduling includes the assignment of tasks, in a specific sequence, to the available resources during a timeline. This assumes special complexity when is verified competition among the different scheduled tasks for the limited available resources. The characterization of the process allows the identification of different constraints that need to be considered in the mathematical model, such as the operational limitations: availability and capacities of equipment/resources; manufacturing of different products using different recipes and tasks' processing times; demand levels, etc.

The case study encompass a paint manufacturing process addressing a scheduling problem: the process management requires an optimal sequence of operations, considering all the aspects of the manufacturing process, while maximizing resource efficiency and satisfying demand levels. Based on the optimal allocation of resources, the analysis of the workforce availability is explored through the run of the presented model for different cases. That represents an important aspect in the operations management, since the labor costs are one of the most significant in the operational structure. And since high flexibility in the production output is demanded, the access to decision-support tools is essential to provide information to the management of the available (or additional) workforce.

The main problem under study can be stated as follows, assuming a uniform discretization of time:

Given:
- Fixed time horizon of scheduling;
- Set of final products composing a production plan and associated demands;
- The RTN representation of the process recipes (tasks and resources required);
- Resources availability (raw materials, equipment and workforce) and existent restrictions;
- Processing time of each task.

Determine:
- Process schedule;
- Resources allocation to each task.

To the objective function: the production output maximization.

3 Modeling framework

Based on the representation proposed by Pantelides (1994), the Resource-Task Network is a uniform and conceptually simple methodology that involves two types of entities, Tasks and Resources. A Task is an abstract operation that consumes and/or produces a specific set of indistinctive Resources. Resources can be classified as non-renewable, such as raw materials or utilities, and renewable, which considers the remaining resources on the plant (equipment or workforce). The Resource-Task Network (RTN) methodology will be applied to our problem for the development of a MILP model, using a discrete time representation, considering the following constrains: excess resource balance, excess resource capacity and operational constrains.

Each task has a fixed duration τ_k and the execution of task k, starting at time t, is characterized by its extent: a pair of variables (N_{kt}, ξ_{kt}). N_{kt} is an integer variable representing the start of task k at event point t, while ξ_{kt} is a continuous variable defining the amount of resource processed by task k at event point t. The amount of resource r consumed or produced at each time $\theta = 0, \ldots, \tau_k$ is expressed by $(\mu_{kr\theta} N_{k,t-\theta} + \nu_{kr\theta} \xi_{k,t-\theta})$. The coefficients $\mu_{kr\theta}$ and $\nu_{kr\theta}$ represent the discrete interaction linked to the two variables considering the resource r consumption/production in task k and time θ. Negative values for the latter indicate consumption of resource, while positive values denote production. Changes to the resource utilization can only occur at interval boundaries. The variable R_{rt} keeps track of the resource availability r at time t and the change in the excess resource level for each resource type, from one time interval to the next, is given by excess resource balance constrain:,

Excess resource balance: the mass balance for each resource r must be satisfied at every instant t, considering R_{r0} the level of resource available initially (K_r is the set of tasks that use resource r, subset of R group of resources, and H the planning horizon).

$$R_{rt} = R_{r,t-1} + \sum_{k \in K_r} \sum_{\theta=0}^{\tau_k} (\mu_{kr\theta} N_{k,t-\theta} + \nu_{kr\theta} \xi_{k,t-\theta}), \ \forall r \in R, t \in H \tag{3.1}$$

Excess resource capacity constraint: the amount of excess resource r at any time cannot be negative and is limited by the demand value of the production plan.

$$0 \leq R_{rt} \leq R_{rt}^{max}, \ \forall r \in R, t \in H \tag{3.2}$$

Operational constrains: for each task k making use of a processing equipment resource r, the amount of material being processed must always be within the maximum and minimum capacities given by V_{kr}, where E is the subset of R resources for the processing units.

$$V_{kr}^{min} N_{kt} \leq \xi_{kt} \leq V_{kr}^{max} N_{kt}, \ \forall r \in E \subset R, k \in K_r, t \in H \tag{3.3}$$

The objective function considered is the maximization of production, where F is a subset of R for the final products, given by:

$$max \sum R_{rt}, \ \forall r \in F \subset R, t \in H \tag{3.4}$$

4 Case Study

Our case study is based on a real chemical industry process: the company is mainly dedicated to the development and production of water-based paint products, providing different compositions concerning its final application. The paint is generally composed by a binder, which is the film forming component; pigments, that give the color; and solvent, that keeps the paint in the liquid form. The process is laid out vertically, composed by an initial step of mixing of the raw materials (dispersion tanks), followed by a finishing step of homogenization of paint properties (finishing tanks) and ending in one of the filling machines (filling lines), where the paint is filled into package containers and sent to storage. The company has installed a considerable number of equipment units, suitable for the current demand, where some are dedicated to the production of a single paint category: *Textured, Smooth* or *Brand* type. In Figure 1 are detailed the possible connections between the set of available equipment resources considered for our study: 4 dispersion tanks $\{M_3, M_5, M_6, M_7\}$, 30 finishing tanks $\{T_1, ..., T_{30}\}$ and 4 filling machine lines $\{L_1, L_2, L_3, L_4\}$, corresponding to the ones with more significance in the manufacturing process.

The different connections between equipment units reveal the possible paths of the process flow. For example, M_3 only produces one category of paint products - *Brand* - that can only be stored in one of the tanks $\{T_{25}, T_{26}, T_{27}, T_{28}, T_{29}, T_{30}\}$, and can only be filled in Line L_4, generating a resource competition among these equipment.

Additional considerations in the characterization of the problem must also account for:

- A production order has to be fully accomplished in 48 hours, with the filling stage taking place in the following shift after the manufacturing step (dispersion and finishing), due to intermediate quality tests procedures. During this elapsed time, the batch is kept stored in the finishing tank.
- The manufacturing step presents a constant average processing time, while the filling step verifies a higher variability in the processing of each order, depending on the batch size and number/types of containers to produce (*Small, Medium* or *Large* containers).
- To guarantee a steady process flow, the management team has decided to restrain the available workforce in different production areas per shift, requiring additional temporary workforce when production demand increases. Two teams of operators, HRa and HRb, are assigned to perform the manufacturing and filling tasks, respectively.

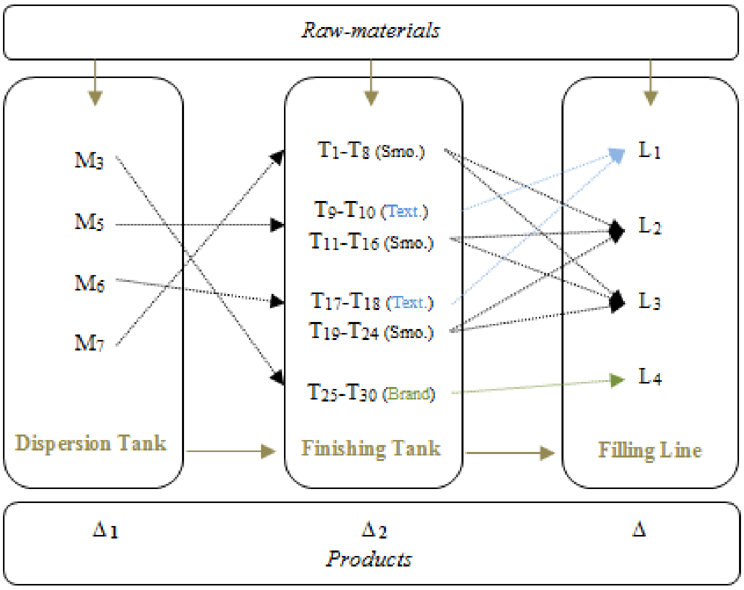


Figure 1: Production process layout

The management wants to get the optimal schedule for a production plan, set for five consecutive shifts of 8 hours, maximizing the production output while guaranteeing the optimal utilization of the resources, in particular the available workforce.

To illustrate our case study, the process recipe for generic product Δ is shown in Figure 2, considering the RTN framework:

- The raw-material, Δ_{rm}, will be used at the *Preparation* stage, Pre_Δ, producing the material, Δ_1, and send it to storage; this *Storage* task, Sto_Δ, consumes Δ_1 and produces Δ_2, which goes to the *Filling* stage, Fil_Δ, until the total quantity of product Δ order is completed. The auxiliary variables - Δ_1 and Δ_2 – are used in the intermediate formulation to represent the connections between the joining tasks, without storage significance.
- Concerning Resources allocation during processing times:
  - The *Preparation* stage requires one equipment (M_i equipment, with $i = 3, 5, 6, 7$) and one team member from HRa;
  - The *Storage* stage requires one of the Tanks (T_i);
  - The *Filling* stage requires one of the Lines (L_i filling lines, with $i = 1, 2, 3, 4$) and one team member from HRb. This task also captives the Tank T_i used in the *Storage* task, because this equipment is only released after the filling order is completed and the tank is emptied.
- Regarding the allocation of Tanks, it were defined three variables T_T, T_F and T_S, corresponding to three groups of dedicated tanks, respectively, for *Textured*, *Smooth* and *Brand* paint types. These variables T_T, T_F and T_S will consider groups of 4, 20 and 6 tanks, respectively, according to Figure 1.

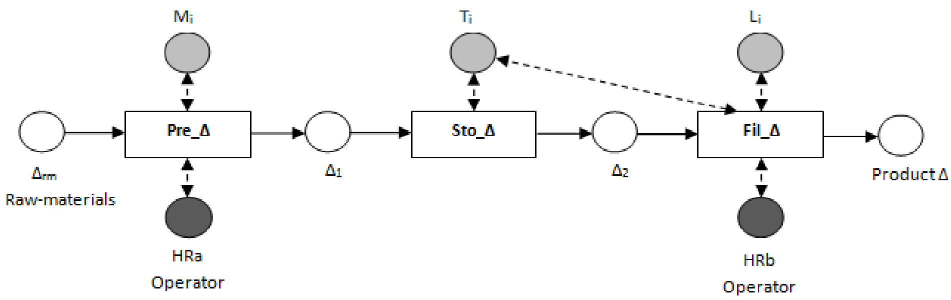


Figure 2: RTN-based single-product process representation

The *Preparation* stage represents all the manufacturing operations (dispersion and finishing) with an average processing time of 150 min, independently of the type of product. Assuming a 48 h production time, it was settled an average time slot for the *Storage* task of 330 min, during which the quality control and/or properties corrections should take place. On the contrary, the processing time of the *Filling* orders is dependent on the selected filling line, type of product, batch volume and type of containers format. Table 1 resumes the data related to each production stage.

Table 1: Processing time and resources allocation in each production stages

Stage	**Processing Time (min)**	**Resources**
Preparation	150	M_i; HRa team member
Storage	330	T_i
Filling	Variable	L_i; T_i(previous task); HRb team member

Regarding the time discretization, the analysis of all processing times has considered 30 min slots as an adequate representation of time. Based on existent production records, the data used in our case-study is shown in Table 2. A plan of 28 production orders are considered for 5 shifts horizon (196 kL), with different types of paint ($Textured, Smooth$ or $Brand$) and different containers format orders ($Large, Medium$ or $Small$ size). For example, product G is a Smooth paint type and has a demand for two orders of 9 kL, which will represent 4 kL produced in Large, 3 kL in Medium and 2 kL in Small containers; it will be manufactured in M_7 using one of the team members of HRa, stored in one of T_F tanks and filled in L_3 by a HRb team member; processing times of each task are 150 min, 330 min and 240 min, respectively.

Table 2: Production plan data with resource allocation and associate processing times

							Resources/Proc.Time(min)		
Production Plan			**Order(kL)**	**Fill.Order/Size(kL)**			Machine	Tank	Fill.Line
Product	Type		.	L	M	S	**[HRa]**		**[HRb]**
A	Text.		10	8	2		M_5 / 150	T_T / 330	L_1 / 390
B	Text.	2x	6	6			M_6 / 150	T_T / 330	L_1 / 210
C	Text.		8	8			M_5 / 150	T_T / 330	L_1 / 270
D	Smo.		6	4	2		M_5 / 150	T_F / 330	L_2 / 120
E	Smo.		9	8		1	M_6 / 150	T_F / 330	L_2 / 180
F	Smo.		6	3,5	2	0,5	M_5 / 150	T_F / 330	L_3 / 120
G	Smo.	2x	9	4	3	2	M_7 / 150	T_F / 330	L_3 / 240
H	Smo.		7	7			M_6 / 150	T_F / 330	L_2 / 90
I	Smo.	2x	6	4	1	1	M_6 / 150	T_F / 330	L_2 / 150
J	Smo.	3x	8	7	1		M_7 / 150	T_F / 330	L_3 / 120
K	Smo.		5		5		M_7 / 150	T_F / 330	L_3 / 90
Y	Smo.		6		5	1	M_7 / 150	T_F / 330	L_3 / 120
M	Smo.	3x	10	8	2		M_5 / 150	T_F / 330	L_2 / 150
N	Smo.		8	7	1		M_6 / 150	T_F / 330	L_2 / 120
O	Smo.		5	4	1		M_7 / 150	T_F / 330	L_3 / 90
P	Brand	4x	5	5			M_3 / 150	T_S / 330	L_4 / 120
Q	Brand		4	4			M_3 / 150	T_S / 330	L_4 / 90
R	Brand		6	6			M_3 / 150	T_S / 330	L_4 / 120
TOTAL (28 ORDERS)			196						

Concerning the scheduling optimization of the proposed case study, with the exhibited production plan on Table 2, two cases were explored, *Optimization Case I* and *II*, where different workforce teams were analyzed addressing a concern of the production manager of the plant. The model was implemented in GAMS (GAMS Rev 237 WIN-VS8 23.7.3 x86/MS Windows) and solved through CPLEX with a Intel Core Duo CPU U2400 1,06 GHz with 2GB RAM.

Optimization Case I

The first case considers teams HRa and HRb with three elements assigned in each, as a practice of the company, to accomplish the total demand plan of 196 kL. The optimal solution obtained is able to satisfy the total demand plan (196 kL). This result is compared to real production records (161 kL) and is verified that the model results provide a significant increase on the production efficiency (196 kL versus 161 kL). Regarding the HR teams allocation, and consenting a delay time gap between the manufacturing and filling verified in first/last shifts, HRa and HRb teams both shows a utilization rate of approximately 58%. This value is extensively low, and so, in terms of resources management, a new case needs to be analyzed (*Case II*).

Optimization Case II

Considering the results of *Case I*, is now analyzed the decrease in terms of team members, proposing two elements assigned to each HRa and HRb. Schedule results are yet shown in a Gantt chart in Figure 3. The Gantt chart illustrates the optimal sequence of production orders and the detailed schedule with starting and finishing times of the involved tasks. Additionally, each resource allocation is identified by task and accounted the utilization of available team workers trough the time line.

The model generates a new integer solution considering the scenario of shortened workforce. This limits the process flow and is unable to fully accomplish the plan in the time horizon: the total output is 175 kL, 89% of the initial plan, with only three orders not satisfied (orders A, one B and one P). However, with this new schedule, we verify the increase of the utilization rate of teams: HRa to 78% and HRb to 73%. Table 3 summarizes the results of teams' utilization rate and Table 4 identifies the main GAMS statistics. A time limit of 3600 CPU seconds was established in accordance with the maximum time allowed by the production manager to get a new schedule.

After discussing the results with the production manager, two new hypotheses were studied that consider elements' decrease in only one of teams: *Case IIA* considers HRa and HRb with three and two elements assigned, respectively; in *Case IIB* the opposite is considered. The optimal model results are shows in Table 3. Consistently with the previous analysis, the increased available workforce lead to a raise on the production output when compared to *Case II*, with a reduction of the respective utilization team rate. Although, the total demand planned of 196 kL was not able to be completed. The results gathered from this study provided important insights to the production manager and were identified as very useful to manage operations at the plant level.

Table 3: Main results of optimization cases

	Case I	**Case II**	**Case IIA**	**Case IIB**	**Real Production Records**
Production Plan (kL)	196	175	186	187	161
#HRa - #HRb elements	3 - 3	2 - 2	3 - 2	2 - 3	3 - 3
HRa utilization rate (%)	58	78	56	81	-
HRb utilization rate (%)	58	73	79	56	-

Table 4: GAMS model statistics

	Single Equations	**Single Variables**	**Discrete Variables**	**Relative Gap (%)**	**MIP Solution (kL)**	**Resource usage (s)**
Case I				0,0	196	7,8
Case II	28.302	17.092	4.374	3,3	175	3.600,0
Case IIA				0,0	186	62,3
Case IIB				0,0	187	7,6

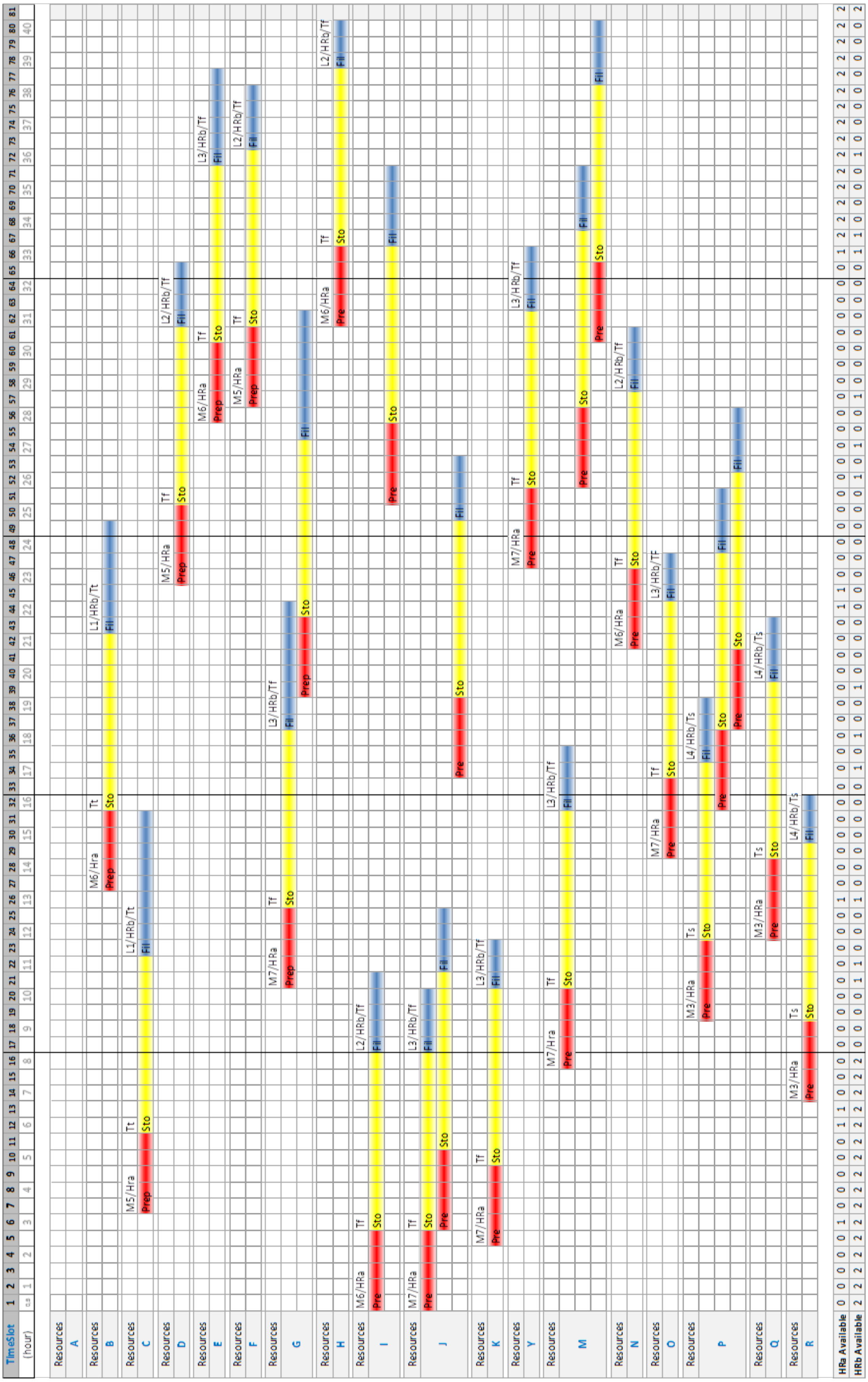


Figure 3: Gantt chart of the scheduling problem, with HRa and HRb teams of two elements each (Optimization Case II)

5 Conclusions

Based on a RTN framework model, a real scheduling production problem was solved. The sequence of production orders, processing tasks and allocated resources was detailed, maximizing the production output. The scheduling results revealed a 22% increase of the output when compared with real production records for similar conditions. Realizing that these results showed a low utilization rate of the teams, a new case was analyzed where a decrease of the team elements was considered. The results with teams HRa and HRb assigned with two elements (*Case II*) have shown an output of 9% above real production records, as well as an increased utilization rate of the workforce to more adequate levels. Consistently, mixing teams of two and three elements have provided improved results in the production output, but with lower utilization rate of the team with more members (*Cases IIA and IIB*). Only the case with two teams of three elements (*Case I*) was able to complete the entire production plan of 196 kL. Nevertheless, the results of the case-study have provided important information that can support the management of the daily process schedule. Further analysis to process conditions that were not yet considered, such as, setup times, processing costs, among others, should also be performed in future works. It is acknowledged that the mathematical formulation of the RTN framework potentiates a uniform formulation of scheduling complex systems and comprehensively problem solving.

Acknowledgments

The authors would like to thank Tintas Robbialac SA for their grateful support and acknowledge the financial support of Fundação para a Ciência e Tecnologia under the grant SFRH/BD/51594/2011.

References

[Barbosa-Póvoa, 2007] Barbosa-Póvoa, A.P. (2007). A critical review on the design and retrofit of batch plants. *Computers and Chemical Engineering*, 31(7), 833–855.

[Chibeles-Martins et al., 2010] Chibeles-Martins, N., Pinto, T., Barbosa-Póvoa, A., Novais, A. Q. (2010). A Meta-Heuristics Approach for the Design and Scheduling of Multipurpose Batch Plants. *Computer Aided Chemical Engineering*, 28, 1315–1320.

[Kondili et al., 1988] Kondili, E., Pantelides, C. C., Sargent, W. H. (1988). A general algorithm for short-term scheduling of batch operations. *Proc. of 3rd International Symposium on Process Systems Engineering*, Sidney, pp. 62-75.

[Méndez et al., 2006] Méndez, A. C., Cerdá, J., Grossmann, I. E., Harjunkoski, I., Fahl, M. (2006). State-of-the-art review of optimization methods for short-term scheduling of batch processes. *Computers and Chemical Engineering*, 30(6-7), 913–946.

[Pantelides, 1994] Pantelides, C. C. (1994). Unified frameworks for optimal process planning and scheduling. *Proc. of 2nd Foundations of Computer-Aided Process Operations*, Cache publications: New York, pp. 253–274.

[Pinto et al., 2008] Pinto, T., Barbosa-Póvoa, A., Novais, A. Q. (2008). Design of multipurpose batch plants: a comparative analysis between the STN, m-STN, and RTN representations and formulations. *Industrial and Engineering Chemistry Research*, 47(16), 6025–6044.

[Simaria et al., 2011] Simaria, AS and Gao, Y and Turner, R and Farid, SS (2011). Designing multi-product biopharmaceutical facilities using evolutionary algorithms. *21st European Symposium on Computer Aided Process Engineering*, pp. 286–290.

The construction of composite indicators with undesirable outputs using DEA models

Andreia Zanella, Ana S. Camanho, Teresa Galvão Dias

Faculdade de Engenharia da Universidade do Porto, Portugal
andreia.zanella@fe.up.pt
acamanho@fe.up.pt
tgalvao@fe.up.pt

Abstract

The construction of composite indicators based on Data Envelopment Analysis (DEA) assumes that the individual output indicators represent good aspects, so they are measured on a scale in which higher values correspond to better performance. However, in real-applications, both desirable and undesirable outputs indicators may be present. Although the literature addresses the construction of DEA models with undesirable outputs, this issue is not discussed in the context of evaluations using composite indicators. This paper discuss two different models that can be used to construct composite indicators with desirable outputs, for which higher levels represent better performance, and undesirable outputs, for which lower levels correspond to better performance. The specificities, strengths and weaknesses of each model are discussed. The implementation of the models is illustrated using a small example.

Keywords: Data Envelopment Analysis, composite indicator, undesirable outputs.

1 Introduction

The Data Envelopment Analysis (DEA) technique uses linear programming to evaluate the relative efficiency of a homogeneous set of Decision Making Units (DMUs) in their use of multiple inputs to produce multiple outputs. In standard DEA models, an inefficient DMU can improve its performance by increasing the levels of outputs (results obtained) or decreasing the levels of inputs (resources used). However, real applications may present both desirable and undesirable outputs and inputs. For example, in environmental performance assessment, we may have an output indicator related to quality of the water, for which more output corresponds to better performance and another output indicator related to the levels of CO_2 emissions, for which less output corresponds to better performance. In this situation, an inefficient DMU should increase the quality of the water or decrease the levels of CO_2 emissions to improve performance.

The performance measurement literature has addressed the measurement of productive efficiency in the presence of undesirable outputs since the 80s. One of the earliest studies addressing the incorporation of undesirable outputs in the assessment of production efficiency was developed by [Pittman, 1983]. This study extended the multilateral productivity indicator proposed by [Caves et al., 1982] to include measures of both desirable and undesirable outputs. The multilateral productivity indicator developed by [Caves et al., 1982] required the specification of the price data, but this information is often unavailable for undesirable factors. Therefore, [Pittman, 1983] proposed an extension of this indicator that assigned a value to the undesirable outputs based on estimates of shadow prices instead of market prices. Some years later, [Fare et al., 1993] proposed an alternative method to estimate shadow prices based on the distance function defined by [Shephard, 1970]. The specification of the shadow prices of undesirable outputs using a linear programming model allowed enhancing the approach proposed by [Pittman, 1983].

[Fare et al., 1989] also proposed a modification of [Farrell, 1957] approach to efficiency measurement to allow an asymmetric treatment of desirable and undesirable outputs. While the multilateral productivity indicator requires the specification of the price information for the undesirable outputs, the nonparametric approach of [Fare et al., 1989] only requires data on quantities of the undesirable outputs. The authors proposed a hyperbolic approach to efficiency measurement to allow considering different assumptions on the disposability of undesirable outputs. The new constraints state that the desirable outputs are strongly disposable (i.e. they can be reduced without cost), while the undesirable outputs are weakly disposable (i.e. they can be reduced only in conjunction with a reduction in the other outputs or an increase in the use of inputs).

Some years later, [Chung et al., 1997] introduced a different approach to deal with undesirable outputs in the efficiency and productivity measurement literature. The authors extended the [Chambers et al., 1996] directional distance function to allow expanding the desirable outputs while simultaneously contracting the undesirable ones. The outputs are expanded or contracted along a path that is defined according to a directional vector. The directional distance function has been widely used in the context of environmental performance assessment, in which the production of waste is often present.

The approaches mentioned above are known as direct approaches to treat undesirable outputs. These approaches allow treating the outputs in their original form, that is, without requiring any modification to the measurement scale. On the other hand, there are indirect approaches that transform the values of the undesirable outputs to allow treating them as normal outputs in traditional DEA models.

[Scheel, 2001], [Dyson et al., 2001] and [Seiford and Zhu, 2002] discussed the different approaches to handle undesirable outputs in DEA models using indirect approaches. One option is to move the variables from the output to the input side. [Scheel, 2001] pointed out that this approach results in the same technology set as incorporating the undesirable outputs as normal outputs, in the form of their additive inverses ($-y_{und}$). Regarding this option, [Seiford and Zhu, 2002] pointed out that to treat undesirable outputs as inputs would not reflect the real production process, as the input-output structure that defines the production process would be lost. The incorporation of the undesirable outputs in the form of their additive inverses was first suggested by [Koopmans, 1951]. Another possibility is to consider the undesirable outputs in the form of their multiplicative inverses ($1/y_{und}$), as proposed by [Golany and Roll, 1989]. Regarding this option, [Dyson et al., 2001] pointed out that this transformation would destroy the ratio or interval scale of the data. The third option is to add to the additive inverses of the undesirable outputs a sufficient large positive number ($-y_{und}+k$), as first suggested by [Ali and Seiford, 1990]. This transformation is the most frequently used in the literature to deal with undesirable outputs using a traditional DEA formulation [Cook and Green, 2005, Oggioni et al., 2011]. It has the advantage of enabling a simple interpretation of results, but it is sensitive to the choice of the constant k, as will be discussed in section 2.

Although the papers mentioned above approach the presence of undesirable outputs and inputs in DEA models, they do not address the modeling of undesirable factors in the construction of composite indicators (CIs). A CI is given by the aggregation of several individual indicators in a single measure. The DEA technique can be applied to construct composite indicators that look only at the achievements of a given set of DMUs, providing an effectiveness measure. The use of DEA for performance assessments focusing only on achievements, rather than the conversion of inputs to outputs, was first proposed by [Cook and Kress, 1990] to construct a preference voting model (for aggregating votes in a preferential ballot). Others relevant studies that support the empirical use of DEA models only with outputs can be found in different fields, such as macroeconomic performance assessment [Lovell et al., 1995], selective examinations for university entrance [Hashimoto, 1996], university quality assessment [Murias et al., 2007], human development [Mahlberg and Obersteiner, 2001, Despotis, 2004, Despotis, 2005], technology achievement [Cherchye et al., 2008] and evaluation of urban quality of life [Morais and Camanho, 2011]. In these studies, all variables were specified as outputs and an identical input level, which for simplicity was assumed to be equal to one, was specified for all DMUs.

In this paper, we approach the construction of CIs that include both desirable and undesirable factors with aggregation procedures based on DEA. Two approaches will be discussed in this context: An indirect approach, based on a traditional DEA model that requires a transformation in the measurement scale of the undesirable outputs and, a direct approach, based on a DEA model specified using a directional distance function. The implementation of these approaches are illustrated using a small example.

2 Undesirable outputs in DEA and the construction of composite indicators

In this section we discuss the main models that have been used to treat undesirable outputs in DEA efficiency assessments. This is followed by the presentation of CI models that can be obtained based on these DEA formulations. As mentioned in the introduction, the models can follow a direct or an indirect approach to handle the undesirable outputs.

2.1 Indirect approach

The DEA output oriented model by [Charnes et al., 1978] can be used for assessments involving undesirable outputs with a transformation in the measurement scale of the undesirable outputs as proposed by [Seiford and Zhu, 2002]. The resulting model, with a constant returns to scale formulation in shown in (2.1).

$$
\begin{aligned}
&\delta = \min \sum_{i=1}^{m} v_i \; x_{ij} && (2.1)\\
&s.t. \sum_{r=1}^{s} u_r \; y_{rj_0} + \sum_{t=1}^{l} p_t \; (M_t - b_{tj_0}) = 1 \\
&\sum_{r=1}^{s} u_r \; y_{rj} + \sum_{t=1}^{l} p_t \; (M_t - b_{tj}) - \sum_{i=1}^{m} v_i \; x_{ij} \leq 0 \quad j = 1, \ldots, n \\
&u_r \geq 0 \qquad r = 1, \ldots, s \\
&p_t \geq 0 \qquad t = 1, \ldots, l \\
&v_i \geq 0 \qquad i = 1, \ldots, m
\end{aligned}
$$

In formulation (2.1), the efficiency score for DMU j_0 is given by $1/\delta$, and it ranges between 0 (worst) and 1 (best). x_{ij} $(i = 1, ..., m) \in \Re_0^+$, y_{rj} $(r = 1, ..., s) \in \Re_0^+$ and b_{tj} $(t = 1, ..., l) \in \Re_0^+$ correspond to the value of the input i, desirable output r and undesirable output t, respectively, for DMU j $(j = 1, ..., n)$. v_i, u_r and p_t are the weights attached to the inputs, desirable outputs and undesirable outputs, respectively, in the performance assessment. M_t is a large positive number greater or equal to the maximum value of the undesirable output t observed in all DMUs. In the transformation proposed by [Seiford and Zhu, 2002] all of the translated outputs are positive, so the constant must be a value larger than the maximum observed in each output indicator. As the model is sensitive to the choice of the value for M_t, a sensitivity analysis of the results for different values of M_t should be undertaken.

In a CI we have only outputs to be aggregated, so we can assume that all DMUs are similar in terms of inputs. Thus, following [Koopmans, 1951], [Lovell, 1995] and [Lovell et al., 1995], we can use a unitary input underlying the evaluation of every DMU, interpreted as a "helmsman" attempting to steer the DMUs towards the maximization of DMUs. By considering a unitary input level for all DMUs in model (2.1) we obtain the CI model presented in (2.2).

$$
\begin{aligned}
&\delta = \min v && (2.2)\\
&s.t. \sum_{r=1}^{s} u_r \; y_{rj_0} + \sum_{t=1}^{l} p_t \; (M_t - b_{tj_0}) = 1 \\
&\sum_{r=1}^{s} u_r \; y_{rj} + \sum_{t=1}^{l} p_t \; (M_t - b_{tj}) - v \leq 0 \quad j = 1, \ldots, n \\
&u_r \geq 0 \qquad r = 1, \ldots, s \\
&p_t \geq 0 \qquad t = 1, \ldots, l \\
&v \geq 0
\end{aligned}
$$

The use of DEA to construct CIs was popularized by [Cherchye et al., 2007]. This approach is known as "benefit of the doubt" construction of composite indicators. The CI proposed by [Cherchye et al., 2007] is equivalent to the input oriented model proposed by [Charnes et al., 1978], assuming constant returns to scale and an unitary input level for all DMUs. A CI can also be obtained from the output oriented model, and this was the approach followed in this paper. As both formulations assume constant return to scale, the performance scores obtained with different orientations are the same. The advantage of using the output oriented formulation is that it leads to a more direct estimation of targets using the dual model, and facilitates the incorporation of weight restrictions.

For benchmarking purposes, the identification of the peers and targets for the inefficient DMUs can be done through the envelopment formulation of model (2.2), shown in (2.3). The objective function value at the optimal solution of the model (2.3) corresponds to the factor δ by which all outputs of the DMU under assessment can be proportionally improved to reach the target output values. The performance score, or composite indicator, of DMU j_0 under assessment is the reciprocal of this value. Therefore, the DMUs with the best performance are those for which there is no evidence that it is possible to expand their outputs, such that the value of δ^* is equal to 1.

$$
\begin{aligned}
&\max \delta && && (2.3)\\
&s.t.\ \ \delta\, y_{rj_0} - \sum_{j=1}^{n} \lambda_j\, y_{rj} \leq 0 && r = 1, \ldots, s \\
&\delta\, (M_t - b_{tj_0}) - \sum_{j=1}^{n} \lambda_j\, (M_t - b_{tj}) \leq 0 && t = 1, \ldots, l \\
&\sum_{j=1}^{n} \lambda_j \leq 1 && j = 1, \ldots, n \\
&\lambda_j \geq 0 && j = 1, \ldots, n
\end{aligned}
$$

The peers for the DMU j_0 under assessment are the DMUs with values of λ_j^* greater than zero in the optimal solution of model (2.3). The targets that a DMU j_0, with a composite indicator score smaller than one, should reach to improve performance are given by the following expressions:

$$
\begin{cases}
{y_{rj_0}}^{target} = \sum_{j=1}^{n} \lambda_j^*\, y_{rj} & r = 1, ..., s \\
{b_{tj_0}}^{target} = M_t - \sum_{j=1}^{n} \lambda_j^*\, (M_t - b_{tj}) & t = 1, ..., l
\end{cases}
\qquad (2.4)
$$

2.2 Direct approach

The efficiency evaluation using the directional distance function (DDF), developed by [Chambers et al., 1996], allows to simultaneously expand outputs and contract inputs according to a directional vector. [Chung et al., 1997] extended this approach to allow including undesirable outputs in the efficiency evaluation. The [Chung et al., 1997] model also assumes weak disposability of undesirable outputs, as proposed in [Fare et al., 1989], but preserves the linearity of the DEA model. The constant returns to scale model of [Chung et al., 1997] is specified as shown in (2.5).

$$
\begin{aligned}
&\max \delta && && (2.5)\\
&s.t. \sum_{j=1}^{n} y_{rj}\ \lambda_j \geq y_{rj_0} + \delta\ g_{y_{rj_0}} && r = 1, \ldots, s \\
&\sum_{j=1}^{n} b_{tj}\ \lambda_j = b_{tj_0} - \delta\ g_{b_{tj_0}} && t = 1, \ldots, l \\
&\sum_{j=1}^{n} x_{ij}\ \lambda_j \leq x_{rj_0} - \delta\ g_{x_{ij_0}} && i = 1, \ldots, m \\
&\lambda_j \geq 0 && j = 1, \ldots, n
\end{aligned}
$$

In formulation (2.5), x_{ij} $(i = 1, ..., m)$ are the inputs used by the DMU j $(j = 1, ..., n)$ to produce y_{rj} $(r = 1, ..., s)$ desirable outputs and b_{tj} $(t = 1, ..., l)$ undesirable outputs. The λ_j are the intensity variables. The components of vector $g = (g_y, -g_b, -g_x)$ indicate the direction of change for the outputs and inputs. Positive values for the components are associated to expansion of desirable outputs and negative values are associated to contraction of inputs and undesirable outputs. The factor δ indicates the extend of DMU's inefficiency. It corresponds to the maximal feasible expansion of desirable outputs and contraction of inputs and undesirable outputs that can be achieved simultaneously.

While inputs and desirable outputs are assumed to be strongly disposable, the undesirable outputs are assumed to be weakly disposable, as it is shown by the equality in the constraint associated to the undesirable outputs. When imposing weak disposability of undesirable outputs we are assuming that they are by-products of the desirable outputs and cannot be reduced without cost, which implies that abatement in an undesirable output is possible if accompanied by a reduction in a desirable output or an increase in an input. The decision on whether to assume strong disposability or weak disposability for the variables of a DEA model depends on the nature of the application under analysis [Liu et al., 2009].

Using a unitary level of input and setting the directional vector as $g = (g_y, -g_b, 0)$ (a vector that allows to, simultaneously, expand the desirable outputs and contract undesirable ones by keeping inputs fixed), the input restriction becomes $\sum_{j=1}^{n} \lambda_j \leq 1$. So, in the context of performance evaluations based on CI, model (2.5) reduces to the formulation shown in (2.6). In the remainder of this paper, we refer to this formulation as Directional CI model.

$$
\begin{aligned}
&\max \delta \\
&s.t. \sum_{j=1}^{n} y_{rj}\ \lambda_j \geq y_{rj_0} + \delta\ g_{y_{rj_0}} && r = 1, \ldots, s \\
&\sum_{j=1}^{n} b_{tj}\ \lambda_j = b_{tj_0} - \delta\ g_{b_{tj_0}} && t = 1, \ldots, l \\
&\sum_{j=1}^{n} \lambda_j \leq 1 \\
&\lambda_j \geq 0 && j = 1, \ldots, n
\end{aligned}
\tag{2.6}
$$

The composite indicator score, ranging from 0 (worst) to 1 (best) is given by $1 - \delta^*$, and corresponds to an efficiency score. As by-products of the assessment using the Directional CI model, it is possible to identify the peers and targets for the inefficient DMUs. These are given by the expressions shown in (2.7).

$$
\begin{cases}
{y_{rj_0}}^{target} = \sum_{j=1}^{n} \lambda_j^*\ y_{rj} & r = 1, ..., s \\
{b_{tj_0}}^{target} = \sum_{j=1}^{n} \lambda_j^*\ b_{tj} & t = 1, ..., l
\end{cases}
\tag{2.7}
$$

3 Illustrative application

This section illustrates the application of the two approaches to construct DEA-based CIs including desirable and undesirable outputs, presented in Section 2.

Our illustrative example consists of a set of 9 DMUs. To allow a graphical illustration of the models, these DMUs are assessed considering two outputs indicators: Y, a desirable output, and B, an undesirable output. Table 1 shows the data for the 9 DMUs.

Table 1: Output indicators for the illustrative example

DMU	Y (desirable)	B (undesirable)
A	5	7
B	25	28
C	22	15
D	10	18
E	30	30
F	20	22
G	19.8	17.8
H	20.8	19.5
I	19.08	19.66

3.1 Indirect approach illustration

Figure 1 illustrates the production possibility set for the illustrative example presented in Table 1, corresponding to an evaluation using model (2.2). As lower values in the output indicator B correspond to better performance, the technically efficient frontier is given by the segments linking DMUs A, C and E.

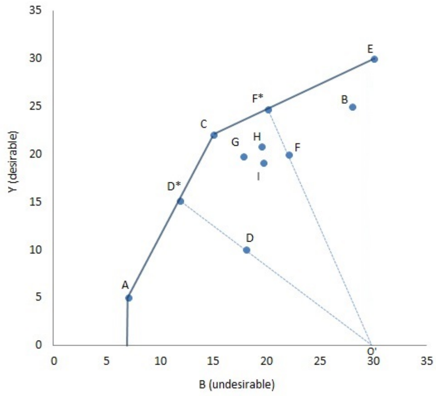


Figure 1: Production possibility set for the indirect approach

Table 2 shows the performance scores, peers and targets for the 9 DMUs obtained using model (2.2) with a value of M equal to 30 (i.e., the largest value observed for the undesirable output indicator B). The composite indicator (or efficiency score) for a DMU is given by $1/\delta$ in the optimal solution of model (2.2). For example, for DMU F the CI is 0.809, which corresponds to the ratio $\frac{O'F}{O'F^*}$ in Figure 1. The target that the DMU F should achieve to improve performance and reach the frontier of the production possibility set is given by the point F*, which corresponds to the value 24.725 for the output indicator Y and 20.110 for the output indicator B. The peers for DMU F are DMUs C and E, with values of λ_C and λ_E equal to 0.659 and 0.341, respectively. The value of λ_j provides an indication of the degree of similarity between a DMU and its peers.

Table 2: Composite indicator, peers and targets obtained using model (2.2), with M equal to 30

DMU	CI ($1/\delta$)	Peers (λ)	Target for Y	Target for B
A	1	A(1)	5	7
B	0.869	C (0.153); E (0.847)	28.772	27.698
C	1	C(1)	22	15
D	0.659	A (0.401); C (0.599)	15.176	11.789
E	1	E(1)	30	30
F	0.809	C (0.659); E (0.341)	24.725	20.110
G	0.877	C (0.928); C (0.072)	22.580	16.087
H	0.880	C (0.795); E (0.205)	23.636	18.068
I	0.820	C (0.841); E (0.159)	23.273	17.387

With the transformation in the measurement scale of the undesirable output, the computation of the composite indicator uses as basis a point that no longer coincides with the origin, as in standard DEA models. Instead, it corresponds to a new reference point that depends on the value of the constant M used in model (2.2). In our illustrative example, the composite indicator is estimated considering that the axis of the undesirable output starts at the worst value observed in the original measurement scale (30, corresponding to DMU E), as shown in Figure 1.

As mentioned in Section 2, the transformation performed by subtracting the values of undesirable outputs from a large positive number (M) has an impact on the results of the DEA model. Table 3 presents the results that would be obtained by specifying different values for M. Note that, as the value of M increases, the discrimination between the DMUs' scores decreases. In addition, the improvement required for the DMUs to reach the frontier becomes more demanding for the undesirable output and less demanding for the desirable output. This effect can be seen in Figure 2, that shows the projection obtained using M equal to 35 and 60, respectively.

By using a value of M larger than the maximum observed for the undesirable output, the DMU's classification, as efficient or inefficient, remains unchanged, but the DMUs' efficiency score changes, and the ranking of DMUs may also be different. For example, Figure 2 shows that when the constant M is specified as equal to 60, DMU G is projected to the segment linking DMUs A and C instead of the segment linking C and E, where it was projected when M was specified as equal to 35. This led to a change in the efficiency ranking of DMUs. Thus, using the indirect approach to construct a composite indicator, it is only possible to ensure that the assessment is classification invariant for different values of the constant M, as stated by [Seiford and Zhu, 2002].

Table 3: Composite indicator and ranks for different values of M

DMU	$M = 30$		$M = 35$		$M = 60$	
	CI	Rank	CI	Rank	CI	Rank
A	1	1	1	1	1	1
B	0.869	6	0.880	6	0.914	6
C	1	1	1	1	1	1
D	0.659	9	0.715	9	0.844	9
E	1	1	1	1	1	1
F	0.809	8	0.824	8	0.875	8
G	0.877	5	0.887	5	0.931	4
H	0.880	4	0.890	4	0.928	5
I	0.820	7	0.834	7	0.891	7

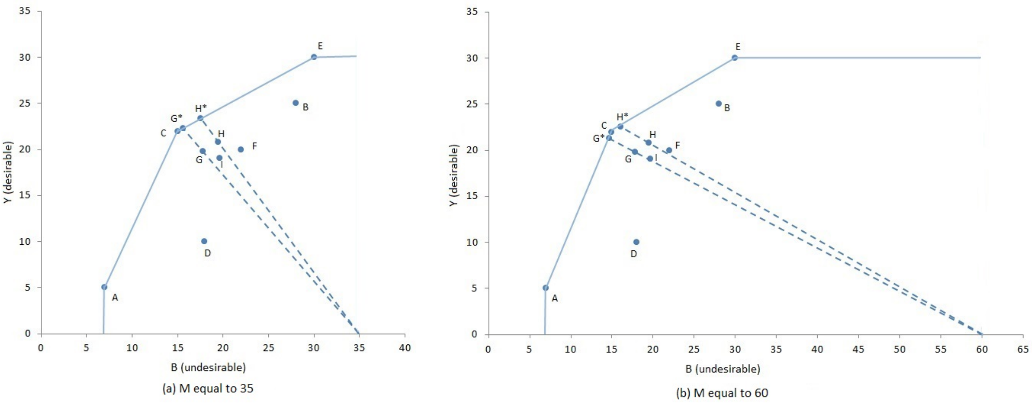


Figure 2: Sensibility analysis for different values of M

3.2 Direct approach illustration

This section illustrates the estimation of composite indicators using the Directional CI model. Figure 3 shows the production frontier that would be obtained for our illustrative example (Table 1) using the Directional CI model. The efficient frontier is defined by the segments linking O, C and E.

By setting the directional vector as $\mathbf{g} = (g_y, -g_b) = (y_{rj_0}, -b_{tj_0})$, i.e. the current value of the outputs for the DMU under assessment, it is possible to simultaneously expand the desirable outputs and contract the undesirable outputs through a path that allows proportional interpretation of improvements. In order to facilitate the interpretation of the DMUs projection to the frontier, Figure 3 illustrates the directional vectors and the projection of DMUs D and F on the frontier, corresponding to points D* and F*, respectively. Note that, for each DMU, the desirable and undesirable outputs are expanded and contracted, respectively, according to a direction that corresponds to proportional changes to the original levels.

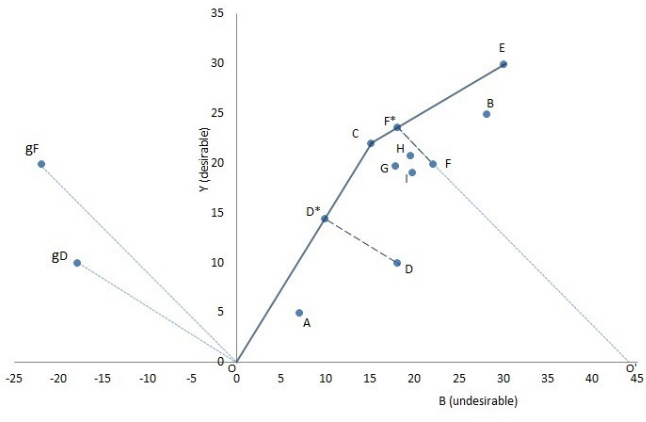


Figure 3: Production possibility set for the direct approach

Table 4 shows the composite indicator, peers and targets obtained using the Directional CI model. The value of δ^* obtained at the optimal solution to the model can be interpreted as the scope for improvement of a given DMU. For example, the value of δ^* for DMU F is 0.181, corresponding to $\frac{FF^*}{O'F}$ in Figure 3. The CI score, equal to $1 - \delta^*$, is given by $\frac{O'F - FF^*}{O'F}$, and is equal to 0.819 for DMU F. The point F* is the target that the DMU F should achieve to become efficient, i.e., to operate at the frontier, which corresponds to the value 23.613 for the output indicator Y and 18.025 for the output indicator B. The peers for DMU F are DMUs C and E, with values of λ_C and λ_E equal to 0.798 and 0.202, respectively.

Table 4: Composite indicator, rank, peers and targets obtained from the Directional CI model

DMU	CI $(1-\delta)$	Rank	Peers (λ)	Target for Y	Target for B
A	0.655	8	C (0.306)	6.725	4.585
B	0.902	3	C (0.317); E (0.683)	27.462	25.242
C	1	1	C(1)	22	15
D	0.549	9	C (0.659)	14.505	9.890
E	1	1	E(1)	30	30
F	0.819	6	C (0.798); E (0.202)	23.613	18.025
G	0.874	5	C (0.963); E (0.037)	22.296	15.556
H	0.885	4	C (0.850); E (0.150)	23.200	17.250
I	0.817	7	C (0.929); E (0.071)	22.567	16.063

Note that if instead of assuming weak disposability of outputs, it was imposed strong disposability, the frontier would be given by the horizontal extension of E. In the context of an assessment using a CI, we believe that weak disposability is more appropriate, since it avoids unrealistic projections (projections involving huge reductions to undesirable outputs and increments to desirable outputs, eventually leading to unrealistic targets corresponding to negative values of the undesirable outputs).

3.3 Discussion regarding the alternative formulations of the CI model

As shown in the illustrative example, the efficient frontiers obtained using the direct and indirect approaches are different (see Figures 1 and 3). The weak disposability assumption imposed to the Directional CI model forces the frontier to pass through the origin. Conversely, the assessment using model (2.2) results in a frontier that does not pass through the origin. This implies that the classification of a DMU as efficient or inefficient depends on the model used. Furthermore, the performance scores and ranking of DMUs obtained with the direct and indirect approaches are also different.

The advantages of using the Directional CI model, shown in (2.6), instead of the DEA model, shown in (2.2), is that it allows to estimate, simultaneously, the inefficiencies associated to desirable and undesirable outputs without any transformation in the measurement scale of undesirable outputs. Furthermore, by setting the components of the directional vector equal to the current value of the outputs for the DMU under assessment, it is possible to preserve the proportional interpretability of the improvements. As the DEA-based CI specified in (2.2) computes all projections in relation to a point that does not coincide with the origin, it does not allow a proportional change to both desirable and undesirable outputs, and the CI score depends on the values chosen for the constants M_t.

4 Conclusions

The traditional DEA based-CI models for performance assessment cannot be used in the presence of both desirable and undesirable outputs indicators, as they do not seek for reductions to the undesirable outputs. In this paper, we discussed two different approaches that can be used for the construction of CI: a direct approach, based on a DEA model specified using a directional distance function, that allows to deal with the undesirable outputs in their original measurement scale and, an indirect approach, that uses a traditional DEA model and includes a transformation in the measurement scale of undesirable outputs. This transformation discussed in this paper consists on subtracting the values of the undesirable output indicators from a sufficiently large positive number, equal or larger than the maximum observed in the original data for each output in the sample.

In order to explain the features of the approaches discussed in this paper we illustrated their implementation using a small example. It was demonstrated that the indirect approach has implications in the direction of the DMUs' projection to the frontier. After the transformation, the reference point used to compute the measure of performance is no longer the origin, as in standard DEA models, but a new reference point, whose value is equal to the positive number used to transform the measurement scale of the undesirable output. As a result, this approach does not allow proportional improvements to both desirable and undesirable outputs. Furthermore, it was shown that the indirect approach is sensitive to the value of the constant used for the transformation. Different values for the constant have implications both in the performance scores and ranking of the DMUs.

Conversely, using the direct approach to deal with the undesirable outputs, it is possible to preserve the proportional interpretability of the improvements by setting the components of the directional vector equal to the values of the desirable and undesirable outputs of the DMU under assessment. This is an important advantage of the direct approach discussed in this paper, and thus we argue that it is the most appropriate for constructing composite indicators in the presence of both desirable and undesirable outputs.

Acknowledgements

The authors acknowledge the financial support of the Portuguese Foundation for Science and Technology (FCT) through project iTeam (Pt/SES-SUES/0041/2008) and Mesur (PTDC/SEN-ENR/111710/2009).

References

[Ali and Seiford, 1990] Ali, A. I. and Seiford, L. M. (1990). Translation-invariance in data envelopment analysis. *Operations research letters*, 9(6):403–405.

[Caves et al., 1982] Caves, D. W., Christensen, L. R., and Diewert, E. (1982). Multilateral comparisons of output, input, and productivity using superlative index numbers. *The Economic Journal*, 92(365):73–86.

[Chambers et al., 1996] Chambers, R. G., Chung, Y., and Fare, R. (1996). Benefit and distance functions. *Journal of Economic Theory*, 70(2):407–419.

[Charnes et al., 1978] Charnes, A., Cooper, W. W., and Rhodes, E. (1978). Measuring the efficiency of decision making units. *European Journal of Operational Research*, 2(6):429–444.

[Cherchye et al., 2007] Cherchye, L., Moesen, W., Rogge, N., and Puyenbroeck, T. V. (2007). An introduction to Benefit of the doubt composite indicators. *Social Indicators Research*, 82(1):111–145.

[Cherchye et al., 2008] Cherchye, L., Moesen, W., Rogge, N., Van Puyenbroeck, T., Saisana, M., Saltelli, A., Liska, R., and Tarantola, S. (2008). Creating composite indicators with DEA and robustness analysis: the case of the technology achievement index. *Journal of the Operational Research Society*, 59(2):239–251.

[Chung et al., 1997] Chung, Y., Fare, R., and Grosskopf, S. (1997). Productivity and undesirable outputs: A directional distance function approach. *Journal of Environmental Management*, 51(3):229–240.

[Cook and Kress, 1990] Cook, W. and Kress, M. (1990). A data envelopment model for aggregating preference rankings. *Management Science*, 36(11):1302–1310. WOS:A1990EN62300002.

[Cook and Green, 2005] Cook, W. D. and Green, R. H. (2005). Evaluating power plant efficiency: a hierarchical model. *Computers & Operations Research*, 32(4):813–823.

[Despotis, 2005] Despotis, D. (2005). Measuring human development via data envelopment analysis: the case of asia and the pacific. *Omega*, 33(5):385–390.

[Despotis, 2004] Despotis, D. K. (2004). A reassessment of the human development index via data envelopment analysis. *Journal of the Operational Research Society*, 56(8):969–980.

[Dyson et al., 2001] Dyson, R. G., Allen, R., Camanho, A. S., Podinovski, V. V., Sarrico, C. S., and Shale, E. A. (2001). Pitfalls and protocols in DEA. *European Journal of Operational Research*, 132(2):245–259.

[Fare et al., 1993] Fare, R., Grosskopf, S., Lovell, C., and Yaisawarng, S. (1993). Derivation of shadow prices for undesirable outputs: A distance function approach. *The review of economics and statistics*, 75(2):374–380.

[Fare et al., 1989] Fare, R., Grosskopf, S., Lovell, C. A. K., and Pasurka, C. (1989). Multilateral productivity comparisons when some outputs are undesirable: A nonparametric approach. *Review of Economics and Statistics*, 71(1):90–98.

[Farrell, 1957] Farrell, M. J. (1957). The measurement of productive efficiency. *Journal of the Royal Statistical Society*, 120:253–281.

[Golany and Roll, 1989] Golany, B. and Roll, Y. (1989). An application procedure for DEA. *Omega-International Journal of Management Science*, 17(3):237–250. WOS:A1989U711000005.

[Hashimoto, 1996] Hashimoto, A. (1996). A DEA selection system for selective examinations. *Journal of the Operations Research Society of Japan*, 39(4):475–485. WOS:A1996WA33300002.

[Koopmans, 1951] Koopmans, T. (1951). Analysis of production as an efficient combination of activities. In *Activity Analysis of Production and Allocation*, pages 33–97. John Wiley and Sons, New York.

[Liu et al., 2009] Liu, W. B., Meng, W., Li, X. X., and Zhang, D. Q. (2009). DEA models with undesirable inputs and outputs. *Annals of Operations Research*, 173(1):177–194.

[Lovell et al., 1995] Lovell, C., Pastor, J., and Turner, J. (1995). Measuring macroeconomic performance in the OECD: a comparison of european and non-european countries. *European Journal of Operational Research*, 87:507–518.

[Lovell, 1995] Lovell, C. K. (1995). Measuring the macroeconomic performance of the taiwanese economy. *International Journal of Production Economics*, 39(1-2):165–178.

[Mahlberg and Obersteiner, 2001] Mahlberg, B. and Obersteiner, M. (2001). Remeasuring the HDI by data envelopment analysis. Laxenburg, austria, International Institute for Applied Systems Analysis, Laxenburg, Austria.

[Morais and Camanho, 2011] Morais, P. and Camanho, A. S. (2011). Evaluation of performance of european cities with the aim to promote quality of life improvements. *Omega*, 39(4):398–409.

[Murias et al., 2007] Murias, P., Miguel, J. C., and Rodriguez, D. (2007). A composite indicator for university quality assesment: The case of spanish higher education system. *Social Indicators Research*, 89(1):129–146.

[Oggioni et al., 2011] Oggioni, G., Riccardi, R., and Toninelli, R. (2011). Eco-efficiency of the world cement industry: A data envelopment analysis. *Energy Policy*, 39(5):2842–2854.

[Pittman, 1983] Pittman, R. W. (1983). Multilateral productivity comparisons with undesirable outputs. *Economic Journal*, 93(372):883–91.

[Scheel, 2001] Scheel, H. (2001). Undesirable outputs in efficiency valuations. *European Journal of Operational Research*, 132(2):400–410.

[Seiford and Zhu, 2002] Seiford, L. M. and Zhu, J. (2002). Modeling undesirable factors in efficiency evaluation. *European Journal of Operational Research*, 142(1):16–20.

[Shephard, 1970] Shephard, R. W. (1970). *Theory of cost and production functions.* Princeton University Press.

Índice de autores

Alcina Nunes, 48
Amélia de Lorena Stanzani, 333
Amaro de Sousa, 323
Ana I. Pereira, 194, 341
Ana Maria A.C. Rocha, 306
Ana Maria Rodrigues, 313
Ana Moura, 120, 275
Ana Paula Barbosa-Póvoa, 27, 88, 98, 262, 286, 367, 378
Ana Paula Monte, 9
Ana Paula Teixeira, 347
Ana Respício, 38
Ana S. Camanho, 108, 185, 386
Ana S. R. Brásio, 67
Andreia Zanella, 386
Andrey Romanenko, 67
Angela Paula Ferreira, 356
António B. Fernandes, 145
António Trigo, 58
Antonio Roberto Balbo, 201, 333
António J.S.T. Duarte, 126
Armando Barbosa, 135

B. Taboada, 341
Bernardo Almada-Lobo, 16
Bruno Bastos, 58

Camila de Lima, 201
Carla A. S. Geraldes, 172
Carla Sofia Pereira, 135
Carlos Balsa, 48
Carlos Lopes, 323
Clara Bento Vaz, 356
Conceição Silva Portela, 108

D. Bento, 341
D. Pinho, 341
Diana F. Lopes, 212
Dorabella Santos, 323

Edite M.G.P. Fernandes, 306
Elisa Barros, 48
Eloísa Macedo, 223

Filipe Alvelos, 38, 244, 323

Guilherme Pereira, 172

Helena Mouta, 271
Helenice de Oliveira Florentino Silva, 201

Isabel M. Horta, 185
Isabel Martins, 244

J. G. Borges, 154
J. Garcia-Gonzalo, 154
J. Orestes Cerdeira, 77, 120
J.M.V. Valério de Carvalho, 126
João Luís H. Matias, 254
João Miranda, 347
João Paulo Costa, 162
Joana Fialho, 162
Joana Matos Dias, 233
Joaquim Jorge Vicente, 367
Jorge Gonçalves, 296
Jorge Oliveira, 275
Jorge P. Sousa, 262
José Gonçalves, 194
José Lima, 194
José Soeiro Ferreira, 313

L. Ferreira, 154

M. Fernanda P. Costa, 306
M. Teresa T. Monteiro, 254
Mónica Oliveira, 88
Manuel Cruz, 120
Maria do Céu Marques, 233
Maria Isabel Gomes, 27, 286
Maurício A. Vaz, 145
Miguel Constantino, 154, 244
Miguel Vieira, 378

Natércia C. P. Fernandes, 67
Nuno Mota, 296
Nuno Moutinho, 271

Paula Fernandes, 9
Paulo Costa, 194
Pedro Amorim, 16
Pedro Fernandes, 135
Pedro Godinho, 162
Pedro Martins, 58

R. Lima, 341
Raul Brás, 77
Ricardo A. S. Castro, 108
Ricardo Magalhães, 244

Sónia R. Cardoso, 98
Sameiro Carvalho, 172
Samuel Moniz, 262
Sandra Afonso, 9
Stefan Nickel, 88
Susana Baptista, 27
Susana Relvas, 98, 296, 367

Tânia Pinto-Varela, 378
Tânia Ramos, 286
Teófilo Miguel M. Melo, 254
Teresa Cardoso, 88
Teresa Galvão Dias, 185, 386
Thiago Pedro Donadon Homem, 201
Tiago Heleno, 58

Vítor Barbosa, 38

Agradecimentos

Os editores deste livro de atas publicamente agradecem aos revisores dos artigos, pelo excelente trabalho realizado na avaliação e melhoria dos trabalhos submetidos:

Adelaide Cerveira (DMAT-UTAD)
Agostinho Agra (DMAT-UA)
Aldina Correia (ESTGF - IPP)
Ana Camanho (FEUP-UP)
Ana Isabel Pereira (ESTG-IPB)
Ana Maria Rocha (EE-UM)
Ana Paula Teixeira (DMAT-UTAD)
Ana Viana (ISEP-IPP)
Anabela Tereso (EE-UM)
António Ascensão Costa (ISEG-UTL)
António Duarte (ESTIG-IPB)
Bernardo Almada-Lobo (FEUP-UP)
Cândida Mourão (ISEG-UTL)
Carina Pimentel (DEGEI-UA)
Carla Amado (FE-UAlg)
Carlos H. Antunes (FCT-UC)
Clara Bento Vaz (ESTG-IPP)
Cláudio Alves (EE-UM)
Cristina Requejo (DMAT-UA)
Eugénia Captivo (FCUL-UL)
Filipe Alvelos (EE-UM)
Francisco Saldanha da Gama (FCUL-UL)
Humberto Varum (DEC-UA)
Inês Marques (FCUL-UL)
Isabel Gomes (FCT-UNL)
Isabel Horta (FEUP-UP)
Isabel Martins (ISA-UTL)
Isabel Proença (ISEG-UTL)
Isabel Ribeiro (FEUP-UP)
Joana Matos Dias (FEUC – UC)
João Andrade e Silva (ISEG-UTL)
João Lourenço (IST-UTL)
João Luís Soares (FCT-UC)
João Matias (DMAT-UTAD)
João Paulo Costa (FEUC-UC)
João Pedro Pedroso (FCUP-UP)
Jorge Orestes Cerdeira (FCT-UNL)
Jose António Oliveira (EE-UM)
José Fernando Oliveira (FEUP-UP)
José Manuel Valério de Carvalho (EE-UM)
Leonor Santiago Pinto (ISEG-UTL)
Lídia Lourenço (FCT-UNL)
Luís Miguel Silva Dias (EE-UM)
Manuel Lopes (ISEP-IPP)
Manuel Matos (FEUP-UP)
Manuel Pina Marques (FEUP-UP)
Margarida Vaz Pato (ISEG-UTL)
Maria Antónia Carravilla (FEUP-UP)
Maria da Conceição da Fonseca (FCUL-UL)
Maria João Alves (FEUC-UC)
Marília Pires (FCT-UAlg)
Maristela Santos (ICMC-USP)
Marta Mateus (FCT-UNL)
Marta Pascoal (FCT-UC)
Michael Poss (CNRS, UTCompiegne)
Miguel Constantino (FCUL-UL)
Mónica Oliveira (IST-UTL)
Paula Rama (DMAT-UA)
Pedro Amorim (FEUP-UP)
Pedro Martins (ISCAC-IPC)
Pedro Oliveira (ICBAS-UP)
Ricardo Oliveira (IST-UTL)
Rosa Figueiredo (DMAT-UA)
Rui Marques (IST-UTL)
Sérgio Dinis Sousa (EE-UM)
Sofia Rodrigues (IPVC)
Susana Relvas (IST-UTL)
Tânia Pinto Varela (IST-UTL)
Tânia Ramos (ISCTE)
Teresa Almeida (ISEG-UTL)
Teresa Gouveia (FEUP-UP)
Teresa Melo (Saarland Univ. of Applied Sciences)
Vera Miguéis (FEUP-UP)

Com o patrocínio

FCT
Fundação para a Ciência e a Tecnologia
MINISTÉRIO DA EDUCAÇÃO E CIÊNCIA

Com o apoio

ISBN: 978-972-745-154-8