Estudo Paramétrico do Método dos Conjuntos Análogos Aplicado a Dados Meteorológicos

Leonardo Oliveira Guth de Araújo - a46677

Dissertação apresentada à Escola Superior de Tecnologia e de Gestão de Bragança para obtenção do Grau de Mestre em Informática.

Trabalho orientado por:

Prof. Dr. Carlos Balsa

Prof. Dr. José Rufino

Esta dissertação não inclui as críticas e sugestões feitas pelo Júri.

Bragança

2020-2021

Estudo Paramétrico do Método dos Conjuntos Análogos Aplicado a Dados Meteorológicos

Leonardo Oliveira Guth de Araújo - a46677

Dissertação apresentada à Escola Superior de Tecnologia e de Gestão de Bragança para obtenção do Grau de Mestre em Informática.

Trabalho orientado por:

Prof. Dr. Carlos Balsa

Prof. Dr. José Rufino

Esta dissertação não inclui as críticas e sugestões feitas pelo Júri.

Bragança

2020-2021

Dedicatória

Dedico essa dissertação a todos que acreditaram em meu potencial e me auxiliaram de alguma forma nessa trajetória.

Agradecimentos

Primeiramente, fica um agradecimento em especial à minha mãe Regiane e ao meu pai Dorinel, que fizeram com que tudo fosse possível, pelo amor que me deram e pelas oportunidades que me proporcionaram.

Aos meus orientadores no Instituto Politécnico de Bragança, Professor Dr. Carlos Jorge da Rocha Balsa e Professor Dr. José Carlos Rufino Amaro. Ambos me ajudaram imensamente. Obrigado pelos seus ensinamentos e por terem feito parte dessa fase tão importante em minha vida.

Agradeço por ter vindo a Portugal com meus colegas Lucas Mendes e Milton Boos, que ser tornaram amigos tão importantes para mim.

Resumo

No presente trabalho é feito um estudo paramétrico de variações do método de Conjuntos Análogos. O método, que é originalmente utilizado para a realização de pós-processamento, é aplicado neste estudo para realizar a previsão de séries temporais. Séries temporais conhecidas são reconstruídas com o intuito de avaliar e identificar os parâmetros ótimos que alcancem a minimização de erros obtidos da comparação entre as séries temporais reais e as previstas. Ao todo foram utilizados nove anos de dados, sete anos para a fase de treinamento e dois para a comparação com as previsões feitas. São utilizados dados de três estações meteorológicas, onde o método é aplicado de forma que a estação A tenha os seus dados previstos a partir das estações B e C.

O estudo demonstrou que é possível identificar valores e padrões de parâmetros que são úteis para a minimização de erros. No entanto, também mostrou é difícil encontrar parâmetros globais que minimizem os erros de qualquer variável em qualquer cenário trabalhado. O estudo de desempenho realizado demonstrou que, na maioria dos casos, utilizar *clusterização* diminui consideravelmente o tempo gasto pelo processo de previsão, alcançando ainda previsões bastante precisas.

Palavras-chave: Conjuntos Análogos; Parametrização; Séries Temporais; Previsão Meteorológica

Abstract

In the present work a parametric study of variations of the Analogs Ensembles method of is conducted. The method, that is originally used to perform post-processing, is applied in this study to the forecast of time series. Known time series are reconstructed in order to evaluate and identify the optimal parameters that achieve the minimization of errors obtained from the comparison between the real and predicted time series. Altogether, nine years of data were used, seven years for the training period and two years for comparison with the predictions made. Data from three weather stations are used, where the method is applied so that station A has its data predicted from stations B and C.

The study showed that it is possible to identify values and patterns of parameters that are useful for the minimization of errors. However, it also showed that it is difficult to find global parameters that minimize the errors of any variable in any scenario worked on. The performance study carried out showed that, in most cases, using clustering considerably reduces the time spent in the forecasting process and still achieves very accurate forecasts.

Keywords: Analogs Ensemble; Parameterization; Time Series; Weather Forecast

Conteúdo

Lista de Tabelas

Lista de Figuras

Capítulo 1

Introdução

1.1 Enquadramento

Informações sobre estados futuros do tempo são usadas diariamente para os mais diversos propósitos em nossa sociedade, tendo grande impacto em muitos processos de tomadas de decisões (por exemplo, em agricultura, gestão de energias renováveis, planeamento de eventos, defesa, etc.). Os métodos utilizados para os cálculos das previsões do tempo apresentaram uma constante evolução ao longo dos anos, sempre com o propósito de aumentar a precisão e o alcance da previsão, bem como a eficiência computacional.

As previsões feitas hoje em dia aproximam-se cada vez mais da realidade. No entanto, os resultados obtidos estão sempre sujeitos a incertezas ou imprecisões, sejam elas geradas pelo modelo de previsão ou por fatores externos relacionados com os dados utilizados [1]. Dito isso, os modelos de previsão são constantemente estudados e aprimorados.

Uma forma de melhorar a qualidade das previsões meteorológicas são os métodos de pós-processamento, usados para realizar a correção das previsões. De entre esses métodos, destaca-se a técnica dos Conjuntos Análogos, introduzida por Luca Delle Monache (2011), que tem mostrado bons resultados em diversas áreas. Apesar de apresentar limitações na previsão de variações repentinas do valor das variáveis físicas associadas à meteorologia, essas limitações podem ser minimizadas pela escolha de dados de treino adequados. Este

método, e as questões ligadas a esta escolha, são o foco principal desta dissertação.

1.2 Objetivos

Esta dissertação tem como objetivo principal aprimorar resultados de reconstrução de séries temporais meteorológicas, com base no método dos Conjuntos Análogos. Pretende-se alcançar esse objetivo através de um estudo paramétrico, que permita identificar os melhores valores para parâmetros chave do método (que possuem um grande impacto durante o processo de cálculo das previsões), a serem usados na reconstrução de três variáveis físicas inter-relacionadas.

1.3 Contribuições

A partir das análises realizadas foi possível identificar padrões resultantes das variações paramétricas que foram feitas buscando identificar os parâmetros ótimos para os diferentes métodos de previsão utilizados. Foram identificados os melhores valores a serem utilizados para três principais parâmetros dos métodos trabalhados. O primeiro avaliado foi o número de *clusters* (N_c) que deve ser utilizado nas variantes K-means e C-means. Foram identificados valores mais propensos a terem resultados bons e ruins para cada variável meteorológica prevista.

O segundo parâmetro avaliado foi o número de análogos (N_a), tendo-se concluído que para as variantes K-means e C-means, não se deve usar um valor limitador para os mesmo, mas sim todos os análogos pertencentes aos melhores *clusters*. Para a variante Monache o valor ideal de N_a varia de acordo com a variável que está sendo prevista; entretanto, $N_a = 150$ foi definido como sendo o valor que minimiza simultaneamente os erros de todas as variáveis testadas.

O terceiro parâmetro analisado foi o tamanho da janela dos análogos (regulado através da variável k), responsável por definir o intervalo de tempo que um análogo abrange. O valor ideal identificado para as variantes K-means e C-means foi $k = 5$, que é independente

da variável prevista. Já na variante Monache o valor ideal de k varia de acordo com a variável prevista; no entanto, $k = 2$ acabou por ser selecionado como a melhor aproximação ao ideal (dada a necessidade de fixar um valor no estudo de outros parâmetros).

Além destes três principais parâmetros, também foram avaliados pesos de relevância para os análogos nas variantes K-means e C-means, modelos de estações dependentes e independentes, a previsão feita com análogos selecionados a partir de variáveis diferentes da variável em foco e o desempenho dos métodos comparando tempos e erros obtidos.

Para os pesos analisados nas variantes K-means e C-means, concluiu-se que não há diferenças significativas entre usar ou não pesos. Portanto, não há porque utilizá-los.

O modelo de estações independentes demonstrou grande impacto em duas das variáveis analisadas, mas apenas quando utilizada a variante Monache. Os restantes das avaliações demonstrou que o modelo de estações dependentes é, na maioria das vezes, mais preciso.

Fazer a previsão de uma variável levando em consideração outras diferentes a ela no processo de seleção dos análogos, demonstrou ser uma boa opção apenas para uma combinação de variáveis. O restante das combinações resultou em previsões muito imprecisas; portanto, realizar a seleção dos análogos apenas com variáveis diferentes da que está sendo prevista, não é uma boa opção.

O desempenho computacional avaliado entre as diferentes variações do método dos Conjuntos Análogos mostrou que a variante K-means sempre obteve os melhores tempos de execução e em apenas uma ocasião não obteve previsões mais previsas em comparação a variante Monache; sendo assim, é mais recomendado o uso da variante K-means.

1.4 Estrutura do Documento

No Capítulo 1 é feita a introdução do trabalho. É apresentada uma breve revisão histórica, que passa pelos principais eventos da evolução da previsão meteorológica, apresenta o estado da arte onde são abordados os trabalhos mais atuais envolvendo o método dos Conjuntos Análogos, e motiva os objetivos da investigação.

O Capítulo 2 aborda mais profundamente a parte lógica do método Conjuntos Análogos. Apresenta como o método pode ser usado para o Hindcast e mostra como o método funciona para suas variações que utilizam clusterização. Além disso, introduz as medidas de similaridade usadas nos processos contidos dentro do método. Também se explica como são calculadas as previsões e termina-se apresentando as métricas usadas para avaliação das previsões obtidas.

No Capítulo 3 são apresentadas informações sobre os dados utilizados nos processos de previsão. Aborda-se a origem dos dados, como estão armazenados e por quais processos eles passam até que possam ser utilizados para a realização das previsões. Além disso, são descritos os processos presentes no algoritmo de previsão, juntamente com os recursos computacionais utilizados.

As análises paramétricas são todas descritas e analisadas no Capítulo 4. O capítulo inicia abordando o parâmetro número de *clusters*, segue com o parâmetro número de análogos e em seguida aborda o parâmetro ligado à dimensão da janela dos análogos. Após esse parâmetros, também avalia medidas de peso ligadas a relevância das observações identificadas no processo de previsão, avalia os métodos de estações dependentes e independentes, avalia previsões feitas com diferentes combinações de variáveis e termina com a análise do desempenho computacional no contexto dos melhores resultado encontrados para cada variante trabalhada.

O Capítulo 5 apresenta as conclusões referentes a todas as análises feitas durante o trabalho e propõe os trabalhos futuros.

Capítulo 2

Contexto e Metodologias

Neste capítulo introduz-se o domínio científico em que se enquadra esta dissertação, assim como a metodologia de análise de dados que constitui o foco principal da mesma.

2.1 Resenha Histórica

No século XIX, o avanço do conhecimento acerca da termodinâmica e da dinâmica dos fluidos possibilitou a compreensão dos princípios físicos fundamentais que governam o fluxo na atmosfera [2]. No final desse século, o meteorologista Cleveland Abbe escreveu [3] que "a meteorologia é essencialmente a aplicação da hidro-dinâmica e da termodinâmica na atmosfera", tendo proposto aproximações matemáticas para a construção de um modelo de previsão de estados atmosféricos.

Pouco tempo depois (1904), o norueguês Vilhelm Bjerknes propôs [4] um modelo de previsão em duas etapas: diagnóstica e prognóstica. Na fase diagnóstica, através de observações, é determinado o estado inicial da atmosfera. Na fase prognóstica, as mudanças que ocorreram nesse estado ao longo do tempo são calculadas utilizando as leis que regem o estado da atmosfera. A fase prognóstica é feita com base em um conjunto de equações que representam princípios físicos de conservação de energia, massa e momento, além de relações diagnósticas entre pressão, temperatura e densidade. Bjerknes desenvolveu então

um método qualitativo para resolver as equações, já que não conseguiria resolvê-las numericamente pois, na época, apenas máquinas de cálculos manuais estavam disponíveis, o que inviabilizava uma solução analítica por conta da complexidade das equações [2], [5].

Em 1913, o inglês Lewis Fry Richardson propôs a discretização da atmosfera numa malha de latitude/longitude e altitude para a resolução desses princípios físicos e, utilizando o banco de dados mais completo disponível, calculou a mudança da pressão atmosférica e dos ventos em dois pontos da Europa Central. Mas os resultados obtidos não foram bons, estando longe da realidade. Os problemas no seu método foram identificados apenas em 1930, por matemáticos que apontaram apenas questões de equilíbrio entre o campo da pressão e do vento nas condições iniciais do modelo previsor [2].

Nessa época, John von Neumann, um dos grandes matemáticos do século XX, incidiu o seu foco sobre fluxos em fluidos turbulentos. Percebeu que as equações utilizadas nessa área dificultavam a análise dos fluxos, assim como a percepção qualitativa dos mesmos, e afirmou que os processos da hidrodinâmica seriam resolvidos mais rapidamente se um processo numérico estivesse disponível. Mas para que isso fosse viável, máquinas computacionais automáticas e muito mais rápidas seriam necessárias [5].

Assim, em 1943, Neumann iniciou o projeto de construção de um computador eletrónico no Instituto de Estudos Avançados (IAS) de Princeton. Esse computador foi o Electronic Numerical Integrator and Computer (ENIAC), que representou um marco na evolução dos sistemas de computação, estando a sua arquitetura na base dos sistemas desenvolvidos desde então até à atualidade. O ENIAC foi finalizado perto do final da segunda Guerra Mundial e foi utilizado, durante esse período, para os cálculos que possibilitaram a construção das primeiras bombas atómicas. Terminada a guerra, foi usado para outros propósitos, sendo um deles a previsão meteorológica [2].

Durante o período da segunda Guerra Mundial, as previsões meteorológicas tiveram grande importância e visibilidade, sendo essas informações vitais no planeamento de ataques (recorde-se, por exemplo, o caso do dia D). Após a guerra, modelos matemáticos começaram a ser desenvolvidos para a previsão do tempo e em 1948/49 foram feitas as primeira previsões com o ENIAC. No entanto, os processos de cálculo eram demasiado

lentos para terem utilidade prática, dificuldade sentida ainda nos anos seguintes [6].

Já no início dos anos 1950, foi finalizado o desenvolvimento do algoritmo numérico usado para representar o sistema de equações quase-geotrófico, um sistema mais complexo do que o usado por Neumann, baseado no gradiente da pressão atmosférica e na força inercial de Coriolis, e que leva em consideração que ambas as forças estão quase em equilíbrio [2], [7]. O sistema foi implementado no ENIAC e então quatro previsões de 24 horas foram feitas, resultando em resultados satisfatórios, indicando que características em grande escala poderiam ser previstas usando esse modelo barotrópico. No entanto, cada uma das previsões demorava 24 horas para ser concluída, significando que os resultados obtidos continuavam a não ter um valor prático, pois apenas acompanhavam o tempo [2], [5]. Desde então, os sistemas computacionais registaram uma tremenda evolução, ao ponto de hoje em dia serem possíveis previsões a 48 horas com elevada fiabilidade.

Numa linha diferente, em 1969, Edward S. Epstein [8] propõe um modelo estocástico dinâmico para tentar lidar com problemas de observação da atmosfera, pois era difícil considerar um estado inicial do modelo suficientemente preciso. Todavia, esta abordagem necessitava de grande poder computacional. Mais adiante, em 1974, Leith [9] propõe uma técnica de "conjuntos análogos" para resolver a previsão dinâmica estocástica usando a aproximação de probabilidade Monte Carlo. A partir desse ponto a técnica dos "conjuntos análogos" começou a ser mais estudada e viria a mostrar bons resultados.

2.2 Previsão por Analogia

Durante algum tempo, a previsão meteorológica assentou em sistemas de equações que usavam informações sobre o estado da atmosfera num determinado instante e projetavam o seu estado futuro. Entretanto, em 1969, Edward N. Lorenz avançou com uma alternativa [10]: o estado atual da atmosfera já foi provavelmente observado no passado, pelo que os estados seguintes tenderão a ser semelhantes aos observados no passado. Mas na altura, após analisar o resultados dos seus esforços baseados nesse pressuposto, Lorenz descartou essa hipótese, acreditando ser impossível encontrar estados da atmosfera análogos.

Mais tarde, em 1988, outros investigadores, como Kimmo Ruosteenoja [11], chegaram à mesma conclusão de Lorenz: seria impossível utilizar a técnica em grande escala e demoraria muitos anos até que fossem encontrados dois estados similares da atmosfera. No entanto, pouco tempo depois, em 1989, H. M. Van den Dool [12] percebeu que o método poderia ser afinal eficaz, desde que a escala do problema fosse reduzida.

Em 2011, Monache propõe [13], pela primeira vez, duas novas abordagens que utilizam a ideia de análogos em métodos de pós-processamento. Uma delas é a combinação do método de pós-processamento Filtro de Kalman (KF) [14] com análogos (ANKF), e a outra é uma abordagem puramente baseada em observações que verificam quando uma previsão análoga é válida (AN) [13].

O método KF utilizado era um algoritmo sequencial de fácil implementação e computacionalmente leve, que realiza a correção de erros de previsões baseando-se em erros passados. Esse método atribui maiores pesos a erros mais recentes, fazendo com que esses erros tenham maior influência na correção do erro atual. No entanto, esse método não consegue prever alterações repentinas dos erros, fazendo com que correções imprecisas possam ser feitas. Essas mudanças repentinas são causadas por variações bruscas do regime meteorológico. O método era aplicado utilizando erros de uma série temporal contínua, onde nenhum tipo de filtro era aplicado aos dados usados. Isso fazia com que essas mudanças bruscas pudessem ser encontradas. Sendo assim, a ideia de Monache foi substituir essa série temporal apenas por valores análogos, ordenados de acordo com a ocorrência de cada um deles. A mudança feita resultou em melhores correções do erro e o aspecto chave do método passou a ser a determinação dos melhores análogos [13].

Posteriormente, em 2013, Monache [15] cita Lorenz [16] na oposição à ideia de que um modelo de Previsão Numérica do Tempo (NWP) pode fornecer informações úteis à tomada de decisões, por ser um método limitado, pois representa apenas um único estado da atmosfera e provém de estados iniciais imperfeitos com deficiências que levam ao crescimento de erros não lineares durante a construção do modelo. Refere ainda o uso de Funções de Densidade de Probabilidade (PDF) como mais úteis para tomar decisões.

No mesmo trabalho, em que introduz o método dos Conjuntos Análogos (AnEN),

Monache menciona também o facto de que esse método não é útil apenas para pós-processamento, e utiliza uma variante do método para estimar a previsão de uma PDF, onde um conjunto de n observações correspondentes aos n melhores análogos a uma previsão determinística (oriunda de um modelo NWP), são usadas para estimar as previsões.

Esse método usa duas séries temporais com dados meteorológicos: uma chamada de *dados históricos*, que é originada de um modelo numérico de previsão do tempo (NWP), e outra chamada de *dados observados*, que representam um conjunto de dados reais. Com o intuito de melhorar a precisão do modelo NWP num determinado período situado nos dados observados, um período igual é selecionado na série de dados históricos e é então comparado com previsões passadas situadas num chamado *período de treinamento*, buscando encontrar previsões passadas mais similares à atual. Essas previsões passadas formam o *conjunto de análogos*. Após selecionados os análogos, os valores reais correspondentes ao mesmo momento no tempo dos análogos, são usados para melhorar a precisão da nova previsão dos dados observados.

2.3 Aplicações do Método AnEn

O método conhecido atualmente como Conjuntos Análogos (AnEn) é uma extensão do método ANKF de Monache [13], e que inicialmente era usado como um método de pós-processamento para correção de erros de previsões numéricas do tempo (NWP). Em [15] é usado pela primeira vez para realizar previsões e não como um método de correção. Neste estudo foram feitas previsões da velocidade do vento e da temperatura para um período de 0-48h. Foram usados dados no período de Abril até Julho de 2011 de 550 estações espalhadas pelos Estados Unidos. As previsões obtidas do AnEn foram comparados com dois conjuntos de dados NWP obtidos de duas fontes diferentes e, de acordo com Monache, os resultados obtidos do AnEn eram melhores, especialmente para eventos não usuais.

Mais adiante, em 2014, Alessandrini utiliza o método AnEn para realizar previsões da força do vento, introduzindo o método na área de energias renováveis [17]. Em 2015, Alessandrini realiza outro trabalho, aplicando de novo o método AnEn, dessa vez na área

da energia solar, realizando previsões de 0-72h em três parques solares na Itália [18].

Em meados de 2017, Cervone[19] realiza uma pesquisa onde o método AnEn é combinado com redes neurais artificiais para prever a potência gerada por placas fotovoltaicas. Os resultados mostram que a versão combinada dos métodos apresenta um melhor resultado em relação às versões onde cada método é usado individualmente, além de mostrar que a solução proposta é boa para computação em larga escala[19].

Em 2019, Wu usa o método AnEn para prever a irradiação solar de um curto período, do nascer ao por do sol do dia seguinte. Combina observações de satélites, radiossondas e estações da superfície para encontrar os análogos e então gerar as previsões [20].

No mesmo ano, Alessandrini realiza outro estudo, dessa vez buscando aprimorar a precisão de previsões de eventos raros da velocidade do vento [21]. Mais recentemente, em 2020, Shahriari [22] realiza pesquisa onde é apresentada uma implementação do AnEn em larga escala usando uma malha 2D de dados, estuda os fatores que podem influenciar a incerteza da previsão do vento e introduz uma medida para identificar locais de alto risco para o desenvolvimento de parques eólicos.

2.4 Hindcasting

O modelo AnEn original pode ser adaptado para reconstruir dados meteorológicos ausentes ou inexistentes de uma estação meteorológica. Nesse caso, ao invés de usar dados obtidos de um modelo NWP, como dados históricos são utilizados dados meteorológicos reais de estações próximas à estação de que se quer obter os valores. Essa estação próxima será usada como uma estação preditora. Neste cenário, os dados históricos são os dados da estação preditora e os dados observados são dados meteorológicos da estação prevista. Esta metodologia é ilustrada na Figura 2.1, onde é usada apenas uma estação preditora.

O processo de reconstrução é dividido em três etapas. Na primeira delas $\mathbb{D}$, um conjunto de dados análogos ao preditor é selecionado no período de treino dos dados históricos. Os análogos são selecionados com base na similaridade entre eles e o preditor. Cada possível análogo P_a (elemento do período de treino) é composto por um vetor de $2k$

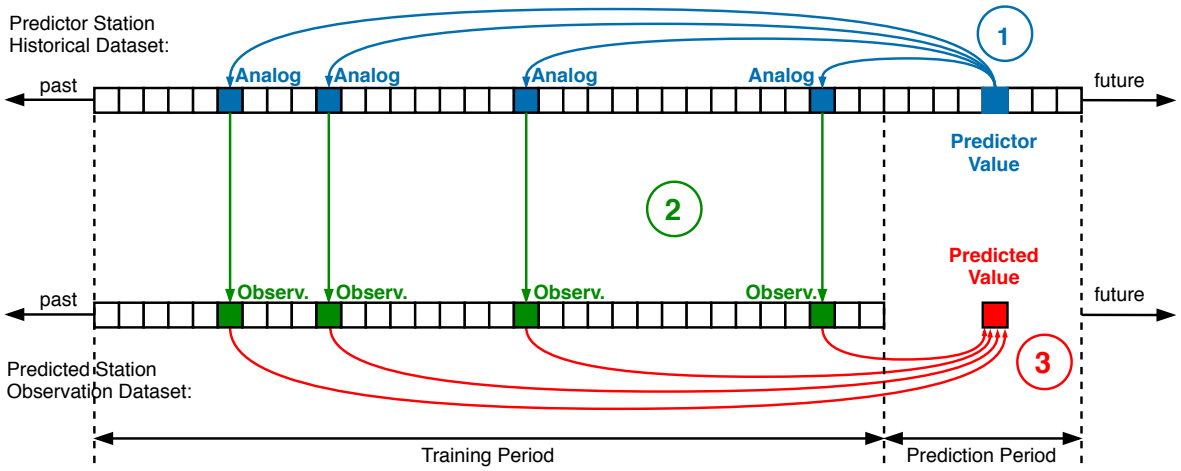


Figura 2.1: Hindcasting com o método dos Conjuntos Análogos (adaptado de [23]).

+ 1 elementos, onde cada elemento representa uma variável meteorológica em um dado instante e k representa metade da janela de tempo coberta pelo análogo.

Na segunda etapa ②, os análogos selecionados são mapeados nas observações correspondentes a eles na estação prevista. Essa correspondência é feita com base no período do tempo em que se encontram. E por fim, na terceira etapa ③, essas observações são usadas para estimar (*hindcast*) o valor ausente no mesmo instante do preditor conhecido.

2.5 Medidas de Similaridade

As medidas de similaridade são usadas para definir quais dos possíveis análogos serão selecionados para escolher as observações a usar nos cálculos das previsões. Originalmente, o método dos conjuntos análogos apresentado por Monache [15] utiliza uma medida de similaridade baseada na distância euclidiana. A fórmula utilizada permite que o cálculo seja feito utilizando múltiplas variáveis (N_v) meteorológicas e procura não apenas por similaridades entre padrões de variação do tempo, mas também por valores numéricos semelhantes nas várias variáveis usadas nas previsões (portanto é feita a adição do desvio padrão σ_{fi} à métrica). Além disso, um peso w_i é atribuído a cada variável usada, permitindo que uma variável possa ter mais relevância nos cálculos do que outra.

A fórmula é apresentada pela Equação 2.1. Nesta, o valor de k representa metade

da janela de tempo utilizada para a seleção dos possível análogos (P_a), e as variáveis F e A representam previsões dado um tempo t, no período de predição e no período de treinamento, respetivamente.

$$\sum_{i=1}^{N_v} \frac{w_i}{\sigma_{fi}} \sqrt{\sum_{j=-k}^{k} (F_{i,t+j} - A_{i,t'+j})^2}. \tag{2.1}$$

Neste trabalho é apenas usada uma variável, o que permite simplificar a equação:

$$\sqrt{\sum_{j=-k}^{k} (F_{i,t+j} - A_{i,t'+j})^2}. \tag{2.2}$$

Além disso, duas estações preditoras são usadas, ao invés de apenas uma. Assim, a Equação 2.2 tem de ser adaptada para considerar valores de ambas as estações:

$$\sqrt{\sum_{j=-k}^{k} [(F^{s1}_{i,t+j} - A^{s1}_{i,t'+j})^2 + (F^{s2}_{i,t+j} - A^{s2}_{i,t'+j})^2]}. \tag{2.3}$$

A Equação 2.3 representa uma abordagem de *estações dependentes* [23], onde os valores de ambas as estações são usados simultaneamente para determinar a similaridade entre os preditores e os P_a e então fazer a escolha dos análogos.

Uma outra abordagem, identificada como *independente*, também é utilizada. Nessa abordagem, os dados das estações preditoras são usados de forma independente, utilizando a Equação 2.2 para avaliar a similaridade entre os preditores e os P_a. Então, os conjuntos de análogos são selecionados separadamente para cada estação preditora. Ao final, após a seleção de todos os análogos de forma separada por estação, são identificadas as observações correspondente a cada análogo selecionado e então uma união das diferentes observações é feita partir de médias aritméticas. Após esse processo, as previsões são calculadas a partir do conjunto final de observações.

Todas as variáveis das Equações 2.1, 2.2 e 2.3 são descritas na Tabela 2.1.

Tabela 2.1: Parâmetros das equações

F_t	Previsão dado um tempo t no período de predição.
$A_{t'}$	Previsão do possível análogo dado um tempo t' no período de treino.
N_v	Número de variáveis meteorológicas consideradas quando comparadas as previsões.
w_i	Peso informado para cada variável i.
σ_{fi}	Desvio padrão da variável i no conjunto de dados histórico.
k	Metade da largura da janela de tempo em sobre a qual a métrica é calculada.
$F_{i,t+j}$	Valor da previsão no tempo $t+j$ na janela de tempo da variável i.
$A_{i,t'+j}$	Valor do análogo no tempo $t'+j$ na janela de tempo da variável i.
$F^{s1}_{i,t+j}$	Valor da previsão no tempo $t+j$ na janela de tempo do preditor da estação 1.
$A^{s1}_{i,t'+j}$	Valor do análogo no tempo $t'+j$ na janela de tempo do preditor da estação 1.
$F^{s2}_{i,t+j}$	Valor da previsão no tempo $t+j$ na janela de tempo do preditor da estação 2.
$A^{s2}_{i,t'+j}$	Valor do análogo no tempo $t'+j$ na janela de tempo do preditor da estação 2.

2.6 Variantes com Clusterização

Realizar a busca pelos análogos pode ser uma tarefa exaustiva (mesmo usando mecanismos de busca paralela), pois deve abranger todos os dados contidos no período de treino dos dados históricos. Uma alternativa para diminuir o esforço necessário é utilizar técnicas de agrupamento (*clusterização*) nos dados. Dessa forma, todos os possíveis análogos (P_a) são previamente agrupados de acordo com o algoritmo de clusterização usado. O processo de clusterização agrupa valores de acordo com a similaridade entre eles, de forma que valores similares fiquem num mesmo *cluster* e haja dissimilaridades entre diferentes *clusters* [24]. Nesse processo, em cada *cluster* é gerado um ponto conhecido como centroide (c). Este ponto pode gerar-se a partir da média de todos os pontos pertencentes ao *cluster* e será usado para determinar a similaridade do *cluster* com o elemento preditor, por comparação deste com o centroide. Realizar essa comparação para todos os *clusters* permitirá identificar o *cluster* mais similar ao preditor. Então, o *cluster* mais similar será utilizado, no todo ou em parte, para determinar as observações que serão utilizadas no cálculo da previsão. Consequentemente, recorrer à clusterização fará com que o número de comparações necessárias para fazer a seleção dos análogos seja reduzida drasticamente.

Este trabalho explora dois métodos de clusterização: K-means e fuzzy C-means.

2.6.1 Clusterização K-means

O método K-means recebe como parâmetros de entrada obrigatórios: os dados que serão agrupados, identificados pela variável x; e a quantidade de *clusters* que serão formados ou os centroides iniciais, representados por uma única variável $centers$. Os dados que serão agrupados são os P_a presentes no(s) período(s) de treino. Já para o número de *clusters* foi utilizado um valor inteiro variado, pois não há valor padrão ou ideal para esse parâmetro.

Existem outros parâmetros que são opcionais, tendo já um valor pré-definido que pode ser alterado de acordo com a necessidade. Um deles, que é utilizado neste trabalho, é o número de iterações máximas ($iter.max$), cujo valor por omissão é 10. Este parâmetro define quantas iteração do algoritmo serão feitas. A cada iteração, as distâncias dos elementos aos centroides são recalculadas e então, se for preciso, os elementos são reagrupados e um novo cálculo para definir o valor dos centroides será feito. Existem duas possibilidades para parar a execução do algoritmo: ou é atingido o número de iterações máximas, ou a variação de um estado de agrupamento passado é tão pequena em relação ao novo estado, que não se justifica continuar com as iterações, pois as próximas iterações não alterarão significativamente os agrupamentos. Neste trabalho, o valor de $iter.max$ foi definido como 40, para garantir a convergência dos agrupamentos para o estado ideal.

Caso não sejam fornecidos os valores dos centroides iniciais, mas sim a quantidade de *clusters* que serão formados, então os primeiros centroides são definidos selecionando aleatoriamente elementos presentes entre os P_a. A partir dai é calculada a distância de cada P_a até aos centroides, utilizando a distância Euclidiana (Equação 2.4). Assim, cada P_a é vinculado ao *cluster* que obteve a menor distância calculada. Após isso, os centroides são recalculados com base na média dos P_a que foram vinculados a eles e então esse processo volta a repetir-se até atingir uma das duas condições de paragem.

$$d = \sqrt{\sum_{i=-k}^{k} (c_i - P_{ai})^2}, \tag{2.4}$$

Como saída da função de clusterização são geradas 9 variáveis, mas apenas duas delas são utilizadas durante o processo de previsão: a variável $clusters$ e a $centers$. A variável

clusters é um vetor de inteiros que vai de 1 a P_a; indica a que *cluster* pertence cada posição do vetor, onde cada posição do vetor representa os P_a fornecidos no parâmetro de entrada x da função. Já a variável *centers* contém uma matriz que indica os valores atribuídos a cada centroide formado; são os valores dessa matriz que serão usados para identificar qual será o melhor *cluster* para selecionar os análogos que serão usado para identificar as observações e então realizar o cálculo das previsões. Uma terceira variável é usada, mas apenas para análises fora do processo de precisão: a variável *size*; trata-se de um vetor de inteiros que indica a quantidade de P_a atribuídos a cada *cluster* formado.

2.6.2 Clusterização C-means

O método C-means é similar ao K-means no sentido de que também gera agrupamentos dos dados de entrada. As suas principais variáveis de entrada são iguais, precisando indispensavelmente do valor de x e do valor de *centers*, que desempenham o mesmo papel que no método K-means. No método C-means também é usada a variável opcional *iter.max* e valor usado também é o mesmo apresentado anteriormente (40). Já as outras variáveis opcionais, que não são alteradas, possuem novas possibilidades, tendo mais configurações disponíveis já que o método C-means é mais complexo e possui mais variáveis.

Uma grande diferença entre os dois métodos está na formação dos agrupamentos e nas variáveis de saída geradas. No método C-means, durante a formação dos *clusters* ocorre um processo adicional para cada P_a informado como dado de entrada: são calculados graus de associação dos P_a aos *clusters*. Diferentemente do método K-means, onde cada P_a está vinculado a apenas um *cluster* o método C-means vincula todos os P_a a todos os *clusters*, com um grau de associação diferente para cada *cluster*. O grau de associação é um valor entre 0 e 1, sendo a soma de todos os graus de um P_a igual a 1. O grau de associação é informado, no final da execução do método, através da variável *membership*; esta é uma matriz com dimensão igual a P_a vezes o número de *clusters*, onde cada coluna representa o grau de associação e as linhas representam cada P_a. Essa variável será considerada durante o processo de cálculo das previsões, sendo usada como um peso baseado no grau

de associação de cada análogo (Equação 2.6). Para mais detalhes sobre os dois métodos de clusterização explorados, consultar [25].

2.7 Métodos de Previsão

Por conta da existência de duas abordagens diferentes – estações *dependentes* vs estações *independentes*–, foi necessário adaptar o cálculo das previsões para ambos os casos. Ao todo quatro fórmulas são usadas, duas para cada abordagem.

2.7.1 Estações Dependentes

Com a abordagem de estações dependentes, a primeira possibilidade para o cálculo da previsão assenta numa média aritmética simples das observações. Após a seleção dos N_a análogos é possível selecionar as O_{t_i} observações e determinar o valor da previsão P_t via:

$$P_t = \frac{1}{N_a} \sum_{i=1}^{N_a} O_{t_i} \tag{2.5}$$

Alternativamente, quando se usam métodos de clusterização, é possível usar outra fórmula (Equação 2.6), que inclui um peso $w(O_{t_i})$. Este peso pode ser calculado de duas formas. Na primeira, o peso é calculado a partir dos valores obtidos das distâncias Euclidianas dos análogos ao centroide do *cluster* a que pertencem; esses valores são normalizados entre 0 e 1 e então usado como pesos para cada análogo usado. Na segunda forma, que serve apenas para o método C-means, o peso é baseado no grau de associação de cada análogo selecionado do melhor *cluster* e, novamente, o valor é normalizado.

$$P_t = \sum_{i=1}^{N_a} [w(O_{t_i}).O_{t_i}] \tag{2.6}$$

2.7.2 Estações Independentes

Na abordagem de estações independentes os cálculos são similares. No entanto, por conta das observações serem selecionadas separadamente em cada estação usada como preditora,

no final do processo é preciso juntar as observações de todas as estações usadas. A Equação 2.5 transforma-se assim na Equação 2.7, sendo N_s a quantidade total de estações usadas.

$$P_t = \frac{1}{N_a N_s} \sum_{i=1}^{N_s} \sum_{j=1}^{N_a} O^i_{t_j} \tag{2.7}$$

Com apenas duas estações, a Equação 2.7 pode ser representada, em alternativa, como:

$$P_t = \frac{1}{2N_a} \left(\sum_{i=1}^{N_a} {O^1}_{t_i} + \sum_{j=1}^{N_a} {O^2}_{t_j} \right) \tag{2.8}$$

Nestas fórmulas, todas as observações de todas as estações estão sendo usadas. Portanto, uma média aritmética das observações é feita para chegar a previsão final. Novamente, é possível utilizar uma versão alternativa do cálculo da previsão onde são levados em conta pesos $w(O_{t_i})$. Assim sendo, a Equação 2.8 seria reescrita na Equação 2.9:

$$P_t = \frac{1}{2N_a} \left(\sum_{i=1}^{N_a} [w({O^1}_{t_i}).{O^1}_{t_i}] + \sum_{j=1}^{N_a} [w({O^2}_{t_j}).{O^2}_{t_j}] \right) \tag{2.9}$$

2.8 Validação e Medições dos Erros

Uma vez obtidas as previsões, a sua qualidade pode ser aferida com base em diversas métricas ou medidas de erro, que traduzem o afastamento dos valores das previsões face aos valores reais. Neste trabalho são consideradas quatro medidas para esse efeito. De facto, Chai e Draxler [26] realizaram uma análise onde mostram que, utilizar medidas de erros em conjunto, possibilita uma avaliação mais robusta da precisão de um modelo.

Uma das medidas de erro usadas é o Bias. A sua expressão é dada pela Equação 2.10 onde x_i representa os valores da previsão e y_i representa valores os reais:

$$\text{Bias} = \frac{1}{n} \sum_{i=1}^{n} (x_i - y_i), \tag{2.10}$$

O Bias (Tendência) mede o erro médio dos valores obtidos em comparação com os valores reais. Trata-se de uma medida de erro muito útil, pois com ela é possível avaliar

se as previsões geradas estão abaixo ou acima da verdade. No entanto, por si só não é suficiente para avaliar a precisão das previsões, pois não mostra o comportamento dos erros, mostrando apenas o erro sistemático das previsões [27].

Outra medida de erro é a Raiz Quadrada do Erro-Médio (RMSE - Root Mean Square Error). Na literatura, esta medida vem sendo usado como uma métrica estatística padrão para medir o desempenho de modelos meteorológicos em pesquisas climáticas [26]. Esta métrica penaliza a variância, o que significa que valores fora do padrão possuem um grande peso no valor final da métrica. A Equação 2.11 fornece a expressão para o seu cálculo:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - x_i)^2}. \tag{2.11}$$

As métricas Bias e RMSE podem usar-se em conjunto na definição de uma outra métrica: o Desvio Padrão dos Erros (SDE), dado pela Equação 2.12. Nesta, o RMSE representa o nível de aleatoriedade do erro e o Bias representa o erro sistemático [28].

$$\mathrm{SDE} = \sqrt{\mathrm{RMSE}^2 - \mathrm{Bias}^2}. \tag{2.12}$$

A última métrica utilizada neste trabalho é o Erro Médio Absoluto (MAE), que representa a distância média absoluta em relação à verdade. Diferentemente do RMSE, o MAE atribui o mesmo peso a todos os valores, impedindo que o seu valor final seja muito distorcido por valores fora do padrão. Um valor alto para o MAE significa que a previsão obtida está longe da verdade. Por definição o seu valor é sempre menor do que o RMSE.

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - x_i|. \tag{2.13}$$

Embora com uma fórmula semelhante ao Bias, o MAE fornece uma medida que representa mais uma aproximação da previsão com a verdade, e não permite que valores positivos e negativos se anulem. Uma análise que é costume fazer combinado o Bias e o MAE é a seguinte: caso o valor obtido para o Bias for baixo e o MAE for alto, então a previsão não é precisa (ora o valor obtido está acima da verdade, ora está abaixo.

Capítulo 3

Dados e Processos Computacionais

Neste capítulo são apresentados os dados que foram utilizados nesta dissertação, incluindo a sua origem, variáveis meteorológicas contempladas, resolução e cobertura temporal, e ainda a metodologia usada para realizar a sua formatação tendo em vista as análises paramétricas posteriormente realizadas. A segunda parte do capítulo incide sobre as questões computacionais, nomeadamente as ligadas aos algoritmos e códigos desenvolvidos, à paralelização da exploração paramétrica e às plataformas computacionais usadas.

3.1 Dados Meteorológicos

Os dados meteorológicos utilizados nesta dissertação foram obtidos do National Data Buoy Center (NDBC), nos Estados Unidos, que, através da National Oceanic and Atmospheric Administration (NOAA), faz a administração das três estações consideradas neste estudo. As três estações encontram-se localizadas na costa da Virgínia (Figura 3.1). Esta seleção foi feita com base no estudo anterior [23], no qual as estações *Dom* (36°57'44"N 76°25'27"W), *Ykr* (37°15'5"N 76°20'33"W) e *Ykt* (37°13'36"N 76°28'43"W) obtiveram os melhores resultados (com a utilização de uma abordagem de estações dependentes durante o cálculo das previsões). Para esta dissertação foram selecionados dados de 2011 a 2019.

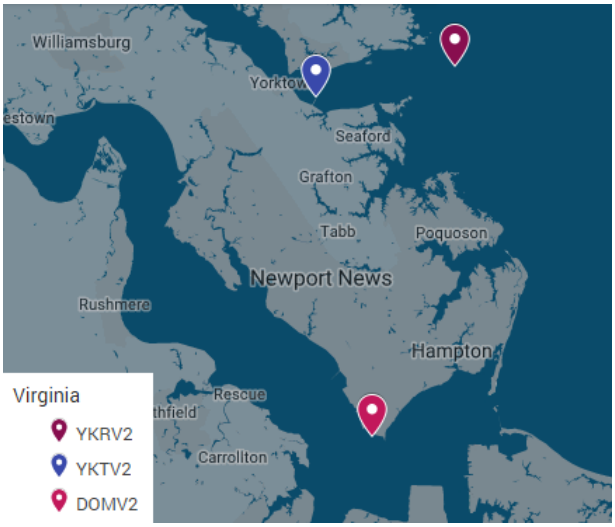


Figura 3.1: Geolocalização das estações seleccionados (Virgínia, Estados Unidos) [29]

3.1.1 Organização dos Dados

Originalmente, os dados são armazenados em arquivos de texto e os valores de cada uma das 18 variáveis disponíveis estão distribuídos em colunas (Tabelas 3.1 e 3.2). Das 18 variáveis apresentadas, 5 são utilizadas para identificação temporal, fornecendo ano, mês, dia, hora e minutos. Das 13 restantes, apenas 6 possuem efetivamente um registro constante de dados: a direção do vento (WDIR), a velocidade do vento (WSPD), a velocidade da rajada de vento (GST), a pressão (PRES), a temperatura do ar (ATMP) e, apenas para as estações *Ykr* e *Ykt*, a temperatura da superfície do mar (WTMP).

Tabela 3.1: Dados meteorológicos originais - estrutura e exemplo (parte 1/2)

YY	MM	DD	hh	mm	WDIR	WSPD	GST	WVHT	DPD
yr	mo	dy	hr	mn	degT	m/s	m/s	m	sec
2011	01	01	00	00	211	0.7	1.0	99.0	99.0

Para que fosse possível uma fácil manipulação dos dados com o RStudio (utilizado para implementar o método AnEn), os dados tiveram que passar por um processo de filtragem e conversão, processo esse que foi realizado utilizando uma *script* construída em Python. Na *script* em questão, os valores omissos, que são originalmente representados por 99 e

Tabela 3.2: Dados meteorológicos originais - estrutura e exemplo (parte 2/2)

APD	MWD	PRES	ATMP	WTMP	DEWP	VIS	TIDE
sec	degT	hPa	degC	degC	degC	mi	ft
99.0	999.0	1023.5	8.8	2.4	999.0	99.0	99.0

999, passaram a ser representados por NA (iniciais de Not Available). Além disso, todos os dados foram armazenados em listas e vetores, para então serem gravados em arquivos no formato NetCDF4 [30]. Todo o processo de conversão é detalhado na seção 3.2.1.

3.1.2 Disponibilidade dos Dados

No período de 2011 a 2019, as três estações possuem todas dados disponíveis. No entanto, há algumas janelas de tempo onde não há nenhum registo. A Tabela 3.3 mostra algumas informações sobre as variáveis usadas para cada uma das estações, como: os menores e maiores valores registados ao longo dos 9 anos, a média de todos os valores, a quantidade exata de valores indisponíveis (#NA) e a percentagem de valores disponíveis.

3.2 Processos Computacionais

Os processos computacionais relevantes nesta dissertação incluem duas grandes etapas de processamento. Na primeira, faz-se a conversão dos dados usados para um formato mais eficiente; esta conversão é feita com uma *script* em Python, que converte os dados do formato original em texto para o formato netCDF4. Na segunda etapa, assente em código R, fazem-se os cálculos das previsões e dos seus erros; os dados, anteriormente convertidos, são carregados para então serem feitos os cálculos e gerados os resultados; esta etapa termina pela exportação dos resultados obtidos na forma de gráficos e tabelas. Foi ainda desenvolvida uma Shell *script* para automatizar a execução da segunda etapa com diferentes parâmetros. As secções seguintes fornecem detalhes sobre estes processos.

Tabela 3.3: Informações diversas sobre as variáveis utilizadas

Yktv2					
Variável	**Min**	**Med**	**Max**	**#NA**	**Disponibilidade (%)**
GST	0.00	5.44	32.80	25156	**96.81%**
WSPD	0.00	4.26	23.80	25127	**96.81%**
ATMP	-13.50	16.06	37.80	26102	**96.69%**
PRES	974.70	1017.34	1044.30	20157	**97.44%**

Domv2					
Variável	**Min**	**Med**	**Max**	**#NA**	**Disponibilidade (%)**
GST	0.00	5.28	32.10	25897	**96.72%**
WSPD	0.00	3.91	24.30	25866	**96.72%**
ATMP	-12.60	16.13	37.20	25194	**96.81%**
PRES	972.80	1017.77	1044.50	19992	**97.47%**

Ykrv2					
Variável	**Min**	**Med**	**Max**	**#NA**	**Disponibilidade (%)**
GST	0.00	6.88	39.60	24910	**96.84%**
WSPD	0.00	5.93	27.60	24795	**96.86%**
ATMP	-12.80	15.86	36.30	22566	**97.14%**
PRES	972.60	1017.35	1043.90	18785	**97.62%**

3.2.1 Conversão dos Dados

O algoritmo para a conversão dos arquivos com dados meteorológicos, baseados na estrutura das Tabelas 3.1 e 3.2, para o formato netCDF4, foi concebido e implementado em Python em trabalhos de investigação anteriores [31] [27]. Em termos computacionais, o processo consiste em três etapas, representadas na Figura 3.2.

O algoritmo inicia realizando a leitura dos arquivos que contêm os dados, efetua a sua conversão e finaliza gravando os dados em novos ficheiros no formato netCDF4.

O objetivo principal deste processo de conversão é possibilitar, durante a execução posterior do algoritmo AnEn, uma leitura rápida dos dados a processar. O processo aplica ainda uma representação alternativa à original para valores omissos (rever secção 3.1.1) e, ao adotar um formato padrão portável (netCFD4), previne possíveis erros que possam ocorrer na leitura dos dados em diferentes linguagens e sistemas [31].

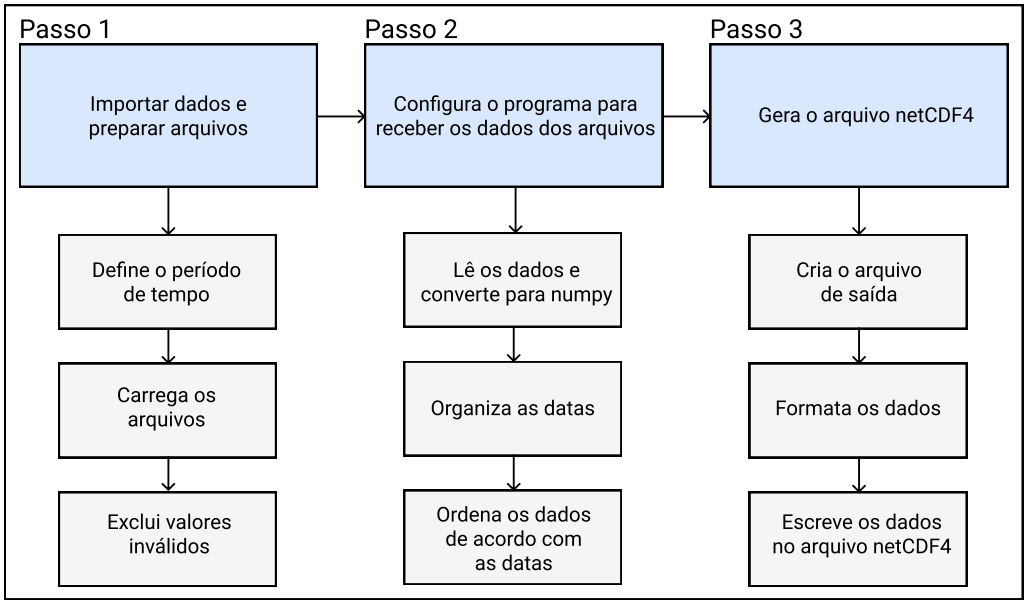


Figura 3.2: Etapas da conversão de dados meteorológicos para o formato netCDF4 (adaptado de [31]).

3.2.2 Cálculos das Previsões e Erros

O código usado no processo de previsão foi escrito em R e tem origem em investigação anterior [27], tendo sofrido as adaptações e modificações necessárias à nova linha de investigação seguida neste trabalho. As principais mudanças feitas no código estão relacionadas com a importação dos dados contidos nos arquivos netCDF4 e com os métodos dependentes e independentes usados durante as previsões. O diagrama na Figura 3.3 representa as principais etapas, e seu encadeamento, do processo realizado pelo código principal de previsão. Estas etapas são descritas com mais detalhe em seguida.

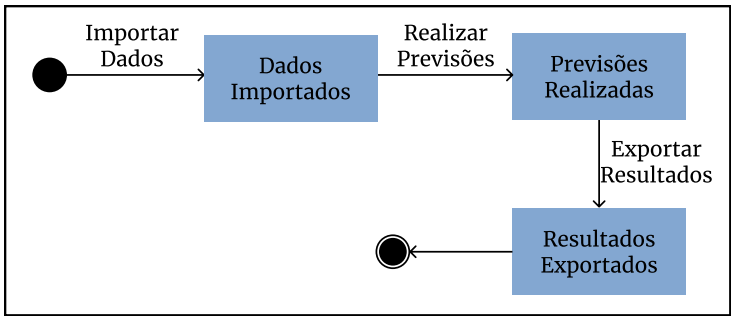


Figura 3.3: Visão geral das principais etapas do código de previsão em R.

Importação dos Dados A fase de importação dos dados não só os carrega na aplicação como também os filtra e organiza de acordo com variáveis inicializadas ainda antes do carregamento. Esse processo possui um elevado custo computacional, pois realiza transformações em todos os dados até estarem efetivamente prontos para a fase de previsão.

Como foi realizada uma quantidade elevada de testes, a importação repetiu-se muitas vezes e, em muitas situações, uma mesma entrada de variáveis era usada. Dessa forma, um mesmo processo de carregamento e formatação repetia-se diversas vezes. Para aumentar a eficiência do processo, optou-se por gravar num ficheiro todos os dados formatados para uma determinada entrada de variáveis. Assim, todas as execuções posteriores à primeira, que usassem uma mesma entrada de variáveis, não necessitariam de realizar o processo de formatação novamente, agilizando-se dessa forma a execução do algoritmo.

O processo de importação de dados está representado no diagrama da Figura 3.4, que mostra todo os passos realizados antes do início propriamente dito das previsões.

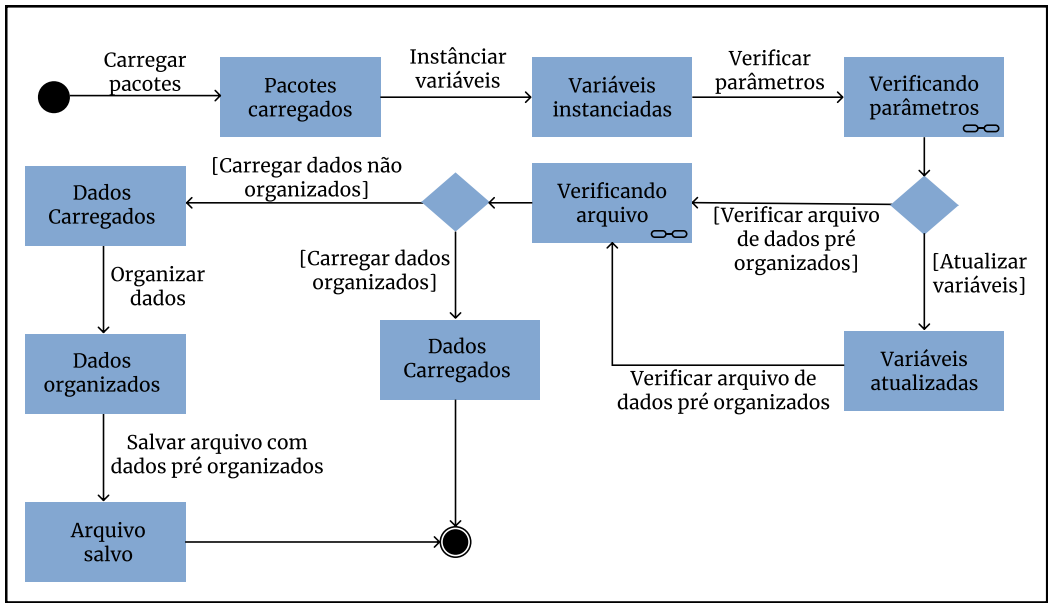


Figura 3.4: Processo de carregamento dos dados

O carregamento de pacotes representa o carregamento de todas as bibliotecas necessárias para a execução do algoritmo. A instanciação de variáveis cria variáveis iniciais com valores padrão que poderão ser alterados ou não pela fase de verificação de parâmetros.

Esses parâmetros podem ser fornecidos externamente ao código, facilitando a criação de *scripts* que executem uma grande quantidade de testes. A verificação de arquivo é feita selecionando os principais parâmetros e verificando se há arquivos pré-organizados com os valores desses parâmetros. Se o arquivo existir é feito o carregamento dos dados contidos no arquivo e então está tudo pronto para a fase de previsão. Caso o contrário, se o arquivo não existir, é invocada a fase de carregamento e organização dos dados, onde arquivos que contém os dados necessários mas que estão organizados de forma mais bruta, são carregados na aplicação e então filtrados. Antes de ir para a fase de previsão, os dados já filtrados são armazenados em novos ficheiros para que execuções futuras não precisem passar novamente por esse processo.

Previsões Ao chegar o momento de realizar as previsões, algumas verificações devem ser feitas para identificar exatamente o método que será utilizado. Ao todo são usáveis três métodos diferentes: o método de Monache, o método K-means e o método C-means. Cada método possui as suas variações, mas uma em comum a todos é que podem usar as variantes de estações dependentes e independentes, que determinarão como serão utilizados os dados das estações preditoras na previsão. Os métodos K-means e C-means podem usar tipos diferentes de pesos adicionais para os análogos encontrados: um deles está ligado à distância do análogo ao centroide do melhor *cluster* e outro relacionado com o grau de associação dos análogos aos *clusters*; de notar que o peso obtido a partir do grau de associação dos análogos só pode ser usado no método C-means, já que é o único método de clusterização usado que gera esse grau de associação. O fluxo da aplicação nessa etapa (etapa de previsão) está representada no diagrama da Figura 3.5.

Primeiramente é verificado qual o método de previsão configurado, após o que é identificado qual o tipo de abordagem de estações (dependentes vs dependentes) a usar, seguindo-se a verificação do tipo de pesos que serão atribuídos aos análogos. Após isso, a previsão ocorrerá, de acordo com a forma como estas propriedades foram configuradas.

Todo o processo até este ponto é feito de forma sequencial mas a partir daqui a previsão é feita de forma paralela, usando a função *parSapply* presente no R, essencial

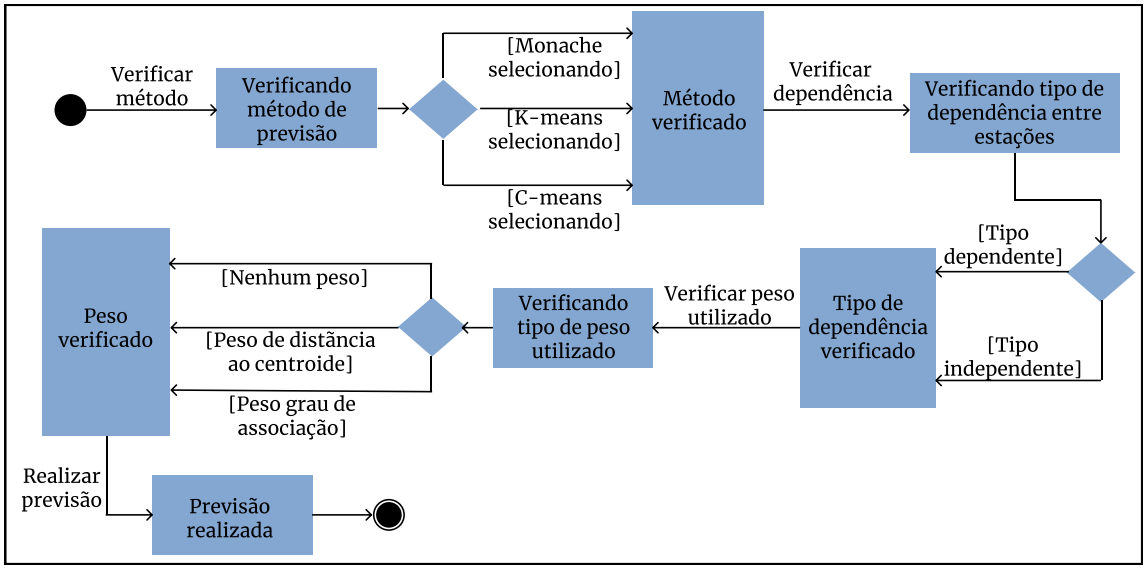


Figura 3.5: Fluxo da aplicação durante a etapa da previsão.

para o desempenho da aplicação. A função *parSapply* explora o paralelismo dividindo um conjunto de dados em fatias e aplica a cada fatia uma mesma função de processamento. No cenário trabalhado as fatias representam vetores de dados que são os preditores e a função aplicada ao preditor é o processo de previsão.

Embora específico do método de Monache, é possível ter uma ideia geral de como todos os processos de previsão funcionam, com base no diagrama da Figura 3.6.

Assim, *A(t[j])* e *B(t[j])* representam as entradas dos dados históricos e dados observados, respectivamente. Um ciclo varre todo o período de treino de *A(t[j])*. Em cada iteração, é criado um vetor *bj* para receber uma janela de dados de tamanho *(2 * k) + 1*, e é calculada a métrica de similaridade *C(bj,bi)* entre o vetor *bj* e o vetor preditor *bi*. Após o ciclo concluir, são selecionados os melhores índices (*ibest*) e então, os índices equivalentes em *B(t[j])* são selecionados para calcular a média de seus valores, *mean(A[ibest])*, e retornar a previsão para cada ponto, *pA(t[j])*.

Para além desta descrição genérica, cada método possui particularidades. Por exemplo, os métodos que usam as estações de forma independente selecionam os melhores índices para cada estação separadamente, tendo ao final de cada iteração do ciclo duas previsões, que são unidas através da média dos seus valores para retornar apenas uma previsão.

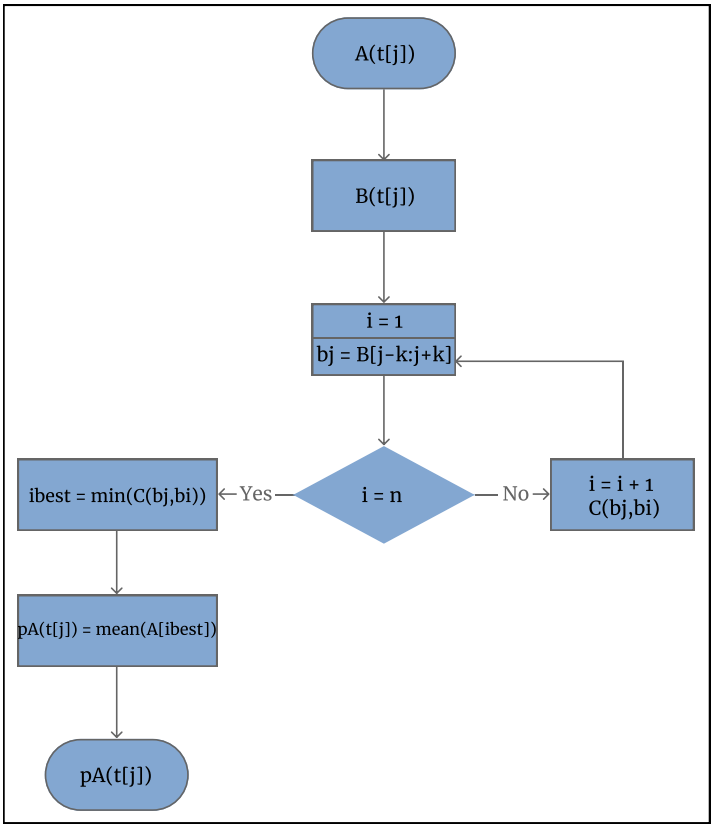


Figura 3.6: Diagrama da previsão com o método de Monache (adaptado de [31])

Adicionalmente, cada tipo de peso usado possui uma forma única de ser calculado e a sua aplicação é feita aos valores obtidos dos melhores índices da estação *B(t[j])* um passo antes da previsão ser calculada.

Exportação de Resultados Na fase de exportação dos resultados, os valores obtidos das previsões, juntamente com as principais configurações, são armazenados em arquivos Rdata. Após as várias execuções do algoritmo, há uma grande quantidade de arquivos Rdata, usados para gerar folhas de cálculo, gráficos e tabelas.

3.2.3 Exploração Paramétrica

Neste trabalho pretende-se melhorar previsões anteriormente obtidas com base no método AnEn (e variantes com clusterização), procurando identificar as melhores combinações de

valores de parâmetros chave desse método. Para automatizar o estudo combinatório dos vários parâmetros considerados, foi desenvolvida uma Shell *script*, que invoca o código R descrito na secção anterior, com diferentes combinações dos parâmetros em estudo.

A *script* consiste basicamente em vários ciclos aninhados, onde cada ciclo faz variar os valores de uma variável de entrada para o código de previsão. Na *script* foi adotado um mecanismo simples de paralelização, que permite lançar vários processos de previsão em simultâneo, até ao máximo correspondente ao número de núcleos de CPU disponíveis.

3.2.4 Recursos Computacionais

O sistema computacional usado para sustentar esta dissertação foi uma máquina virtual alojada no *cluster* de virtualização do CeDRI[1], correndo Linux Ubuntu 18.04.5 LTS. Os recursos associados à máquina virtual, relevantes para esta dissertação, foram: 64 vCPUs (fornecidos por 2 CPUs físicos AMD EPYC 7452 de 32 núcleos cada, com frequência de 2.35/2.65GHz), 128 GB de vRAM (fornecidos a partir de 256GB de RAM DDR4-3200), disco virtual de 512 GB. Esta características permitiram executar todas as previsões sem quaisquer estrangulamentos; em particular, durante o estudo paramétrico, permitiram a execução simultânea de múltiplas previsões com diferentes parâmetros; e, na fase final do trabalho, permitiram também avaliar a escalabilidade dos métodos de previsão num contexto de execução paralela, em função de diferentes níveis de consumo dos recursos.

[1] `https://cedri.ipb.pt/about/cluster`

Capítulo 4

Estudo Paramétrico

O algoritmo dos Conjuntos Análogos contém um conjunto de parâmetros que é necessário calibrar em função do problema. O valor a atribuir a cada uma destas variáveis é desconhecido à partida. Uma forma de contornar esta incerteza consiste na avaliação do efeito da variação dos seus valores na qualidade da previsão.

O estudo paramétrico realizado neste capítulo procura contribuir para se perceber o efeito de variações em três das principais variáveis do método dos Conjuntos Análogos: i) o tamanho da janela dos análogos (W_s), ii) a quantidade de análogos usados (N_a) e, para as variantes do método com clusterização, iii) o número de *clusters* formados (N_c). Adicionalmente, são testados os efeitos dos pesos das observações na qualidade da previsão através da fórmula 2.6, nas seguintes situações: a) um peso igual para todas as observações (utilizado em todos os métodos), b) um peso baseado na distância do análogo ao centro do *cluster* (utilizado apenas nos métodos com clusterização) e c) um peso baseado no grau de pertença dos análogos ao *cluster* (apenas para o método C-means).

As primeiras análises paramétricas são feitas no contexto da previsão de uma só variável atmosférica como, por exemplo, a pressão atmosférica (ATMP). A previsão utiliza apenas dados dessa variável disponíveis nas estações preditoras. Posteriormente, após se identificar os melhores parâmetros para cada variável, é feita uma análise onde variáveis diferentes da variável em foco são utilizadas durante o processo de seleção dos análogos.

Todas as análises foram feitas levando em consideração as estações *Dom* e *Ykr* como

sendo estações preditoras e considerando a estação *Ykt* como sendo a estação prevista.

A seleção dos melhores resultados é feita a partir da identificação dos testes que minimizam mais erros em simultâneo, considerando que a minimização dos erros MAE e RMSE é mais relevante (dado estarem diretamente ligados à magnitude dos erros), seguindo-se a minimização do erro SDE e por último do BIAS.

4.1 Número de Clusters

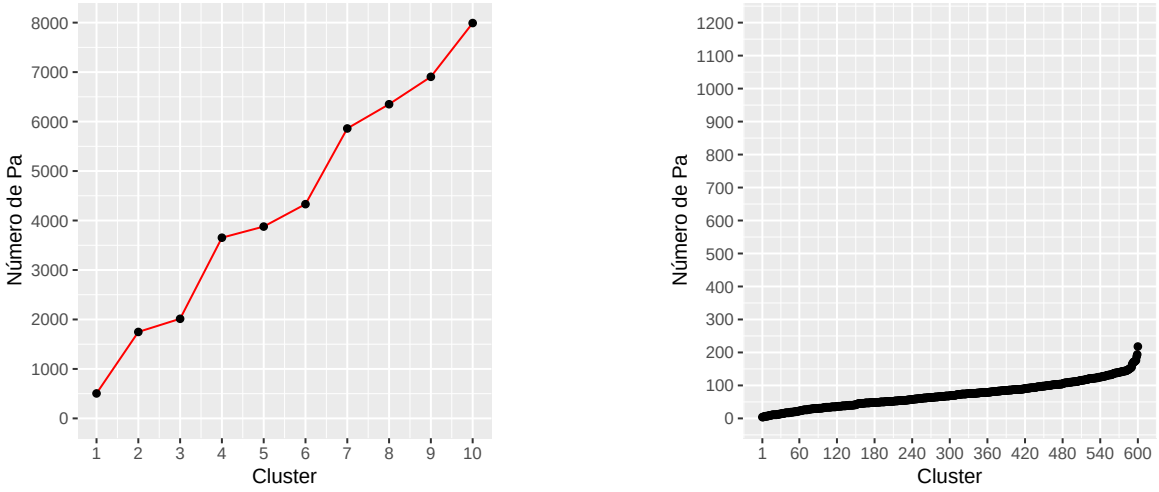
O número de *clusters* (N_c) é uma variável que aparece apenas nas variantes do método dos Conjuntos Análogos em que a clusterização K-means ou C-means é utilizada para identificar os possíveis análogos (P_a) – valores encontrados no período de treino dos dados históricos e que depois serão mapeados em observações que, por sua vez, serão usadas para fazer a previsão/reconstrução. Os agrupamentos resultantes são designados por *clusters*. Os *clusters* são formados levando em consideração o valor da distância euclidiana de cada P_a ao centroide do *cluster*. Os P_a são atribuídos ao *cluster* cujo centroide está mais próximo. Donde, os P_a similares entre eles (e o centroide) agrupam-se no mesmo *cluster*.

Teoricamente, N_c pode variar entre 1 e o número total de possíveis análogos (P_{a_t}) (total de elementos do período de treino). É também de esperar que quanto menor for N_c, maior será o número de P_a por *cluster*, e vice-versa. Mas a distribuição do número de P_a por *cluster* não é necessariamente homogénea, podendo co-existir *clusters* de dimensões variadas. O número de P_a em cada *cluster* varia ainda em função do tamanho do período de treino e é influenciado pelo valor de k (que determina o tamanho da janela do análogo).

4.1.1 Distribuição do Número de Possíveis Análogos por *Cluster*

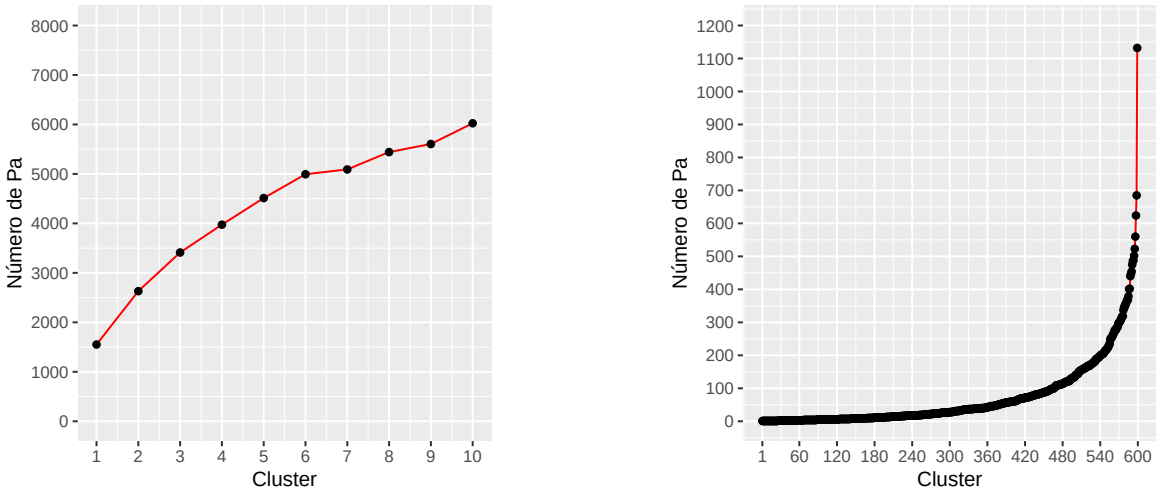
As Figuras 4.1 e 4.2 mostram a distribuição dos P_a por *cluster* (ou seja, a distribuição das cardinalidades possíveis dos *clusters*, vistos como conjuntos de possíveis análogos), para os métodos de clusterização K-means e C-means, quando $N_c = 10$ e $N_c = 600$, para um total de possíveis análogos $P_{a_t} = 43238$. Nos gráficos, cada abcissa identifica um *cluster* através de um número de 1 a N_c, e a soma dos valores de P_a dos vários *clusters* perfaz P_{a_t}.

Os *clusters* estão dispostos de forma a que os valores correspondentes de P_a apareçam por ordem crescente, disposição que facilita a compreensão da distribuição desses valores.



(a) $N_c = 10$ (b) $N_c = 600$

Figura 4.1: Distribuição dos possíveis análogos por *cluster* com K-means.



(a) $N_c = 10$ (b) $N_c = 600$

Figura 4.2: Distribuição dos possíveis análogos por *cluster* com C-means.

Analisando as Figuras 4.1 e 4.2, é possível perceber que há diferenças relevantes na quantidade de P_a afetos aos diferentes *clusters* formados pelos dois métodos.

Assim, para um número reduzido de *clusters* ($N_c = 10$), já se sabe que os *clusters* tenderão a ser maiores. No entanto, os valores de P_a com K-means são mais dispersos e cobrem uma maior amplitude (variando entre 505 e 7992 - Figura 4.1a) do que com o método C-means (onde variam entre 1552 e 6024 - Figura 4.2a). Em ambos os casos, sendo o valor de N_c pequeno, se as previsões fossem calculadas com base em todos os P_a de um *cluster*, os seus resultados provavelmente não estariam muito próximos da verdade pois, muito possivelmente, haveria uma grande dispersão dos análogos incluídos no melhor *cluster*. Isso faria com que as previsões feitas a partir desses análogos tivessem maiores erros, ficando dessa forma longe da verdade. Isto mesmo pode ser observado nos gráficos da secção 4.1.2 , onde os erros são tanto maiores quanto menor for o valor de N_c.

Por outro lado, para um número substancial de *clusters* ($N_c = 600$), estes tenderão a ser menores (i.e., a terem menos possíveis análogos); todavia, neste caso, é no método C-means que os valores de P_a se espalham por um intervalo maior (entre 1 e 1132 - Figura 4.2b), ao passo que no método K-means a gama de variação é menor (entre 4 e 218 - Figura 4.1b); contudo, também é possível observar que no método C-means cerca de 90% ($\approx$1-(600-540)/600) dos valores de P_a se enquadram na mesma gama de variação do método K-means ($0<P_a \leq 200$), sendo uma minoria (10%) os valores de P_a que sobem acima do limite superior desse intervalo ($P_a > 200$) e já com alguma dispersão; adicionalmente, a comparação das figuras 4.1b e 4.2b parece também indicar haver uma maior quantidade de *clusters* pequenos gerados com o método C-means, do que com o método K-means.

4.1.2 Erros para Diferentes Valores do Número de *Clusters*

Nesta secção mostram-se os erros obtidos em previsões com diferentes valores de N_c, utilizando todos os possíveis análogos (P_a) presentes no *cluster* selecionado para a previsão, e assumindo $k = 5$ (o que corresponde ao tamanho $W_s = 11$ da janela dos análogos). Para cada variável meteorológica (GST, ATMP, WPSD e PRES) mostram-se os erros BIAS, MAE, RMSE e SDE obtidos pela aplicação do método K-means, e as diferenças entre esses erros e os obtidos com o método C-means (os erros gerados por ambos os métodos

são muito semelhantes, recorrendo-se a gráficos com a escala apropriada de forma a tornar perceptíveis as diferenças entre os erros). Em cada gráfico, o eixo horizontal corresponde ao valor de N_c usado em cada teste e o eixo vertical apresenta o valor do erro obtido na previsão, ou a diferença dos erros entre o método K-means e o método C-means.

Variável ATMP

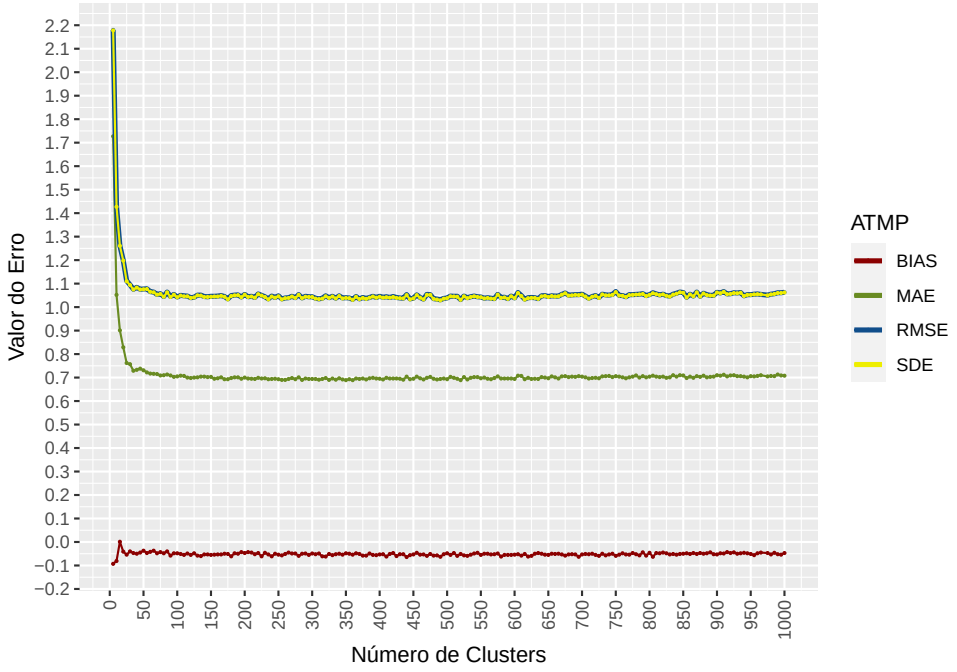


Figura 4.3: Erros das previsões da variável ATMP em função de N_c - método K-means

A Figura 4.3 mostra a distribuição dos erros das previsões da variável ATMP com o método K-means. É possível notar um comportamento similar na variação dos erros: com exceção do BIAS (que se mantém relativamente constante), verifica-se uma diminuição dos erros à medida que N_c aumenta de 1 até 100 (sendo essa quebra mais pronunciada no intervalo $1 \leq N_c \leq 25$); a partir de $N_c = 100$ os valores estabilizam, exibindo oscilações muito ligeiras. Pode-se também observar que os erros RMSE e SDE são muito próximos entre si. Os dois cenários que minimizaram mais erros foram $N_c = 520$ e $N_c = 490$.

A distribuição dos erros gerados pela aplicação do método C-means é muito semelhante

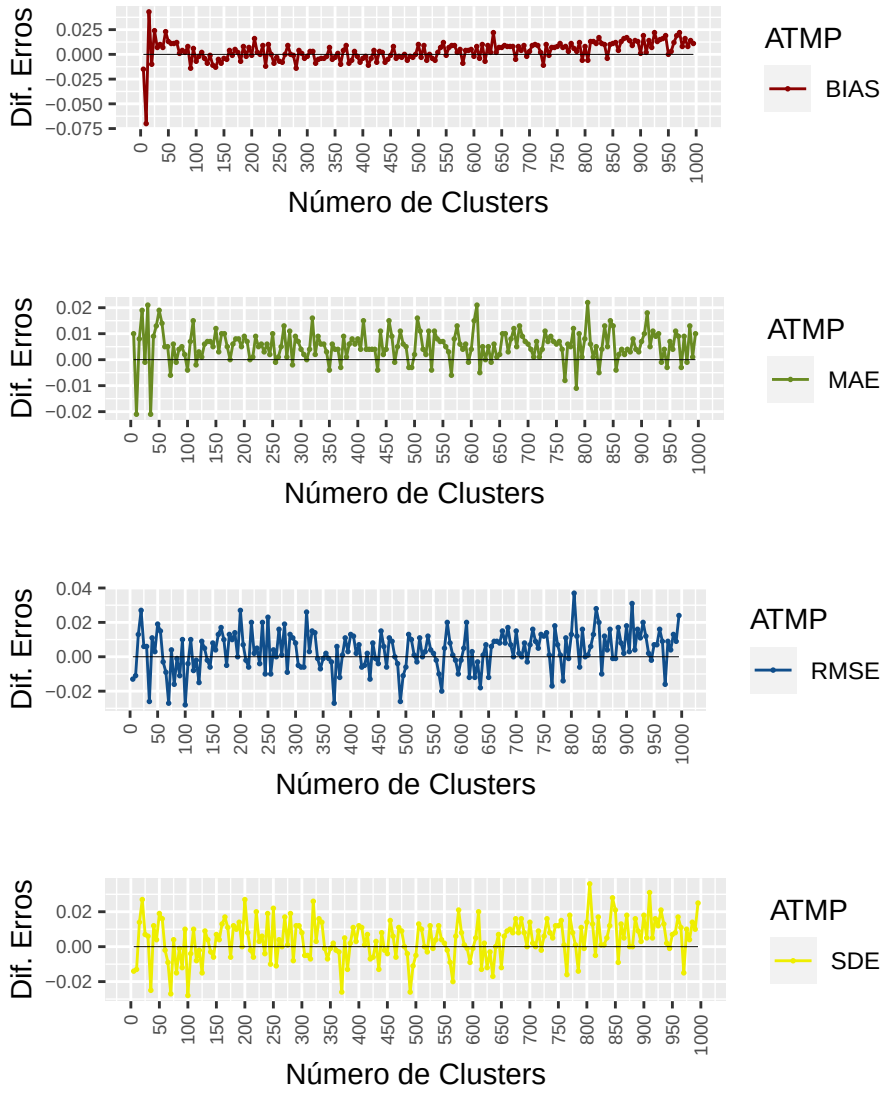


Figura 4.4: Diferenças entre os erros gerados por K-means e por C-means - variável ATMP

à obtida pelo método K-means, sendo as diferenças entre esses erros muito pequenas. Isso pode ser confirmado pela análise dos gráficos da Figura 4.4. Neste caso, para a variável ATMP, as diferenças são maioritariamentre positivas, sinal de que o método K-means tende a gerar erros ligeiramente maiores que o método C-means. Em termos absolutos, com C-means, os cenários que minimizaram mais erros foram $N_c = 320$ e $N_c = 250$.

O gráfico da distribuição dos erros usando C-means está no apêndice C (Figura C.1).

Variável GST

O gráfico da Figura 4.5 mostra os erros das previsões da variável meteorológica GST com o método K-means. O número de *clusters* que minimiza mais erros é agora $N_c = 280$.

A distribuição dos erros tem semelhanças com a observada para a variável ATMP: os erros estabilizam com $N_c \geq 100$, sendo mais pronunciados no intervalo $1 \leq N_c \leq 25$, com os erros mais altos exibidos para valores de N_c menores (correspondentes a *clusters* de maior dimensão) – isto também com excepção do BIAS, que se mantém sempre relativamente constante. No entanto, em termos absolutos, os erros observados estabilizam em valores maiores que os da variável ATMP, o que pode indicar que as previsões feitas para a variável GST não são tão precisas como as previsões da variável ATMP.

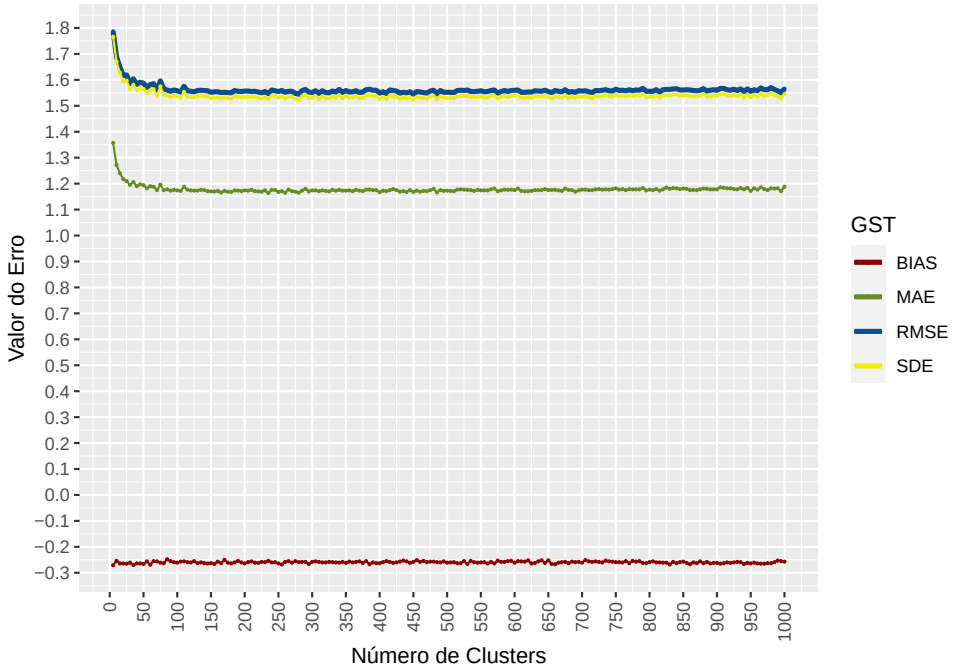


Figura 4.5: Erros das previsões da variável GST para diferentes N_c - método K-means

Aplicando o método C-means, obtém-se ainda um gráfico similar ao obtido com o método K-means (ver Figura C.2 no apêndice C). As semelhanças são comprováveis pela consulta dos gráficos da Figura 4.6, que mostram as pequenas diferenças entre os erros obtidos pelo método K-means e pelo método C-Means, agora sem uma vantagem clara para

qualquer um destes métodos (com excepção do erro MAE, que é sistematicamente superior no método K-means). Com o método C-means, o número de *clusters* que minimiza mais erros em simultâneo na previsão da variável GST é de $N_c = 550$.

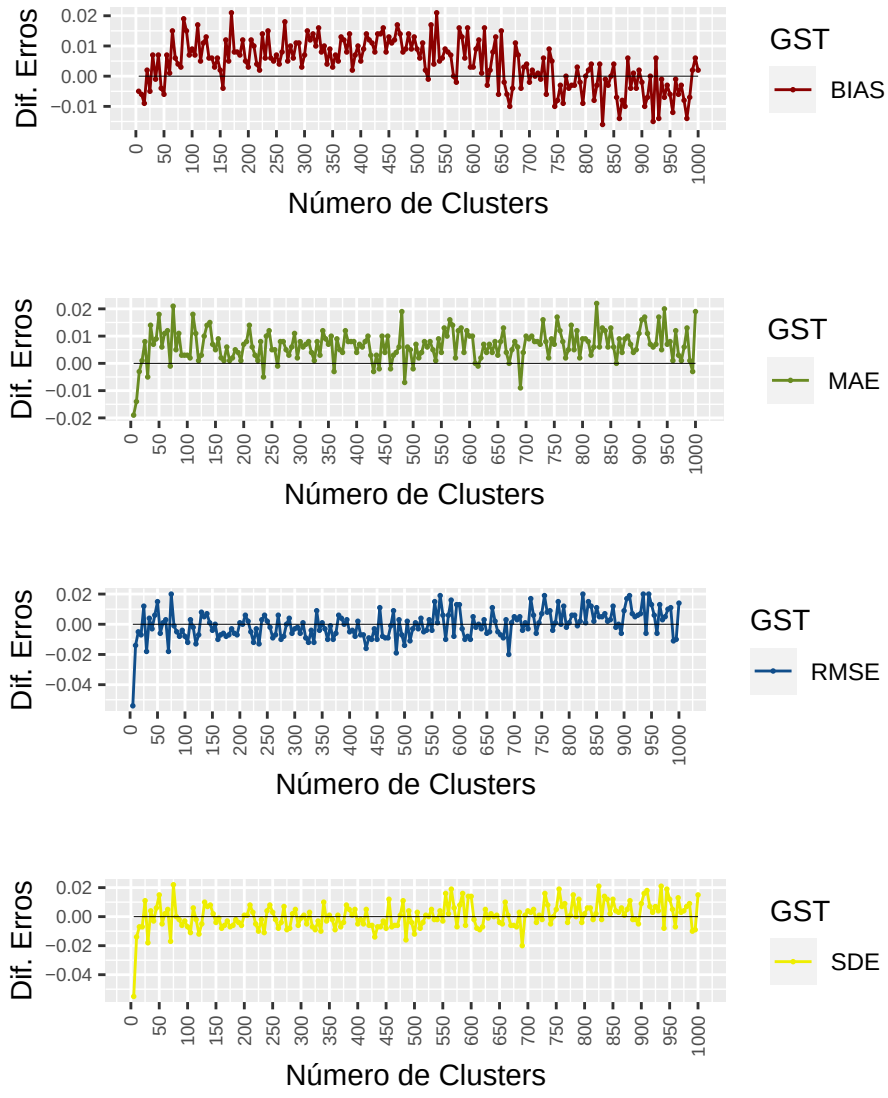


Figura 4.6: Diferenças entre os erros gerados por K-means e por C-means - variável GST

Variável PRES

Para a variável PRES, os diferentes erros nas previsões com o método K-means, para diferentes valores de N_c, estão representados na Figura 4.7. Novamente, o gráfico gerado

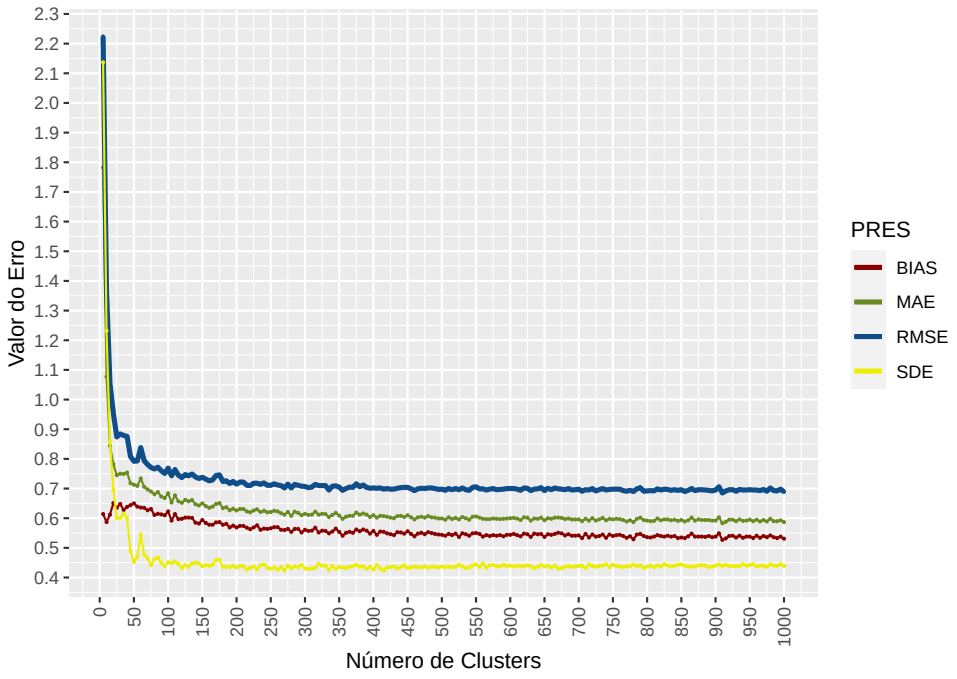


Figura 4.7: Erros das previsões da variável PRES para diferentes N_c - método K-means

pelo método C-means tem uma configuração semelhante, tendo-se remetido para apêndice (Figura C.3). Desde logo, é perceptível uma grande diferença em comparação com as variáveis ATMP e GST: observa-se uma maior variação dos erros antes de estes atingirem um patamar de alguma estabilização (em que continuam a diminuir ligeiramente).

Além disso, desta vez, os valores obtidos para o erro BIAS são positivos, indicando um desvio sistemático da previsão, em que os valores previstos ficam além dos valores reais. Por outro lado, como foi referido na seção 2.8, quanto menores os valores obtidos para os erros MAE e RMSE, mais próximas da verdade estão as previsões. Desta forma, pode considerar-se que as previsões feitas da variável PRES são as que apresentam maior exatidão, pois os valores observados para estes dois erros são inferiores aos restantes casos estudados. Globalmente no caso da variável PRES, a minimização dos erros é obtida com $N_c = 910$ e $N_c = 995$, para as variantes K-means e C-means, respectivamente.

Analisando ainda os gráficos da Figura 4.8, que mostram as diferenças entre os erros gerados pelo método K-means e pelo método C-means, é visível que o método K-means

quase sempre (quando $N_c \geq 75$) gera menores erros (com excepção do SDE).

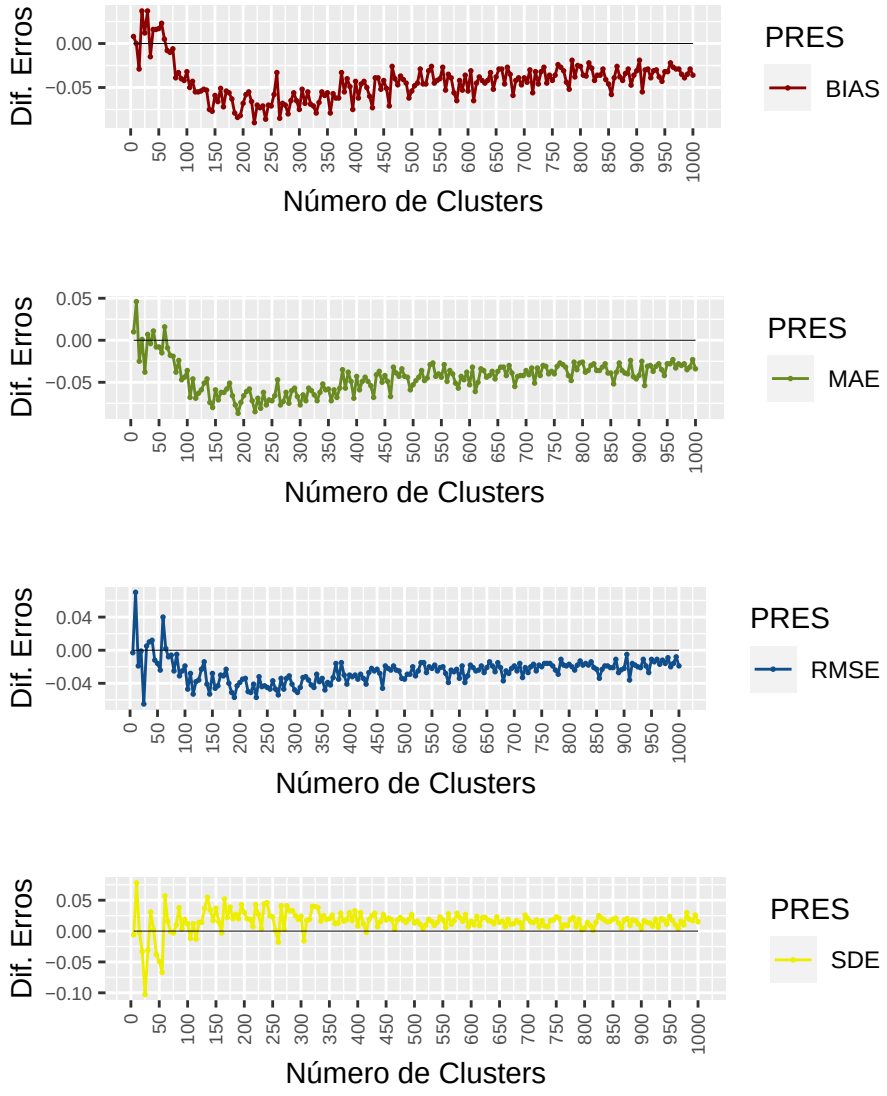


Figura 4.8: Diferenças entre os erros gerados por K-means e por C-means - variável PRES

Variável WPSD

Por fim, analisam-se os erros em função do número de *clusters* na previsão da variável WSPD. Os resultados estão apresentados nas Figuras 4.9 e C.4, onde é possível observar um comportamento bastante semelhante ao já observado para variável a GST. Muito provavelmente, isto deve-se ao fato das duas variáveis estarem fortemente correlacionadas.

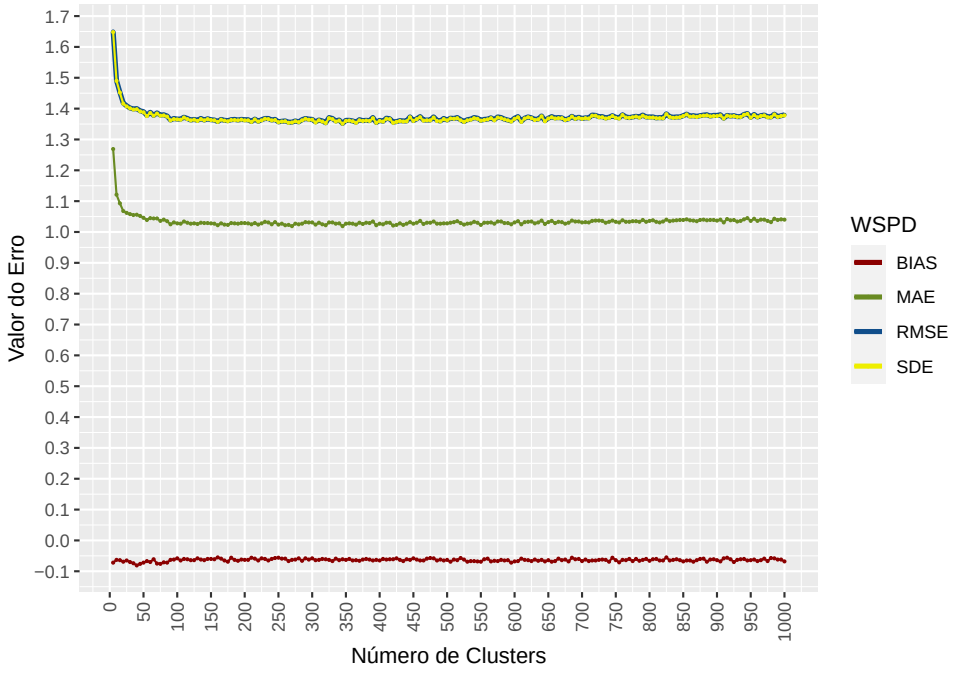


Figura 4.9: Erros das previsões da variável WSPD para diferentes N_c - métodos K-means

Novamente há uma estabilização da média dos valores dos erros para testes com mais de $N_c = 100$ *clusters*. Para o gráfico da Figura 4.9, montado a partir dos resultados obtidos utilizando a variante K-means, é perceptível um leve aumento dos erros (com exceção do BIAS), à medida que N_c cresce. Para o método C-means (Figura C.4), a média dos vários erros mantém-se mais constante com $N_c > 100$. Globalmente, os erros mais baixos obtidos com os métodos K-means e C-means ocorrem com $N_c = 345$ e $N_c = 195$, respetivamente.

Focando a atenção nas diferenças dos erros obtidos pelo método K-means e pelo método C-means (Figura 4.10), o primeiro gera erros BIAS menores, sendo os restantes erros contrabalançados entre o K-means e o C-means, dependendo da faixa de valores de N_c.

Discussão

A partir dos gráficos apresentados é possível notar uma semelhança no comportamento dos erros, com a variação de N_c, para os casos das variáveis ATMP, GST e WSPD. Mas no caso da variável PRES observa-se um comportamento distinto. Isso ocorre possivelmente

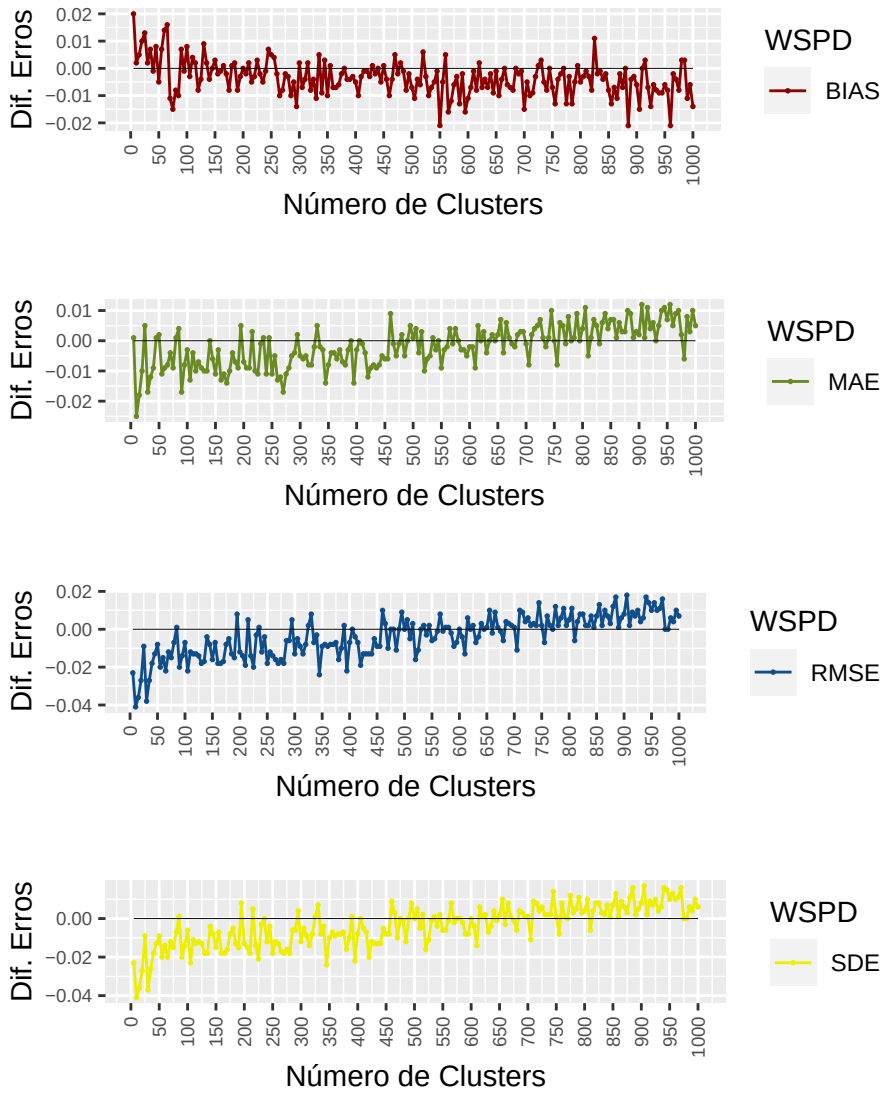


Figura 4.10: Diferenças entre os erros gerados por K-means e por C-means - variável WSPD

pelo fato da variável PRES não possuir grandes oscilações ao longo de um curto período de tempo, fazendo com que sejam formados *clusters* mais bem definidos, que possuam possíveis análogos muitas vezes com os mesmos valores ou a mesma variação.

As previsões com menores erros, que poderão corresponder a previsões mais exatas, foram obtidas com a variável PRES, seguido-se as variáveis ATMP, WSPD e GST.

Dito isto, determinar um único valor de N_c, afirmando que ele seria o melhor para

qualquer um dos cenários, é uma tarefa difícil por conta de cada variável possuir um comportamento diferente. Claramente, selecionar um N_c abaixo de 100 não é uma boa opção, para qualquer uma das variáveis. Para as variáveis ATMP, GST e WSPD há uma pequena tendência de aumento dos erros à medida que o valor de N_c cresce, o que justificaria não escolher grandes valores para N_c. Mas isso não se aplica à variável PRES, uma vez que possui uma contante diminuição dos erros à medida que N_c aumenta.

Uma heurística apresentada em investigação anterior relacionada [27], consiste em escolher N_c como a raiz quadrada do número total de possíveis análogos ($N_c \approx \sqrt{P_{a_t}}$), sendo P_{a_t} o número de possíveis análogos presentes no período de treino. Essa revela-se uma escolha bastante acertada tendo em conta os estudos aqui apresentados, em que esse valor estaria próximo de 220. $N_c = 220$ é superior ao limite observado de diminuição dos erros que ocorre até cerca de $N_c = 100$ *clusters*. Valores inferiores a $N_c = 100$ devem claramente ser evitados. Ainda assim não seria um valor ótimo para a variável PRES, pois para esta variável os menores erros ocorrem para elevados números de *clusters*.

4.2 Número de Análogos

O Número de Análogos (N_a) define quantos elementos do período de treino (os N_a elementos mais similares ao preditor), serão usados para definir um número igual de observações, que por sua vez serão usadas para fazer a previsão. Durante os testes realizados, o parâmetro N_a assumiu diversos valores para que fosse possível entender o impacto da sua variação. Para todas as variantes do método AnEn (Monache, K-means e C-means), o valor de N_a pode teoricamente variar de 1 ao número total de possíveis análogos (P_{a_t}).

Nos métodos de clusterização, porém, não é comum ser formado um único *cluster*, com P_{a_t} elementos, mas sim vários *clusters*, com menos elementos; quando um desses *clusters* é selecionado, devido à similaridade do seu centroide com o preditor, todos os P_a elementos desse *cluster* passam a ser considerados possíveis análogos; é então aqui que entra em jogo o parâmetro N_a: se $P_a \geq N_a$, então apenas os N_a elementos do *cluster* que estão mais próximos do centroide serão selecionados como análogos; se $P_a < N_a$, então todos os P_a

elementos do *cluster* serão selecionados como análogos.

No método de Monache, a escolha dos análogos é feita a partir do valor da métrica apresentada na Equação 2.2. Já nas variantes com clusterização, a escolha dos análogos é feita a partir do valor obtido da métrica apresentada na Equação 2.4.

Mais adiante (secção 4.5), é feita uma comparação entre métodos que utilizam a abordagem de estações dependentes e independentes, onde o valor escolhido para N_a tem que ser visto de forma diferente para cada um dos métodos. Com efeito, há uma diferença relevante na quantidade de análogos selecionados entre os métodos com estações dependentes e independentes. Por exemplo, ao especificar-se um valor $N_a = 10$, no método dependente, como os valores meteorológicos das estações preditoras são agrupados em um mesmo vetor, apenas 10 análogos são selecionados, enquanto que, no método independente, seriam selecionados 10 análogos para cada uma das estações preditoras.

Para a análise da variável N_a, foi assumido $k = 5$ e uma gama de variação de N_a entre 50 e 400, em intervalos de 50 unidades. Para as variantes com clusterização foi realizado um teste adicional em relação ao método de Monache, onde não foi especificado nenhum valor para N_a, implicando que todos os P_a do melhor *cluster* foram usados para os cálculos das previsões. Foi também necessário definir um valor fixo para o número de *clusters* (N_c); como a variável PRES demonstrou ser uma exceção em relação ao comportamento dos erros durante a variação de N_c, o valor escolhido para N_c teve um foco maior na tendência dos melhores erros das variáveis ATMP, GST e WSPD; assim, assumiu-se $N_c = 350$, um valor próximo da média (368,75) dos valores de N_c que se verificou produzirem os menores erros na previsão das variáveis ATMP, GST e WSPD (ver secção anterior).

4.2.1 Erros para Diferentes Valores do Número de Análogos

Os resultados apresentados nas Tabelas 4.1 a 4.4, são de testes realizados para as quatro variáveis meteorológicas, a fim de avaliar o impacto da variação de N_a nos erros das

previsões. Nas tabelas são destacadas (a negrito) as linhas com os menores erros[1], indicando uma maior proximidade com os valores reais, i.e., melhores resultados. As linhas onde o valor de N_a se representa por '—' indicam os testes onde nenhum valor de N_a foi especificado, ditando que fossem selecionados todos os P_a possíveis análogos do *cluster*.

Tabela 4.1: Erros de previsão da variável GST para diferentes valores de N_a

Na	Método	BIAS	RMSE	MAE	SDE
—	K-means	-0,263	1,559	1,172	1,537
—	C-means	-0,260	1,558	1,166	1,536
50	**Monache**	**-0,733**	**2,255**	**1,711**	**2,132**
50	K-means	-0,341	1,579	1,191	1,542
50	C-means	-0,351	1,570	1,18	1,530
100	Monache	-0,733	2,259	1,713	2,136
100	K-means	-0,317	1,564	1,179	1,532
100	C-means	-0,329	1,566	1,173	1,531
150	Monache	-0,731	2,263	1,715	2,142
150	K-means	-0,284	1,571	1,184	1,545
150	C-means	-0,308	1,563	1,172	1,532
200	Monache	-0,730	2,268	1,719	2,147
200	K-means	-0,269	1,563	1,178	1,539
200	C-means	-0,292	1,561	1,167	1,533
250	Monache	-0,728	2,271	1,721	2,151
250	K-means	-0,260	1,554	1,174	1,532
250	C-means	-0,294	1,562	1,171	1,534
300	Monache	-0,725	2,273	1,722	2,155
300	K-means	-0,257	1,561	1,176	1,540
300	**C-means**	**-0,284**	**1,55**	**1,161**	**1,524**
350	Monache	-0,722	2,275	1,724	2,158
350	K-means	-0,261	1,557	1,172	1,535
350	C-means	-0,275	1,556	1,166	1,531
400	Monache	-0,720	2,277	1,725	2,161
400	**K-means**	**-0,260**	**1,553**	**1,175**	**1,531**
400	C-means	-0,277	1,556	1,167	1,531

De forma sumária (e ignorando os erros BIAS), na previsão das variáveis GST, ATMP, PRES e WSPD, os métodos que implicam menores erros absolutos de previsão são, respetivamente, C-means com $N_a = 300$, K-means com $N_a = 300$, Monache com $N_a = 150$ e K-means com $N_a = 300$. Ou seja, para o *dataset* usado, são métodos de clusterização a

[1]Para esta escolha são considerados os menores erros RMSE, MAE, SDE e BIAS, por esta ordem.

Tabela 4.2: Erros de previsão da variável ATMP para diferentes valores de N_a

Na	Método	BIAS	RMSE	MAE	SDE
—	K-means	-0,037	1,040	0,695	1,039
—	C-means	-0,051	1,034	0,686	1,033
50	Monache	0,002	1,078	0,739	1,078
50	K-means	-0,056	1,049	0,703	1,047
50	C-means	-0,043	1,049	0,700	1,048
100	Monache	0,001	1,071	0,730	1,071
100	K-means	-0,055	1,048	0,696	1,046
100	C-means	-0,038	1,041	0,691	1,040
150	Monache	0,003	1,067	0,727	1,067
150	K-means	-0,058	1,047	0,696	1,045
150	C-means	-0,042	1,043	0,688	1,042
200	Monache	0,005	1,067	0,727	1,067
200	K-means	-0,049	1,03	0,691	1,029
200	**C-means**	**-0,051**	**1,032**	**0,687**	**1,031**
250	**Monache**	**0,007**	**1,066**	**0,726**	**1,066**
250	K-means	-0,060	1,048	0,696	1,046
250	C-means	-0,057	1,044	0,693	1,042
300	Monache	0,008	1,067	0,726	1,067
300	**K-means**	**-0,052**	**1,027**	**0,691**	**1,026**
300	C-means	-0,056	1,047	0,690	1,045
350	Monache	0,010	1,068	0,727	1,068
350	K-means	-0,050	1,03	0,688	1,028
350	C-means	-0,057	1,045	0,691	1,043
400	Monache	0,0120	1,069	0,727	1,069
400	K-means	-0,053	1,041	0,697	1,040
400	C-means	-0,051	1,035	0,691	1,033

fazer as melhores previsões e tal é conseguido com um número de análogos $N_a = 300$.

É importante no entanto perceber o impacto da escolha de valores de N_a que não os ótimos, pois esse impacto pode até ser bastante limitado, ao ponto de não se justificar o esforço da busca do valor de N_a que minimiza os erros em termos absolutos. Para o efeito, podem-se calcular as diferenças percentuais entre os erros obtidos com diferentes valores de N_a, relativamente aos obtidos com o N_a ótimo (o que minimiza o erro). Basicamente, se x for o menor erro, a diferença percentual de um outro erro y face a x é dada por $(y-x)/x \times 100\%$. As tabelas C.3a a C.3c (variável PRES), C.2a a C.2c (variável ATMP), C.1a a C.1c (variável GST) e C.4a a C.4c (variável WSPD), no apêndice C, contêm essas

Tabela 4.3: Erros de previsão da variável PRES para diferentes valores de N_a

Na	Método	BIAS	RMSE	MAE	SDE
—	K-means	0,561	0,712	0,618	0,439
—	C-means	0,635	0,754	0,69	0,406
50	Monache	0,278	0,500	0,391	0,416
50	**K-means**	**0,528**	**0,679**	**0,580**	**0,427**
50	**C-means**	**0,546**	**0,682**	**0,594**	**0,408**
100	Monache	0,274	0,495	0,389	0,412
100	K-means	0,562	0,702	0,613	0,420
100	C-means	0,595	0,719	0,647	0,404
150	**Monache**	**0,273**	**0,494**	**0,390**	**0,411**
150	K-means	0,559	0,703	0,614	0,427
150	C-means	0,608	0,729	0,658	0,403
200	Monache	0,273	0,495	0,390	0,413
200	K-means	0,561	0,711	0,617	0,436
200	C-means	0,615	0,742	0,673	0,415
250	Monache	0,273	0,498	0,391	0,417
250	K-means	0,555	0,706	0,612	0,436
250	C-means	0,608	0,737	0,665	0,417
300	Monache	0,273	0,501	0,393	0,421
300	K-means	0,557	0,710	0,614	0,440
300	C-means	0,600	0,738	0,661	0,429
350	Monache	0,273	0,505	0,394	0,425
350	K-means	0,553	0,705	0,610	0,436
350	C-means	0,615	0,741	0,666	0,414
400	Monache	0,272	0,510	0,395	0,431
400	K-means	0,565	0,706	0,618	0,423
400	C-means	0,625	0,742	0,674	0,399

diferenças considerando os erros das tabelas 4.1 a 4.4. Para facilitar a análise que se pretende, atente-se na tabela 4.5, que mostra a máxima diferença percentual registada.

Analisando a tabela 4.5 na vertical, percebe-se que, com excepção da variável PRES, os erros nos três métodos de previsão não variam muito quado N_a se afasta do ótimo, pois a máxima diferença percentual dos erros face aos menores erros é relativamente baixa (entre 0,91% e 2,14%); significa isto, portanto, que a escolha de outro N_a que não o ótimo irá fazer crescer os erros muito pouco; no entanto, para a variável PRES, os erros crescem de forma mais evidente (e mais pronunciada nos métodos de clusterização).

Para finalizar esta análise, considera-se qual seria impacto nos erros de não indicar

Tabela 4.4: Erros de previsão da variável WSPD para diferentes valores de N_a

Na	Método	BIAS	RMSE	MAE	SDE
—	K-means	-0,059	1,355	1,024	1,354
—	C-means	-0,062	1,369	1,031	1,368
50	**Monache**	**-0,35**	**2,107**	**1,578**	**2,078**
50	K-means	-0,121	1,377	1,032	1,372
50	C-means	-0,117	1,380	1,039	1,375
100	Monache	-0,342	2,109	1,580	2,081
100	K-means	-0,097	1,366	1,025	1,362
100	C-means	-0,099	1,374	1,032	1,371
150	Monache	-0,335	2,109	1,581	2,083
150	K-means	-0,075	1,362	1,026	1,360
150	**C-means**	**-0,086**	**1,361**	**1,026**	**1,358**
200	Monache	-0,331	2,113	1,584	2,087
200	K-means	-0,067	1,357	1,024	1,355
200	C-means	-0,076	1,368	1,029	1,366
250	Monache	-0,328	2,115	1,586	2,089
250	K-means	-0,068	1,365	1,027	1,364
250	C-means	-0,074	1,378	1,037	1,376
300	Monache	-0,325	2,117	1,588	2,092
300	**K-means**	**-0,060**	**1,352**	**1,019**	**1,351**
300	C-means	-0,070	1,379	1,037	1,377
350	Monache	-0,322	2,12	1,59	2,095
350	K-means	-0,060	1,358	1,024	1,357
350	C-means	-0,062	1,371	1,034	1,369
400	Monache	-0,320	2,122	1,591	2,097
400	K-means	-0,057	1,364	1,028	1,362
400	C-means	-0,069	1,374	1,034	1,372

Tabela 4.5: Máxima diferença percentual dos erros face aos menores erros

Método/Variável	GST	ATMP	PRES	WPSD
Monache	1,36%	1,79%	4,87%	0,91%
K-means	1,67%	2,14%	6,55%	1,85%
C-means	1,64%	1,89%	16,16%	1,4%

nenhum valor de N_a para os métodos de clusterização (utilizando-se assim todos os P_a possíveis análogos do melhor *cluster*). Esse impacto pode ser aferido comparando os erros obtidos nessa situação, com os obtidos quando se usa o valor de N_a que minimiza os erros. Essa comparação pode ser feita nas tabelas 4.1 a 4.4, mas para facilitar a mesma,

resume-se na tabela 4.6 a informação pertinente. Nesta tabela, para cada variável e para cada método de clusterização, apresentam-se os menores erros absolutos e o N_a respetivo, e a diferença percentual face a esses erros quando não se especifica N_a.

Tabela 4.6: Menores erros absolutos com indicação de N_a, e diferença percentual dos erros face aos menores erros absolutos quando não se indica o número de análogos (N_a='—')

Variável	Método	Na	RMSE	MAE	SDE
GST	K-means	—	0,39%	-0,26%	0,39%
		400	1,553	1,175	1531%
	C-means	—	0,52%	0,43%	0,79%
		(*) 300	1,55	1,161	1,524
ATMP	K-means	—	**1,27%**	0,58%	**1,27%**
		(*) 300	1,027	0,691	1,026
	C-means	—	0,19%	**-0,15%**	0,19%
		200	1,032	0,687	1,031
PRES	K-means	—	4,86%	6,55%	2,81%
		50	0,679	0,580	0,427
	C-means	—	10,56%	**16,16%**	**-0,49%**
		50	0,682	0,594	0,408
WSPD	K-means	—	0,22%	0,49%	0,22%
		(*) 300	1,352	1,019	1,351
	C-means	—	0,59%	0,49%	0,74%
		150	1,361	1,026	1,358

Como se pode observar, para as variáveis GST, ATMP e WSPD, as diferenças percentuais dos erros quando não se define N_a são bastante reduzidas, variando (em módulo) entre 0,15% e 1,27% no conjunto das três variáveis. Já para a variável PRES, a gama de variação dessas diferenças percentuais é maior, oscilando (em módulo) entre 0,49% a 16,16%. Daqui se pode concluir que em relação às variáveis GST, ATMP e WSPD, para as quais os métodos C-means, K-means e K-means, respetivamente, produzem os menores erros com valores específicos (*) de N_a, não veriam a sua previsão especialmente penalizada caso esses métodos fossem usados sem especificar valor de N_a. Já no caso da variável PRES, é o método de Monache, e não nenhum método de clusterização, que gera os menores erros; a propensão a maiores erros com clusterização poderá explicar porque é que, caso se usassem esses métodos, não seria indiferente indicar ou não N_a.

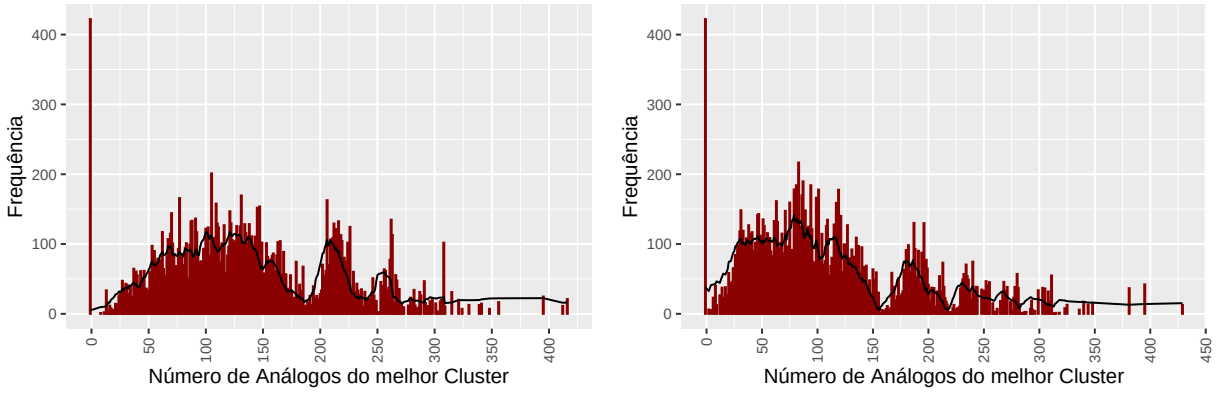
Em suma, e considerando que nesta secção se procurou averiguar o impacto nos erros para diferentes valores de N_a, observou-se que há efetivamente valores que minimizam os erros em termos absolutos, mas com métodos de clusterização não há grande vantagem em especificar um valor de N_a face ao uso de todos os possíveis análogos (P_a).

4.2.2 Possíveis Análogos versus Número de Análogos

Como já foi antes referido (início da secção 4.2), especificar um valor de N_a para os métodos com clusterização pode não ser vinculativo: o número dos elementos agrupados em cada *cluster* pode ser diferente do valor de N_a fornecido; desta forma, o número de possíveis análogos (P_a) do melhor *cluster* poderá ser superior ou inferior a N_a, retendo-se apenas os melhores N_a possíveis análogos, ou todos os P_a possíveis análogos, respetivamente.

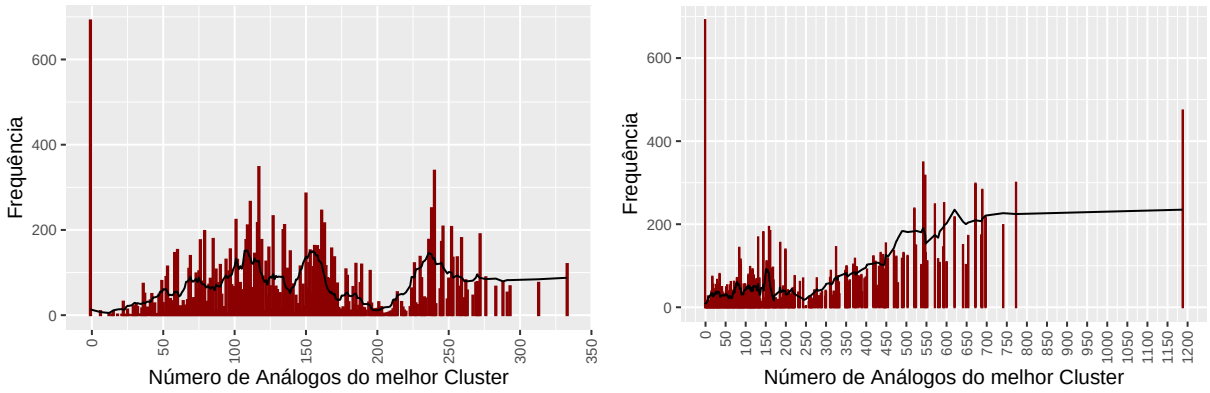
Tendo isto em mente, fizeram-se testes com o objetivo de descobrir os diferentes valores de P_a e a sua frequência, quando se usam os métodos de clusterização sem indicação de um valor de N_a, ou seja, aproveitando sempre todos os P_a análogos do melhor *cluster*. Nestes testes foram usados, como anteriormente, os valores $k = 5$ e $N_c = 350$.

Os resultados destes testes podem ser observados nos gráficos das Figuras 4.11 a 4.14. Estes gráficos mostram a frequência com que ocorre um certo valor, valor esse que corresponde ao número de análogos do melhor *cluster*, durante a previsão das quatro variáveis em estudo, com os métodos K-means e C-means.



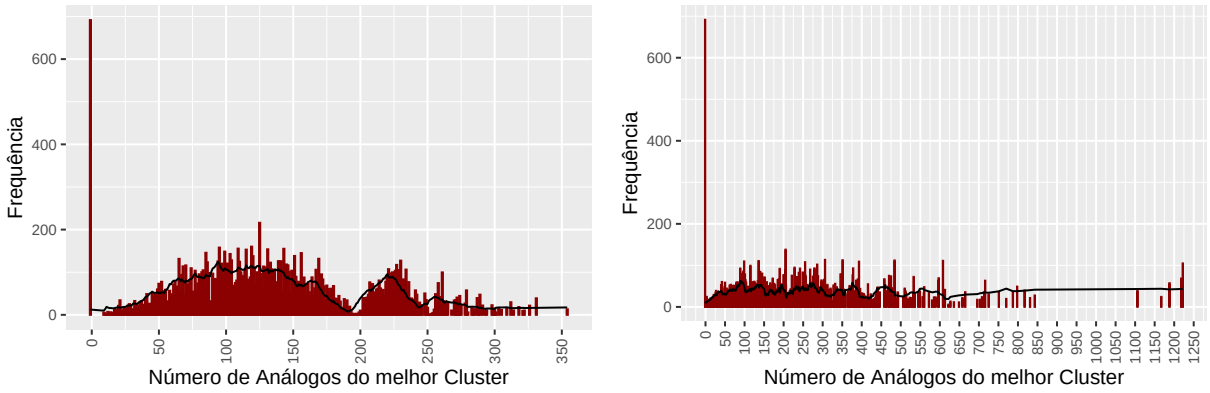
(a) K-means (b) C-means

Figura 4.11: Frequência dos Números de Análogos do melhor Cluster (variável ATMP)



(a) K-means (b) C-means

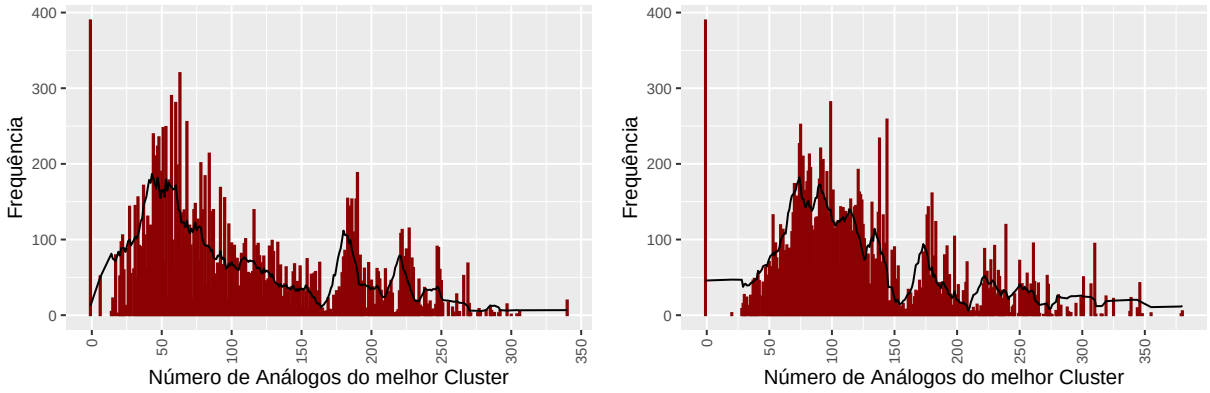
Figura 4.12: Frequência dos Números de Análogos do melhor Cluster (variável GST)



(a) K-means (b) C-means

Figura 4.13: Frequência dos Números de Análogos do melhor Cluster (variável WSPD)

Antes de discutir os gráficos, esclarece-se um detalhe comum a todos: naturalmente, a abcissa zero deveria ter também zero como ordenada, pois nunca são criados *clusters* vazios; no entanto, nos vários gráficos, a ordenada associada à abcissa zero contabiliza o número de situações em que nenhum *cluster* foi selecionado como sendo o melhor; isso ocorre quando a janela de dados meteorológicos, que é selecionada das estações preditoras, e é utilizada para realizar as comparações com os centroides de cada *cluster*, não possui ao menos 50% dos seus valores como válidos (por exemplo, se esse vetor de dados for de tamanho 10 e possuir mais de 5 valores NA – *Not Available* – em sua composição, a previsão que seria feita com esse conjunto de dados não será feita e o valor final atribuído à



(a) K-means (b) C-means

Figura 4.14: Frequência dos Números de Análogos do melhor Cluster (variável PRES)

previsão será NA). Refira-se que apesar da coluna que emerge da abcissa zero se destacar nos gráficos, esta situação tem pouca expressão (por exemplo, nos testes que deram origem ao gráfico da Figura 4.11.a), as previsões que não foram realizadas correspondem a menos de 2.5% do total de previsões tentadas).

Em conjunto, estes gráficos mostram que os melhores *clusters* podem ter uma grande variedade de tamanhos (número de análogos que os compõem) e que esses tamanhos ocorrem com frequências muito distintas, embora seja possível observar que alguns tamanhos têm mais prevalência que outros. Toda esta variabilidade sugere que é difícil definir valores específicos para N_a, e que defini-los em intervalos de 50 (como se fez no estudo da secção 4.2.1) é porventura uma aproximação demasiado grosseira aos valores ótimos de N_a, dada a forma como se distribuem nos gráficos (com zonas de grande densidade).

Por outro lado, testar exaustivamente todos os valores possíveis de N_a (entre 1 e P_{a_t}) é demasiado pesado, em termos computacionais. Neste contexto, os gráficos em causa são bastante importantes, porque na prática indicam quais os valores de N_a que vale a pena considerar (caso se queiram indicar de forma explícita), de forma a não correr o risco de perder nenhum análogo do melhor *cluster*: deve-se considerar qualquer valor para N_a que seja maior do que a última abcissa do gráfico para a qual a ordenada seja maior que zero; por exemplo, na Figura 4.11, seria de considerar $N_a >= 416$; de facto, arriscando um valor inferior para N_a, poderão não se considerar alguns análogos do melhor *cluster*

(porque poderia acontecer $N_a < P_a$), significando que também algumas observações não se iriam considerar na previsão, o que poderia aumentar o erro da mesma.

Dito isto, os gráficos também permitem responder a outras questões, quando cruzados com as tabelas produzidas na secção 4.2.1. Em particular: qual seria a consequência de indicar para N_a o valor encontrado como ótimo nessas tabelas ? Nas secções seguintes faz-se essa análise para as variantes K-means e C-means.

Variante K-means

Para a variável ATMP, na Tabela 4.2 os menores erros do método K-means foram obtidos com $N_a = 300$; nessa situação, se para o melhor *cluster* houver mais que 300 possíveis análogos ($P_a > N_a$), apenas $N_a = 300$ são selecionados para o cálculo da previsão, mas se houver menos de 300 possíveis análogos ($P_a < N_a$) todos eles serão usados. Ora, na Figura 4.11.a) foram poucas as situações (na realidade, apenas cerca de 2% do total de situações registadas) em que os melhores *clusters* tinham mais de 300 possíveis análogos; isso significa que na maioria das vezes todos os análogos pertencentes ao melhor *cluster* foram utilizados. Este tipo de análise pode-se estender às restantes variáveis meteorológicas:

- **variável WSPD**: apenas 1% dos *clusters* utilizados (Figura 4.13.a)) possuía mais de 300 possíveis análogos (valor ótimo de N_a na Tabela 4.4);
- **variável GST**: nenhum *cluster* utilizado (Figura 4.12.a)) possuía mais de 400 possíveis análogos (melhor valor N_a na Tabela 4.1); isso explica a grande proximidade entre os erros obtidos com a utilização de todos os possíveis análogos e os erros obtidos com $N_a = 400$; nesse caso, a diferença entre esses dois cenários justifica-se apenas na aleatoriedade dos centroides iniciais escolhidos para a formação dos *clusters*; essa é a única fase possível em que esses dois cenários podem divergir; esta fase influencia a formação dos *clusters* e consequentemente a escolha dos análogos utilizados para as previsões;
- **variável PRES**: a percentagem de *clusters* utilizados (Figura 4.14.a)) que possuem mais de 50 possíveis análogos (valor ótimo de N_a com K-means na Tabela 4.3) é de

78%; sendo assim, é a única variável, no caso da variante K-means, para a qual se justificaria o uso de um valor de N_a diferente da dimensão P_a do *cluster*.

Variante C-means

Analisando agora a variante C-means, foi possível perceber que houve uma grande diferença entre esta e os resultados obtidos com a variante K-means.

- **variável ATMP**: variável em que se verificou a menor percentagem de utilização de *clusters* com um número P_a superior ao melhor valor de N_a identificado na secção 4.2.1; assim, considerando a previsão da variável ATMP com a variante C-means, a Tabela 4.2 regista os melhores resultados com $N_a = 200$; por seu turno, com base nos dados que sustentam a Figura 4.11.b), consegue-se saber que os casos em que os melhores *clusters* possuem mais de 200 análogos correspondem a 13%;

- **variável GST**: na variante C-means obteve os menores erros com um $N_a = 300$ (Tabela 4.1); com base no gráfico da Figura 4.12.b) é possível perceber que a percentagem de *clusters* usados com mais de 300 possíveis análogos foi de 50%;

- **variável WSPD**: essa percentagem foi ainda maior (Figura 4.13.b)): para 64% dos *clusters* usados $P_a \geq 150$, com 150 identificado como melhor N_a (Tabela 4.4);

- **variável PRES**: novamente se destaca com uma percentagem de 94% dos *clusters* utilizados tendo mais de 50 possíveis análogos ($N_a = 50$ é o melhor valor na variante C-means na Tabela 4.3), sendo possível confirmar isso na Figura 4.14.b).

Discussão

Levando em consideração a afirmação feita anteriormente para a variável PRES da variante K-means, era possível dizer que faria sentido especificar um valor de N_a na variante C-means não apenas para a variável PRES mas sim também para as variáveis GST e WSPD que obtiveram elevadas percentagens de uso de clusters com dimensões superiores aos respetivos números recomendados de análogos. No entanto, essa não é a conclusão

a que se chega ao observar as porcentagens de diferença entre os valores dos erros das variáveis GST e WSPD, presentes nas Tabelas C.1b e C.4b.

A Tabela C.4b mostra essas percentagens para os erros da variável WSPD, que claramente mostra que os erros obtidos são muito próximos uns dos outros. Novamente, os valores dos erros obtidos com o melhor número de análogos estão próximos dos valores dos erros obtidos com outros valores de N_a, principalmente dos obtidos sem especificar N_a. Não especificar um valor de N_a causa um impacto nos erros de cerca de 0.6% (ver Tabela C.4b). Para a variável GST esse impacto é de em média 0.5% (ver Tabela C.1b).

Já a variável PRES na variante C-means, possui as mais elevadas percentagens de diferenças entre os erros, variando negativamente em média 14%. Essa percentagem é entre os testes feitos com o valor recomendado do N_a e o caso com a utilização de todos os possíveis análogos do *cluster* (ver Tabela C.3b).

Com base nos resultados dos testes realizados conclui-se não ser possível recomendar um N_a ótimo, comum às variáveis ATMP, GST e WSPD, qualquer que seja o método de clusterização. O melhor é não especificar nenhum N_a e utilizar todos os possíveis análogos incluídos no *cluster* escolhido. As diferenças entre os erros obtidos com os melhores valores de N_a e sem indicação destes valores são bastante pequenas. Além disso, não indicar um valor para N_a implica que não será necessário considerar o tamanho dos *clusters* que serão usados e também não será necessário realizar cálculos extras para identificar os melhores análogos dentro dos *clusters*, reduzindo assim o esforço computacional.

Para a variável PRES, verificou-se na seção 4.1 que os menores erros foram obtidos com um elevado número de *clusters* (N_c) e isso implica que os *clusters* formados nesse cenário teriam um valor médio de análogos por *cluster* muito menor em relação aos testes da secção 4.2. Ou seja, faz sentido a necessidade de especificar um valor de N_a pequeno quando há bastante análogos por *cluster*. Caso se use um valor mais elevado de N_c, a especificação de N_a não é necessária. Portanto a especificação ou não de um valor de N_a quando se estiver trabalhando com a variável PRES depende da quantidade de *clusters* formados e, no caso desta não ser elevada, pode-se especificar um valor do número de análogos relativamente pequeno como, por exemplo, $N_a = 50$.

4.3 Dimensão da Janela do Análogo

Nesta secção é feito um estudo do impacto da variação do parâmetro k. Como referido anteriormente, k é um parâmetro que determina a dimensão dos vetores com os dados meteorológicos que constituem os possíveis análogos: esses vetores têm $2k + 1$ elementos. Uma vez que os dados meteorológicos utilizados neste trabalho são separados no tempo de 6 em 6 minutos, um incremento de k em uma unidade representa um aumento da janela de tempo do análogo em 12 minutos. Por exemplo, dizer que o parâmetro k possui um valor igual a 5 significa que o vetor cobre uma janela de tempo de 1 hora (30 minutos antes e 30 minutos depois do seu valor central).

Nos testes realizados fez-se variar k entre 1 e 20 (com incrementos unitários) – correspondendo a períodos de tempo que variam de 18 minutos a 4 horas –, ou entre 1 e 10 (também com incrementos unitários) – com janelas de tempo entre 18 minutos e 2 horas –. Tal como na seção anterior, os restantes parâmetros, que já foram testados, possuem valores fixos. Para o método de Monache o valor de N_a foi fixado em 150, por ser um valor médio de N_a entre os melhores identificados para o método Monache. Esse valor foi verificado como sendo o melhor para a variável PRES e para as outras variáveis são encontrados erros bastante próximos aos melhores com esse valor de N_a (ver tabelas 4.1 a 4.4); já para as variantes K-means e C-means nenhum valor de N_a foi especificado, ou seja, todos os possíveis análogos pertencentes ao *clusters* são usados. Para o parâmetro N_c, o seu valor foi novamente fixado em 350 seguindo (o mesmo adotado na seção anterior). Em todos os testes faz-se a reconstituição de uma variável da estação *Ykt* usando os dados da mesma variável das estações *Dom e Ykr*.

Variação de k entre 1 e 20

A Tabela 4.7 mostra os erros na reconstituição da variável meteorológica GST com a variante C-means. O valor de k que gerou o menor conjunto de erros foi k=6, onde só o erro BIAS tem um valor ligeiramente superior face a testes com outros valores de k. É também possível observar que para $k \geq 10$ os erros crescem constantemente. Esta observação

constitui regra para quase todas as variáveis testadas com as diferentes variantes.

Tabela 4.7: Erros de previsão da variável GST com C-means, para $k = 1..20$

k	BIAS	RMSE	MAE	SDE
1	-0,249	1,557	1,170	1,537
2	-0,252	1,565	1,176	1,545
3	-0,257	1,555	1,168	1,534
4	-0,262	1,557	1,167	1,535
5	-0,269	1,560	1,166	1,537
6	**-0,272**	**1,556**	**1,165**	**1,532**
7	-0,269	1,565	1,171	1,542
8	-0,266	1,572	1,173	1,550
9	-0,259	1,576	1,179	1,554
10	-0,268	1,576	1,182	1,553
11	-0,264	1,582	1,188	1,559
12	-0,262	1,579	1,184	1,557
13	-0,269	1,586	1,193	1,563
14	-0,270	1,588	1,194	1,565
15	-0,270	1,599	1,202	1,576
16	-0,278	1,601	1,205	1,577
17	-0,279	1,611	1,214	1,587
18	-0,282	1,613	1,218	1,588
19	-0,283	1,613	1,223	1,588
20	-0,283	1,624	1,224	1,599

Apenas a variável ATMP, com a utilização do método Monache, não seguiu essa regra, sendo que os melhores resultados foram obtidos para múltiplos valores de k, de 10 a 13. Os resultados dessa variável estão na Tabela 4.8, onde é possível ver que para vários valores de k os erros se mantêm (quase) iguais.

Variação de k entre 1 e 10

Dado os testes feitos com a variável ATMP utilizando o método Monache terem alcançado bons resultados com $k = 10$, o estudo do impacto da variação de k restringiu-se, nos restantes testes, ao intervalo entre 1 e 10 (onde $k = 10$ representa uma janela de tempo de duas horas). Os resultados destes testes constam das tabelas 4.9 a 4.11.

Um comportamento similar, onde múltiplos erros têm valores muito similares pode ser visto na Tabela 4.9a. Nela, os resultados obtidos para a variável PRES com o método

Tabela 4.8: Erros de previsão da variável ATMP com Monache, para $k = 1..20$

k	BIAS	RMSE	MAE	SDE
1	0,002	1,072	0,733	1,072
2	0,000	1,070	0,731	1,070
3	0,001	1,070	0,729	1,070
4	0,002	1,069	0,728	1,069
5	0,003	1,067	0,727	1,067
6	0,005	1,066	0,727	1,066
7	0,006	1,066	0,726	1,066
8	0,008	1,066	0,726	1,066
9	0,010	1,066	0,726	1,066
10	**0,011**	**1,065**	**0,726**	**1,065**
11	**0,013**	**1,065**	**0,726**	**1,065**
12	**0,014**	**1,065**	**0,726**	**1,065**
13	**0,016**	**1,065**	**0,726**	**1,065**
14	0,018	1,065	0,727	1,065
15	0,019	1,066	0,728	1,066
16	0,020	1,067	0,728	1,067
17	0,021	1,068	0,729	1,068
18	0,021	1,069	0,730	1,069
19	0,020	1,070	0,731	1,070
20	0,021	1,072	0,732	1,072

Monache são apresentados e os quatro melhores conjuntos de erros, obtidos com valores de k de 7 a 10, estão destacados por conta da proximidade entre eles.

Há pois variáveis onde a variação de k não afeta significativamente as previsões, fazendo com que os valores dos erros obtidos sejam muito próximos uns dos outros. No entanto, também há variáveis em que a variação de k afeta mais nitidamente os valores dos erros.

A Tabela 4.10a mostra resultados de testes onde a variável ATMP foi prevista utilizando K-means. Nessa tabela, os menores erros foram obtidos com $k = 5$. Essa variável, juntamente com a variável PRES, são as duas variáveis mais afetadas pela variação de k na variante K-means (ver também Tabela 4.9b relativa à variável PRES); ainda assim, os erros obtidos para os diferentes valores de k continuam próximos uns dos outros.

A Tabela 4.10b mostra os erros obtidos na previsão da variável ATMP com C-means, tendo-se obtido os menores erros $k = 6$. Novamente, no geral, os erros obtidos nos múltiplos testes são próximos e o resultado $k = 6$ destaca-se ligeiramente no meio de

k	BIAS	RMSE	MAE	SDE
1	0,283	0,499	0,398	0,411
2	0,278	0,497	0,394	0,412
3	0,275	0,497	0,392	0,414
4	0,274	0,497	0,391	0,415
5	0,273	0,494	0,390	0,411
6	0,273	0,491	0,389	0,408
7	**0,274**	**0,488**	**0,388**	**0,404**
8	**0,274**	**0,488**	**0,389**	**0,403**
9	**0,275**	**0,488**	**0,389**	**0,403**
10	**0,275**	**0,488**	**0,389**	**0,404**

(a) Método Monache.

k	BIAS	RMSE	MAE	SDE
1	-0,574	0,717	0,628	0,430
2	-0,571	0,712	0,622	0,425
3	-0,557	0,704	0,612	0,430
4	-0,570	0,713	0,623	0,428
5	-0,550	0,703	0,607	0,438
6	-0,554	0,708	0,610	0,441
7	**-0,548**	**0,701**	**0,605**	**0,437**
8	-0,549	0,705	0,608	0,442
9	-0,554	0,708	0,612	0,441
10	-0,557	0,713	0,618	0,445

(b) Método K-means.

k	BIAS	RMSE	MAE	SDE
1	-0,594	0,734	0,656	0,431
2	**-0,587**	**0,733**	**0,650**	**0,439**
3	-0,616	0,743	0,667	0,416
4	-0,612	0,744	0,671	0,422
5	-0,627	0,750	0,684	0,411
6	-0,625	0,746	0,675	0,408
7	-0,644	0,750	0,691	0,385
8	-0,633	0,748	0,687	0,399
9	-0,641	0,752	0,691	0,393
10	-0,625	0,752	0,686	0,419

(c) Método C-means.

Tabela 4.9: Erros de previsão da variável PRES para $k = 1..10$

erros muito próximos, levando a pensar que essa vantagem é casual. Para aferir essa possibilidade, o teste com $k = 6$ foi repetido várias vezes e os resultados obtidos mostraram que essa combinação de erros ocorre normalmente para $k = 6$, sendo também possível obter erros um pouco maiores, fazendo com que o valor ótimo de k flutue entre 6 e 8.

Com exceção da variável PRES, as variantes K-means e C-means seguem um padrão de melhor valor para k e esse valor fica quase sempre entre 5 e 6. Analisando a variável PRES na variantes K-means (ver Tabela 4.9b), nota-se que os resultados de $k = 5$ e $k = 6$ são bastante próximos do resultado de $k = 7$, podendo assim serem escolhidos sem que haja uma grande alteração na precisão da previsão. Com a variante C-means o cenário para a variável PRES muda um pouco: os erros obtidos com $k = 5$ e $k = 6$ são em média

k	BIAS	RMSE	MAE	SDE
1	-0,067	1,043	0,690	1,041
2	-0,060	1,038	0,690	1,036
3	-0,059	1,046	0,697	1,044
4	-0,058	1,049	0,699	1,047
5	**-0,053**	**1,031**	**0,685**	**1,029**
6	-0,049	1,039	0,693	1,038
7	-0,047	1,048	0,699	1,047
8	-0,043	1,045	0,700	1,044
9	-0,034	1,044	0,705	1,043
10	-0,040	1,042	0,699	1,041

(a) Método K-means.

k	BIAS	RMSE	MAE	SDE
1	-0,064	1,044	0,689	1,042
2	-0,060	1,046	0,690	1,045
3	-0,062	1,049	0,691	1,047
4	-0,053	1,036	0,691	1,035
5	-0,049	1,038	0,690	1,037
6	**-0,050**	**1,028**	**0,685**	**1,027**
7	-0,050	1,037	0,691	1,036
8	-0,046	1,034	0,685	1,033
9	-0,036	1,037	0,69	1,036
10	-0,042	1,034	0,692	1,033

(b) Método C-means.

Tabela 4.10: Erros de previsão da variável ATMP para $k = 1..10$

4% maiores dos que obtidos com $k = 2$ (ver Tabela 4.9c).

Apesar da excepcionalidade do melhor valor de k ser 2 para a variável PRES na variante C-means, a regra parece ser que o valor recomendado de k deve ser 5 ou 6 quando se usam métodos de clusterização. Assim, optou-se por assumir $k = 5$ para as variantes K-means e C-means nos testes seguintes, omitindo-se as tabelas dos erros para as variáveis GST (K-means) e WSPD (K-means e C-means) com esses métodos.

Para o método Monache aplicado às variáveis GST e WSPD, os menores erros obtiveram-se com $k = 2$ e, como se pode observar nas Tabelas 4.11a e 4.11b, destacam-se dos restantes. Além disso, o valor ótimo de k difere do encontrado para as variáveis PRES e ATMP. Para a variável PRES (Tabela 4.9a), a variante Monache obteve os menores erros com k próximo a 7, com erros similares para $7 \leq k \leq 10$. Para a variável ATMP (Tabela 4.8), os menores erros obtiveram-se com k=10, mas com resultados similares para $10 \leq k \leq 13$.

Recomendar um valor de k para o método Monache é pois um pouco mais difícil do que para as variantes K-means e C-means, onde há uma constância dos menores erros em torno de $k = 5$ e $k = 6$. Dado que para as variáveis ATMP e PRES se obteve bons resultados para vários valores de k, parece razoável recomendar $k = 2$ para a variante de Monache, pois duas das quatro variáveis testadas obtiveram melhores resultados com $k = 2$ e as outras duas possuem uma variação pouco significativa entre os melhores resultados e os com $k = 2$. Esse será então o valor de k assumido doravante com o método Monache.

k	BIAS	RMSE	MAE	SDE
1	-0,398	2,190	1,689	2,153
2	**-0,53**	**2,197**	**1,685**	**2,132**
3	-0,623	2,223	1,695	2,134
4	-0,686	2,245	1,707	2,137
5	-0,731	2,263	1,715	2,142
6	-0,763	2,278	1,723	2,146
7	-0,778	2,288	1,729	2,151
8	-0,793	2,297	1,733	2,156
9	-0,805	2,305	1,736	2,160
10	-0,814	2,301	1,739	2,162

(a) Variável GST.

k	BIAS	RMSE	MAE	SDE
1	-0,12	2,08	1,598	2,077
2	**-0,206**	**2,075**	**1,578**	**2,064**
3	-0,273	2,088	1,576	2,071
4	-0,306	2,100	1,579	2,077
5	-0,335	2,109	1,581	2,083
6	-0,35	2,116	1,584	2,087
7	-0,361	2,123	1,586	2,092
8	-0,371	2,127	1,586	2,094
9	-0,38	2,129	1,585	2,095
10	-0,386	2,131	1,584	2,096

(b) Variável WSPD.

Tabela 4.11: Erros de previsão com Monache para $k = 1..10$

Síntese

Os resultados apresentados na Tabela 4.12 são aqueles que obtiveram os melhores conjuntos de erros nas previsões das variáveis meteorológicas (ATMP, GST, PRES e WSPD), tendo em conta as três variantes do método AnEn (Monache, K-means e C-means).

Assim, os melhores valores de k identificados para o método Monache são 2, 2, 7 e 10. Com a variantes K-means esses valores foram 5, 5, 6 e 7. E por fim, para a variante C-means, os melhores resultados foram obtidos com k valendo 2, 5, 6 e 6.

Pese embora alguma variabilidade destes resultados, como referido anteriormente decidiu-se fixar $k = 5$ para as variantes de clusterização, e $k = 2$ para a variante Monache.

4.4 Peso das Observações

Como referido no início deste capítulo, os pesos nas Equações 2.6 e 2.9 definem a importância que as observações, encontradas a partir dos análogos, têm na previsão. Nesta secção estuda-se o impacto de duas formas diferentes de atribuir pesos às observações, algo que apenas faz sentido quando se usam métodos de clusterização.

Numa forma, utilizável em ambas as variante K-means e C-means, os pesos (P_1) são calculados a partir da distância dos análogos selecionados ao centroide do *cluster*; quanto mais próximo um análogo for do centroide mais relevância terá na previsão a observação

Tabela 4.12: Valores de k que produziram os menores erros

k	Variável	Método	BIAS	RMSE	MAE	SDE
10	ATMP	Monache	-0,011	1,065	0,726	1,065
5	ATMP	K-means	-0,053	1,031	0,685	1,029
6	ATMP	C-means	-0,050	1,028	0,685	1,027
2	GST	Monache	-0,530	2,197	1,685	2,132
5	GST	K-means	-0,261	1,556	1,171	1,534
6	GST	C-means	-0,272	1,556	1,165	1,532
7	PRES	Monache	0,274	0,488	0,388	0,404
7	PRES	K-means	0,548	0,701	0,605	0,437
2	PRES	C-means	0,587	0,733	0,650	0,439
2	WSPD	Monache	-0,206	2,075	1,578	2,064
6	WSPD	K-means	-0,064	1,357	1,025	1,356
5	WSPD	C-means	-0,058	1,364	1,030	1,363

correspondente. A outra forma de atribuir os pesos apenas pode ser utilizada na variantes C-means, dado apenas esta gerar um valor de *membership* ou grau de associação; os pesos (P_2) são obtidos a partir dos graus de associação de cada análogo; quanto maior for esse valor, maior será relevância da observação no valor da previsão.

Todos os resultados apresentados nesta secção assumem $N_c = 350$, N_a igual ao número total de possíveis análogos que compõem o *cluster* e $k = 5$, parâmetros definidos levando em consideração os resultados obtidos das seções anteriores.

Impacto dos Pesos P_1

As Tabelas 4.13 e 4.14 mostram os erros resultantes da previsão obtidos com (Peso = Sim) ou sem (Peso=Não) a primeira forma de ponderação da observações. Da Tabela 4.13 constam os erros nas previsões das variáveis ATMP, GST, PRES e WSPD com a variante K-means. Na Tabela 4.14 apresentam-se os erros obtidos com a variante C-means.

Na Tabela 4.13 nota-se que a variável PRES é a que apresenta maior diferença entre os testes feitos com e sem a utilização dos pesos P_1. Contudo, a diferença entre os erros não passa em média os 4%. Já para as variáveis ATMP, GST e WSPD, a utilização dos pesos P_1 faz apenas com que o valor do erro BIAS se torne mais negativo, indicando que as previsões feitas têm uma tendência ainda maior de estarem abaixo dos valores reais.

Tabela 4.13: Erros de previsão com/sem aplicação dos pesos P_1 na variante K-means

Peso	Variável	BIAS	RMSE	MAE	SDE
Sim	ATMP	-0,048	1,042	0,696	1,041
Não	ATMP	-0,037	1,04	0,695	1,039
Sim	GST	-0,282	1,556	1,173	1,530
Não	GST	-0,263	1,559	1,172	1,537
Sim	PRES	0,541	0,687	0,592	0,424
Não	PRES	0,561	0,712	0,618	0,439
Sim	WSPD	-0,087	1,359	1,022	1,357
Não	WSPD	-0,059	1,355	1,024	1,354

A Tabela 4.14 mostra a diferença entre a utilização ou não dos pesos P_1 com a variante C-means. Novamente, é para a variável PRES que se observa o maior impacto da utilização dos pesos. Verifica-se aí a diminuição dos erros BIAS, RMSE e MAE, em cerca de 7% em média, e a manutenção do SDE num valor semelhante ao obtido sem pesos.

Tabela 4.14: Erros de previsão com/sem aplicação dos pesos P_1 na variante C-means

Peso	Variável	BIAS	RMSE	MAE	SDE
Sim	ATMP	-0,047	1,035	0,686	1,034
Não	ATMP	-0,051	1,034	0,686	1,033
Sim	GST	-0,295	1,557	1,166	1,529
Não	GST	-0,260	1,558	1,166	1,536
Sim	PRES	0,587	0,716	0,641	0,409
Não	PRES	0,635	0,754	0,690	0,406
Sim	WSPD	-0,077	1,360	1,025	1,358
Não	WSPD	-0,062	1,369	1,031	1,368

Impacto dos Pesos P_2

O impacto da utilização dos pesos P_2 é mostrado na Tabela 4.15, onde se apresentam os erros obtidos na previsão das quatro variáveis meteorológicas utilizando a variante C-means. Como se pode observar, a variável PRES é a única que apresenta uma redução nos erros BIAS, RMSE e MAE em conjunto com a utilização do peso P_2. As variáveis GST e WSPD apresentam um aumento no valor dos quatro erros, enquanto que a variável ATMP apresenta uma variação mínima nos erros BIAS e MAE.

Tabela 4.15: Erros de previsão com/sem aplicação dos pesos P_2 na variante C-means

Peso	Variável	BIAS	RMSE	MAE	SDE
Sim	ATMP	-0,032	1,034	0,691	1,033
Não	ATMP	-0,051	1,034	0,686	1,033
Sim	GST	-0,335	1,584	1,182	1,548
Não	GST	-0,260	1,558	1,166	1,536
Sim	PRES	0,612	0,745	0,673	0,425
Não	PRES	0,635	0,754	0,690	0,406
Sim	WSPD	-0,097	1,385	1,038	1,382
Não	WSPD	-0,062	1,369	1,031	1,368

Síntese

A partir dos resultados das tabelas anteriores é possível afirmar que, em geral, os pesos aplicados sobre as observações que são usadas para os cálculos das previsões não influenciam de forma significativa os valores destas. Mas, tal como observado para outros parâmetros, a variável PRES é uma exceção; no seu caso, há uma diminuição dos erros com a aplicação dos dois tipos de pesos, em ambas as variantes K-means e C-means.

4.5 Estações Dependentes versus Independentes

Como explicado na seção 2.7, há duas formas de se utilizar as estações preditoras durante o processo de previsão: a forma dependente e a forma independente. Nesta secção faz-se uma análise comparativa desta duas formas de utilizar os dados das estações preditoras.

As estações utilizadas foram, novamente, a *Ykt* como sendo a estação que está sendo prevista, e as estações *Dom* e *Ykr* como estações preditoras. Também são usados os parâmetros já definidos nas secções anteriores; portanto, para as variantes K-means e C-means utilizou-se $N_c = 350$, $k = 5$, não foi especificado um valor para N_a e não foi utilizado nenhum peso; já para o método Monache, foi definido $N_a = 150$ e $k = 2$.

As Tabelas 4.16, 4.17 e 4.18 mostram os erros das previsões das variáveis meteorológicas (ATMP, GST, PRES e WSPD) feitas com os dados dependentes (Dependência = Sim) e independentes (Dependência = Não).

Variante Monache

A Tabela 4.16 mostra o impacto dessa escolha para o caso da variante Monache. Na tabela é possível ver que as variáveis GST e WSPD obtiveram melhores resultados com a utilização do método das estações independentes; já as variáveis ATMP e PRES tiveram os melhores resultados com o método de estações dependentes.

Tabela 4.16: Erros c/ estações preditoras dependentes e independentes - variante Monache

Dependência	Variável	BIAS	RMSE	MAE	SDE
Sim	ATMP	0,001	1,071	0,733	1,071
Não	ATMP	-0,208	1,437	0,799	1,422
Sim	GST	-0,53	2,197	1,685	2,132
Não	GST	-0,421	1,733	1,314	1,681
Sim	PRES	0,278	0,497	0,394	0,412
Não	PRES	0,579	0,853	0,721	0,626
Sim	WSPD	-0,206	2,075	1,578	2,064
Não	WSPD	-0,166	1,572	1,192	1,563

Variante K-means

Na Tabela 4.17 estão apresentados os resultados deste teste com a variante K-means. Diferentemente dos resultados obtidos para o método Monache, as quatro variáveis não apresentaram diferenças tão expressivas nos erros das previsões com e sem a dependência entre estações. Adicionalmente, exibem um comportamento comum: para todas as variáveis, o uso do método das estações dependentes produziu menores erros.

Tabela 4.17: Erros c/ estações preditoras dependentes e independentes - variante K-means

Dependência	Variável	BIAS	RMSE	MAE	SDE
Sim	ATMP	-0,037	1,04	0,695	1,039
Não	ATMP	-0,051	1,047	0,685	1,045
Sim	GST	-0,263	1,559	1,172	1,537
Não	GST	-0,278	1,681	1,290	1,658
Sim	PRES	0,561	0,712	0,618	0,439
Não	PRES	0,653	0,732	0,685	0,332
Sim	WSPD	-0,059	1,355	1,024	1,354
Não	WSPD	-0,086	1,558	1,193	1,555

Variante C-means

Para a variante C-means, os resultados apresentados na Tabela 4.18 mostram que apenas a variável PRES teve um resultado mais satisfatório sem a utilização da dependência entre estações (e, ainda assim, a melhora não foi expressiva). Já as restantes variáveis apresentaram melhores resultados quando se usou o método das estações dependentes.

Tabela 4.18: Erros c/ estações preditoras dependentes e independentes - variante C-means

Dependência	Variável	BIAS	RMSE	MAE	SDE
Sim	ATMP	-0,051	1,034	0,686	1,033
Não	ATMP	-0,063	1,053	0,682	1,052
Sim	GST	-0,260	1,558	1,166	1,536
Não	GST	-0,255	1,681	1,287	1,661
Sim	PRES	0,635	0,754	0,690	0,406
Não	PRES	0,651	0,737	0,693	0,345
Sim	WSPD	-0,062	1,369	1,031	1,368
Não	WSPD	-0,081	1,577	1,207	1,575

Síntese

Com base nas tabelas anteriores, pode-se concluir que especificar o método de estações que é utilizado causa um impacto relevante, tanto positivo quanto negativo, nos erros das previsões do método Monache. Os resultados que apresentaram uma melhoria no desempenho das previsões foram obtidos dos testes com as variáveis GST e WSPD. Essas duas variáveis são obtidas de medições feitas da velocidade do vento, e o comportamento de seus valores registados ao longo do tempo é similar e possuem grandes variações repentinas. O mesmo ocorre com as variáveis ATMP e PRES que, apesar de representarem entidades físicas bastante diferentes, possuem uma variação de seus valores similar entre elas e essa variação é mais gradativa ao longo do tempo, sem mudanças bruscas em um curto período de tempo. Essa diferença entre os comportamentos pode ser um dos motivos que levou os testes a obterem melhores e piores resultados com o método Monache.

Para a variante K-means não há outra conclusão senão a de que não é recomendada a utilização do método de estações independentes para esse conjunto de estações e variáveis. Na variação K-means não é possível visualizar o mesmo comportamento das variações dos

erros obtidos com o método Monache, levando a crer que o método é que está influenciando o comportamento dos erros obtidos e não o comportamento das variáveis em si.

A variante C-means novamente se assemelha à variante K-means; apesar de ter obtido um melhor resultado para o método de estações independentes com a variável PRES, o restante das variáveis manteve o comportamento de piorar os resultados. Sendo assim, o uso do método de estações dependentes é recomendado para as variantes K-means e C-means, levando em consideração esses conjuntos de estações e variáveis.

4.6 Previsão com Variáveis Diferentes da Prevista

Até agora, em todas as previsões feitas, os dados das estações preditoras utilizados para realizar a seleção dos análogos correspondem a dados da variável que está sendo prevista, ou seja, se a variável em foco é a variável ATMP, apenas dados dessa mesma variável são utilizados para realizar os processos de seleção dos análogos. No entanto, esse processo de seleção de análogos pode ser feito a partir de outras variáveis, mesmo que essas não sejam a variável em foco (que será prevista). Fazer essa alteração para o processo de seleção dos análogos não influencia o tipo dos dados que serão usados para realizar a previsão final, uma vez que as observações selecionadas precisam sempre de corresponder à variável que está sendo prevista. Um possível cenário para a utilização dessa abordagem, onde a variável preditora é diferente da variável prevista, é quando as estações que são utilizadas como preditoras não possuem registros históricos da variável que será prevista.

Os testes discutidos nesta secção mostram o desempenho das previsões quando essa troca de variáveis é feita em ambas as estações que estão sendo utilizadas como preditoras (*Dom e Ykr*). Os parâmetros considerados nesses testes foram: para as variantes K-means e C-means, $N_c = 350$, $k = 5$, nenhum valor foi especificado para N_a, nenhum peso para as observações foi especificado e foi utilizada a abordagem de estações dependentes; para o método Monache, $N_a = 150$, $k = 2$ e foi utilizada a abordagem de estações dependentes (apesar de se ter sido verificado que as variáveis GST e WSPD obtiveram melhores desempenhos com estações independentes) – essa abordagem de estações foi

especificada dessa forma para que a diferença entre os testes resida exclusivamente nas variáveis que estão sendo utilizadas para a seleção dos análogos.

Os resultados dos testes constam das Tabelas 4.19, 4.20, 4.21 e 4.22. Nestas, a variável que foi utilizada para a seleção dos análogos é indicada pela coluna "Variável Preditora" e os resultados também estão separados por método utilizado.

Previsão da variável ATMP

A Tabela 4.19 mostra os erros obtidos na previsão da variável ATMP variando o método e também a variável preditora. A partir desta tabela é possível ver que realizar a previsão da variável ATMP a partir de qualquer outra variável que não seja ela própria é inviável, pois os erros obtidos com as diferentes variáveis foram bastante elevados em comparação ao teste onde a variável ATMP foi utilizada como preditora.

Tabela 4.19: Erros da previsão da variável ATMP a partir de diversas variáveis preditoras

Variável Preditora	Método	BIAS	RMSE	MAE	SDE
ATMP	Monache	0,000	1,070	0,731	1,070
GST	Monache	-0,049	8,855	7,775	8,855
PRES	Monache	0,019	7,614	6,424	7,614
WSPD	Monache	-0,397	8,795	7,757	8,786
ATMP	K-means	-0,050	1,038	0,693	1,037
GST	K-means	-0,519	8,916	7,802	8,901
PRES	K-means	0,343	7,998	6,621	7,991
WSPD	K-means	-0,558	8,771	7,670	8,754
ATMP	C-means	-0,054	1,035	0,692	1,034
GST	C-means	-0,519	8,837	7,772	8,822
PRES	C-means	-0,279	7,776	6,604	7,771
WSPD	C-means	-0,570	8,743	7,668	8,724

Previsão da variável GST

A Tabela 4.20 mostra os erros das previsões feitas tendo como alvo a variável GST. E, novamente, os resultados mostram que realizar a previsão utilizando outras variáveis como base para a seleção dos análogos não produz bons resultados. Entretanto, também mostra que utilizar a variável WSPD como preditora para a previsão da variável GST não é mau de todo, principalmente quando utilizadas as variantes K-means e C-means, que

obtiveram resultados muito próximos em relação aos resultados obtidos com a utilização da própria GST como variável preditora.

Tabela 4.20: Erros da previsão da variável GST a partir de diversas variáveis preditoras

Variável Preditora	Método	BIAS	RMSE	MAE	SDE
ATMP	Monache	-0,208	3,042	2,378	3,035
GST	Monache	-0,530	2,197	1,685	2,132
PRES	Monache	-0,359	3,008	2,344	2,986
WSPD	Monache	-0,428	2,520	1,907	2,484
ATMP	K-means	-0,198	3,073	2,402	3,067
GST	K-means	-0,257	1,554	1,171	1,533
PRES	K-means	0,012	3,196	2,549	3,196
WSPD	K-means	-0,261	1,609	1,203	1,588
ATMP	C-means	-0,186	3,033	2,373	3,027
GST	C-means	-0,272	1,559	1,166	1,535
PRES	C-means	-0,205	3,048	2,405	3,041
WSPD	C-means	-0,259	1,622	1,207	1,602

Previsão da variável PRES

Os resultados relativos à previsão da variável PRES, presentes na Tabela 4.21, seguem um padrão similar ao da variável ATMP. Nenhuma outra variável, quando utilizada como variável preditora, se compara, na eficácia da previsão, aos resultado obtidos quando é utilizado a própria PRES como variável base da previsão.

Tabela 4.21: Erros da previsão da variável PRES a partir de diversas variáveis preditoras

Variável Preditora	Método	BIAS	RMSE	MAE	SDE
ATMP	Monache	-0,364	6,599	5,154	6,589
GST	Monache	-0,276	6,629	5,177	6,623
PRES	Monache	0,278	0,497	0,394	0,412
WSPD	Monache	-0,074	6,503	5,085	6,503
ATMP	K-means	-0,642	6,651	5,216	6,620
GST	K-means	-0,162	6,590	5,169	6,588
PRES	K-means	0,564	0,713	0,619	0,435
WSPD	K-means	-0,210	6,518	5,085	6,515
ATMP	C-means	-0,657	6,660	5,212	6,628
GST	C-means	-0,215	6,581	5,145	6,578
PRES	C-means	0,615	0,745	0,674	0,421
WSPD	C-means	-0,234	6,505	5,095	6,501

Previsão da variável WSPD

Nas previsões da variável WSPD, o cenário é bem diferente em relação à capacidade preditora das outras variáveis. Assim, a Tabela 4.22, mostra que realizar a previsão da variável WSPD utilizando a variável GST como preditora alcança melhores resultados em comparação com a utilização das outras variáveis, incluindo a própria WSPD. Comparando os cenários onde a variável GST e WSPD são usadas como variável preditora, é possível ver que os erros aleatórios são menores nas previsões feitas com a variável GST como preditora, indicando uma maior precisão das previsões; no entanto, o erro sistemático BIAS tem um pequeno aumento, indicando que as previsões feitas, apesar de serem mais precisas, tendem a obter resultado mais inferiores à verdade.

Tabela 4.22: Erros da previsão da variável WSPD a partir de diversas variáveis preditoras

Variável Preditora	Método	BIAS	RMSE	MAE	SDE
ATMP	Monache	-0,025	2,436	1,901	2,436
GST	Monache	-0,302	1,817	1,400	1,792
PRES	Monache	-0,147	2,410	1,876	2,405
WSPD	Monache	-0,206	2,075	1,578	2,064
ATMP	K-means	-0,015	2,443	1,901	2,443
GST	K-means	-0,070	1,327	1,005	1,325
PRES	K-means	0,222	2,587	2,067	2,577
WSPD	K-means	-0,059	1,359	1,024	1,358
ATMP	C-means	-0,007	2,435	1,911	2,435
GST	C-means	-0,077	1,321	0,994	1,319
PRES	C-means	0,007	2,467	1,945	2,467
WSPD	C-means	-0,055	1,365	1,028	1,364

4.7 Eficiência Computacional

Após a identificação dos melhores parâmetros para cada cenário estudado, realizou-se uma análise do desempenho computacional para determinar as formas mais eficientes de se obter bons resultados. Esta análise baseia-se na medição do tempo que cada método demora para ser executado em diferentes cenários de disponibilidade de recursos computacionais, dentro dos limites impostos pela configuração do sistema descrito na secção

3.2.4. Assim, apresentam-se tempos para uma execução sequencial (1 só núcleo) e para uma execução paralela (com o número de núcleos dobrando até ao máximo de 64). A execução paralela, recorde-se (secção 3.2.2), ocorre apenas em certas fases do processo de previsão, baseando-se na função *parSapply* da plataforma R. Apresentam-se ainda medidas de Speedup (Aceleração) e Eficiência; a primeira é dada por $S_n = T_1 \ / \ T_n$, sendo T_1 e T_n o tempo de execução com 1 só núcleo e com n núcleos, respetivamente; a segunda é dada por S_n/n, medindo quão eficientemente se estão a usar os n núcleos de CPU.

Na secção 3.2.2, elaborou-se sobre a fase de importação dos dados, onde há a possibilidade de ser feito apenas um carregamento rápido de um ficheiro já previamente organizado ou há também a possibilidade de ser necessário fazer a organização dos dados. Os resultados apresentados a seguir foram obtidos no contexto da primeira possibilidade.

Como já referido, os parâmetros utilizados nas execuções dos testes seguintes são os identificados como melhores, sob o ponto de vista de minimização dos erros, ao longo do estudo feito. Portanto, para o método Monache foi utilizado o valor de $N_a = 150$ e $k = 2$. Para as variantes K-means e C-means usou-se $N_c = 350$, N_a não é especificado (usa-se todos os elementos do melhor *cluster*), $k = 5$, e não se especifica nenhum tipo de peso para as observações. Adicionalmente, foi utilizada a abordagem de estações dependentes.

Previsão da variável PRES

O gráfico da Figura 4.15 e a tabela 4.23 apresentam o resultado do estudo de escalabilidade da previsão da variável PRES, recorrendo a diferentes métodos e a diferente número de núcleos de CPU envolvidos.

O gráfico mostra que a quantidade de núcleos usada influencia bastante nas execuções feitas com o método Monache; já para as variantes K-means e C-means não é visível uma grande variação nos tempos de execução quando se varia a quantidade de núcleos. Diferentemente do método de Monache, essas variantes possuem a fase de clusterização dos P_a possíveis análogos e esse processo não é feito de forma paralelizada; apenas o processo dos cálculos das previsões é feita de forma paralela e esse é o processo que demanda menos

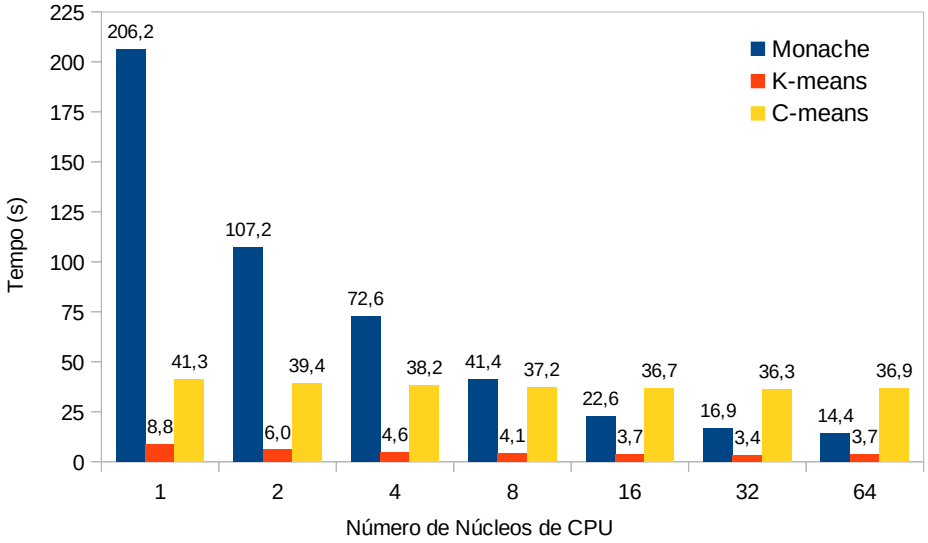


Figura 4.15: Tempos (s) da previsão da variável PRES em função do número de núcleos de CPU usados

		Número de Núcleos de CPU						
		1	2	4	8	16	32	64
Monache	**Speedup**	1,00	1,92	2,84	4,98	9,12	12,20	14,32
	Eficiência	100,0%	96,2%	71,0%	62,3%	57,0%	38,1%	22,4%
K-means	**Speedup**	1,00	1,47	1,91	2,15	2,38	2,59	2,38
	Eficiência	100,0%	73,3%	47,8%	26,8%	14,9%	8,1%	3,7%
C-means	**Speedup**	1,00	1,05	1,08	1,11	1,13	1,14	1,12
	Eficiência	100,00%	52,41%	27,03%	13,88%	7,03%	3,56%	1,75%

Tabela 4.23: Speedup e Eficiência da previsão da variável PRES em função do número de núcleos de CPU usados

tempo nas variantes K-means e C-means.

No método Monache, a maior parte do tempo é gasto na fase das previsões, que envolvem um elevado número de cálculos, pois o preditor é comparado com todos os possíveis análogos. Como estas comparações são em grande número e são feitas de forma paralelizada, é possível notar uma grande melhoria nos tempos do método Monache com o aumento de núcleos utilizados. Nos métodos de clusterização, o preditor é comparado apenas com os centroides dos clusters formados, pelo que a paralelização deste processo exibe menos benefícios; além disso, a maior parte do tempo é gasto no processo de clusterização; dessa forma, o impacto da paralelização, embora visível e semelhante nos dois

métodos (K-means e C-means), acaba por ser modesto.

Apesar do grande aumento de desempenho do método Monache com o aumento de núcleos utilizados, tal não é rápido o suficiente para ultrapassar os tempos obtidos com a variante K-means. No entanto, a redução do tempo de sua execução é suficiente para ultrapassar os tempos da variante C-means que possui tempos superiores ao método Monache a partir da utilização de 16 núcleos. Para um mesmo conjunto de entradas, os tempos de execução da variante K-means sempre serão mais rápidos do que os tempos da variante C-means, pelo fato do processo de clusterização desta ser sempre mais demorado.

A Tabela 4.23 mostra mais uma vez que o método de Monache é o que tira melhor partido da paralelização, com o speedup a chegar a 14,32, embora isso fique muito aquém dos 64 núcleos usados nesse cenário, como traduzido pela fraca eficiência, que nesse caso é apenas de 22,3%; para os restantes métodos, as acelerações são muito mais pequenas ou quase inexistentes, com os consequentes reflexos em eficiências ainda mais baixas.

A Tabela 4.24 mostra três conjuntos de erros retirados das execuções que foram utilizadas para fazer a medição do tempo de cada método. Na tabela é visível que os erros obtidos para o método Monache são inferiores aos erros obtidos das variantes K-means e C-means. De notar que os erros não são apresentados em função do número de núcleos usados, pois mantêm-se muito semelhantes quando se varia esse número.

Tabela 4.24: Erros na previsão da variável v=PRES

	Erros individuais (e_j)				Erros combinados
Métodos (m_i)	BIAS	RMSE	MAE	SDE	$\sum_j \lvert e_j \rvert$ sem BIAS
Monache	0,278	0,497	0,394	0,412	**1,303**
K-means	0,555	0,702	0,608	0,431	1,741
C-means	0,611	0,734	0,661	0,407	1,802

Dependendo da importância relativa da minimização do tempo das previsões, e da minimização dos erros, é possível definir uma métrica linear simples, que auxilie na escolha do método de previsão que, simultaneamente, satisfaz estes dois objetivos, cada um com um grau de importância diferente. Assim, seja α a importância dada ao tempo de execução e $\beta = 1-\alpha$ a relevância dado à minimização do erro, com $\alpha \in [0, 1]$. A métrica pretendida,

que se optou designar por *ranking*, é dada simplesmente por *ranking* $= \alpha\times$ (erro combinado normalizado) $+ \beta\times$ (menor tempo normalizado), em que a normalização se faz por divisão pelo menor erro combinado, ou pelo menor tempo de execução, conforme o caso.

Tabela 4.25: *Rankings* dos métodos de previsão da variável PRES

Métodos	erro combinado normalizado	menor tempo normalizado	número de núcleos (n)	*ranking* c/ $\alpha = 0.75$	*ranking* c/ $\alpha = 0.5$	*ranking* c/ $\alpha = 0.25$
Monache	1	4,235	64	1,809	2,618	3,426
K-means	1,336	1	32	**1,252**	**1,168**	**1,084**
C-means	1,383	10,676	32	3,706	6,030	8,353

A tabela 4.25 mostra os *rankings* para diferentes valores de α (0.75, 0.5 e 0.25), que traduzem uma importância decrescente da minimização do tempo de execução em detrimento da minimização do erro. Naturalmente, se a minimização do tempo de execução ou do erro forem critérios a considerar de forma isolada, o gráfico 4.15 e a tabela 4.24 são suficientes para a escolha do método de previsão e do número de núcleos de CPU a usar.

Com base no gráfico e tabelas anteriores, pode-se então concluir que para a previsão da variável PRES: i) se interessa apenas minimizar os erros ($\alpha = 1$), então deve-se optar pelo método Monache; mas, adicionalmente, este deve ser executado com 64 núcleos, pois isso minimizará também o tempo de execução; ii) se só interessa minimizar o tempo de execução ($\alpha = 0$), então deve-se optar pelo método K-means com 32 núcleos; iii) se minimizar os erros é mais importante que minimizar o tempo de execução ($\alpha = 0.75$), então novamente deve-se optar pelo método K-means; iv) se minimizar o erro tem igual importância que minimizar o tempo de execução ($\alpha = 0.5$), a melhor opção é novamente executar a previsão com o método K-means; v) finalmente, se minimizar o erro tem menos importância que minimizar o tempo de execução ($\alpha = 0.25$), o melhor continua sendo prever com o método K-means. As melhores opções para as situações iii), iv) e v) foram escolhidas com base nos menores valores (a negrito) dos rankings na tabela 4.25, pressupondo assim sempre o uso de 32 núcleos.

Com a Tabela 4.24 e com o gráfico da Figura 4.15 é possível concluir que, se se optar por usar a variante K-means (por ter o menor tempo de execução) ao invés do método

Monache é possível ter uma redução no tempo de execução de aproximadamente 74%, utilizando 64 núcleos. Por outro lado, a combinação dos erros (ver tabela 4.24) aumenta em aproximadamente 33%. Esse comportamento de ganho de tempo e perda de precisão só ocorre para a variável PRES. Assim como já vem sendo observado, as variáveis ATMP, GST e WSPD possuem uma maior precisão quando são previstas utilizado as variantes de clusterização ao invés do método Monache.

Previsão das demais variáveis

Para as demais variáveis analisadas (ATMP, GST e WSPD), há grande semelhança entre elas no que diz respeito aos tempos de execução, *speedup*, eficiência e *ranking* dos métodos de previsão. Por conta disso, a análise destes resultados é realizada nesta secção apenas para a variável ATMP, sendo que as considerações feitas podem ser aplicadas também às variáveis GST e WSPD, cujos gráficos e tabelas constam do apêndice C.3.

O gráfico da Figura 4.16 e a tabela 4.26 apresentam o resultado do estudo de escalabilidade da previsão da variável ATMP. É possível ver que a mesma análise feita anteriormente para a variável PRES pode ser reconsiderada: aumentar a quantidade de núcleos de CPU utilizados faz com que as execuções do algoritmo de previsão, com o uso do método Monache, tenham uma alta redução no tempo de execução, ultrapassando os tempos obtidos com o método C-means mas não os tempos da variante K-means.

		Número de Núcleos de CPU						
		1	2	4	8	16	32	64
Monache	**Speedup**	1,00	1,85	3,94	5,27	9,67	12,06	13,58
	Eficiência	100,0%	92,7%	98,4%	65,9%	60,5%	37,7%	21,2%
K-means	**Speedup**	1,00	1,51	1,93	2,12	2,62	2,70	2,41
	Eficiência	100,0%	75,4%	48,4%	26,5%	16,4%	8,4%	3,8%
C-means	**Speedup**	1,00	1,03	1,12	1,11	1,15	1,12	1,14
	Eficiência	100,00%	51,35%	27,94%	13,93%	7,20%	3,50%	1,77%

Tabela 4.26: Speedup e Eficiência da previsão da variável ATMP em função do número de núcleos de CPU usados

Já os valores apresentados na Tabela 4.27 mostram que a variante K-means obteve os

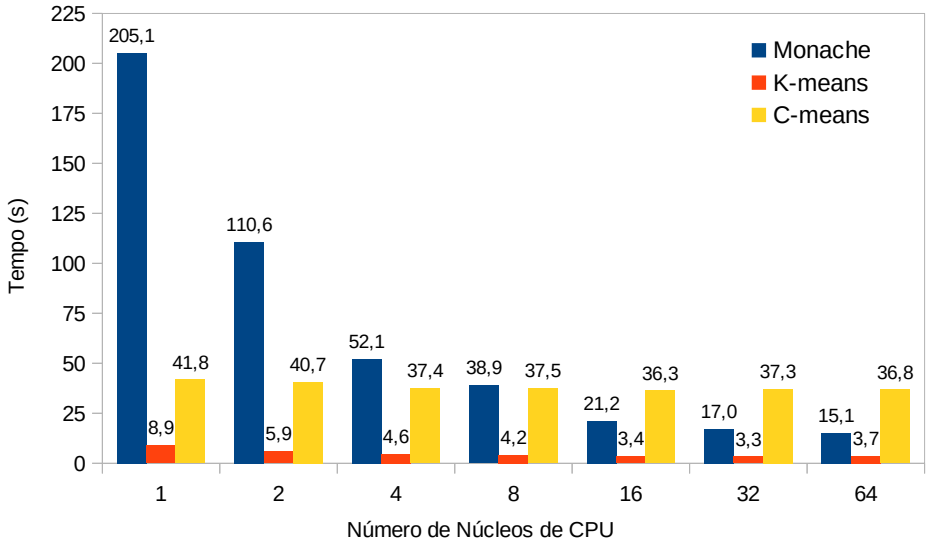


Figura 4.16: Tempos (s) da previsão da variável ATMP em função do número de núcleos de CPU usados

menores erros na previsão da variável ATMP. Além disso, verifica-se também que os erros obtidos para a variante C-means foram muito próximos aos erros obtidos com a variante K-means, ao passo que o método Monache passa a ser o mais impreciso dos três.

Dessa vez, é possível ver que para além da redução no tempo de execução também há uma redução nos erros obtidos por parte da variante K-means, fazendo com que essa variante passe a ser a melhor opções em relação ao método Monache, não só por ser mais rápida como também por ser mais precisa. Desta forma, uma tabela com os *rankings* é neste caso dispensável, pois é clara a prevalência do método K-means em todas as situações, independentemente da importância conferida ao tempo de execução e aos erros.

Tabela 4.27: Erros na previsão da variável v=ATMP

	Erros individuais (e_j)				Erros combinados
Métodos (m_i)	BIAS	RMSE	MAE	SDE	$\sum_j \lvert e_j \rvert$ sem BIAS
Monache	0,000	1,070	0,731	1,070	2,871
K-means	-0,049	1,028	0,686	1,027	**2,741**
C-means	-0,039	1,031	0,688	1,030	2,749

Síntese

Considerando as observações anteriores, é então possível concluir que:

- Na previsão da variável PRES, a minimização dos erros é conseguida pelo método Monache, o qual deve executar-se, se possível, com todos os núcleos de CPU disponíveis; para a mesma variável, se minimizar o tempo de previsão for prioritário (prejudicando os erros, ainda que de forma modesta), deve-se usar o método K-means, preferencialmente com oito núcleos (a partir daqui os ganhos são mínimos).
- Para a previsão das restantes variáveis consideradas neste estudo, o método K-means garante simultaneamente a minimização dos erros e dos tempos de previsão, sendo que para este último objetivo é suficiente usar oito núcleos de CPU.
- Em geral, pode-se dizer que o método K-means, com oito núcleos de CPU, acaba por representar o melhor compromisso para as quatro variáveis. Esta conclusão é particularmente interessante porquanto hoje em dia cada vez mais sistemas computacionais vêm já providos da capacidade de executar 8 fios de execução (seja porque possuem pelo menos 8 núcleos de execução, seja com recurso a *hyperthreading*).

Capítulo 5

Conclusão

O estudo realizado utilizou a metodologia dos Conjuntos Análogos para realizar a previsão de séries temporais. O método foi originalmente usado para a realização de pós-processamento e no decorrer de sua evolução passou a ser utilizado também para a previsão e *hindcasting* de dados meteorológicos.

O trabalho procurou determinar os valores dos parâmetros que devem ser utilizados em diferentes variações do método dos Conjuntos Análogos, para que haja a minimização dos erros das previsões realizadas. Foram analisadas três variações do método, denominadas Monache, K-means e C-means.

Dentre os parâmetros analisados estão o número de *clusters* formados nas variantes K-means e C-means, o número de análogos selecionados durante as previsões, o tamanho da janela de dados que indica a janela de tempo que um análogo abrange, pesos de relevância atribuídos as observações selecionadas nas variantes K-means e C-means, o impacto da opção pela abordagem de estações meteorológicas dependentes vs independentes, e da utilização de variáveis meteorológicas diferentes das que estão sendo previstas para realizar a seleção dos conjuntos de dados análogos durante o processo de previsão. Além disso, uma análise de desempenho foi feita comparando as diferentes variantes utilizadas.

Na análise feita com o número de *clusters* que deve ser utilizado nas variantes K-means e C-means, ficou claro que o valor ótimo desse número é difícil de determinar, pois varia dependendo da variável que está sendo trabalhada. Entretanto, também foi possível

concluir que utilizar um número de *clusters* inferior a 100 não era recomendado.

O valor ideal para a quantidade de análogos depende da variante em uso. Na variante Monache, apesar do valor ideal para esta quantidade também depender da variável trabalhada, foi possível chegar a um valor médio de 150 que atendesse a todas as variáveis, sem uma grande perda da precisão das previsões. Já nas variantes K-means e C-means há a possibilidade do valor do número de análogos não ser especificado e de se utilizar todos os possíveis análogos incluídos no *cluster* selecionado. Essa é justamente a escolha recomendada para essas variantes, que utilizam clusterização no seu processo de previsão.

O parâmetro que determina o tamanho das janela de dados que serão utilizados pôde ser facilmente identificado como tendo um valor ótimo de 5 para as variantes K-means e C-means. Já na variante Monache, o valor ideal tem que ser identificado para cada variável, mas pode ser facilmente encontrado, pois segue um padrão onde o valor ideal tem elevada probabilidade de estar entre 1 e 10.

Os pesos aplicados às observações utilizadas nos processos de previsão das variantes K-means e C-means não introduziram, no geral, melhorias relevantes na previsão e apenas com a variável PRES foi possível identificar um leve aumento na precisão da previsão. Para as outras variáveis analisadas, a diferença entre os erros dos testes onde os pesos foram aplicados e não aplicados é irrelevante.

A opção pela abordagem de estações dependentes ou, em alternativa, de estações independentes, demonstrou ter grande impacto na variante Monache. A abordagem deve ver escolhida tendo em conta a variável que esta sendo trabalhada, pois a precisão das previsões feitas podem aumentar ou diminuir dependendo da variável em foco. Para as variantes K-means e C-means concluiu-se que melhores resultados são obtidos levando em consideração a abordagem de estações dependentes.

As variantes estudadas permitem que parte do processo de previsão seja feito utilizando variáveis diferentes da variável em foco. Tal possibilita que estações que não possuem dados registados da variável que está sendo prevista, possam usar-se no processo de previsão. Após analisar este processo com diferentes combinações de variáveis, concluiu-se que o mais recomendado é fazer as previsões a partir de dados da própria variável que está

sendo prevista. No entanto, houve uma exceção: as variáveis GST e WSPD obtiveram resultados satisfatórios utilizando uma combinação de diferentes variáveis.

Os testes relacionados com a análise da eficiência computacional (limitados a sistemas de memória partilhada) demonstraram ser vantajosa a execução do código da previsão em sistemas com vários núcleos de execução, embora praticamente sem vantagens com mais de oito núcleos. Ficou também claro que é possível conjugar a aceleração da previsão, com a minimização dos erros, escolhendo criteriosamente o método de previsão sendo que, para a maioria das situações, a variante K-means impôs-se como a escolha mais acertada.

5.1 Trabalhos Futuros

O estudo realizado focou-se em determinar os melhores parâmetros para um conjunto específico de estações e variáveis. Portanto, ainda é necessário validar se os mesmos padrões e resultados são observados em mais diferentes combinações desses dois parâmetros. Além disso, todos os testes foram realizados levando em consideração duas estações preditoras e apenas uma variável para essas estações, ou seja, é necessário validar os resultados com o uso de múltiplas variáveis e com o uso de apenas uma estação preditora ou mais de duas.

Os diversos testes realizados demonstraram que os resultados obtidos a partir das variantes K-means e C-means foram bons (em termos de erros) e foram obtidos de forma rápida. Por esses motivos vale apena realizar uma maior exploração e comparações de desempenho entre diferentes métodos de clusterização. Adicionalmente, impõe-se considerar a comparação das diferentes abordagens do método de Conjuntos Análogos com técnicas de *machine learning*, que apresentam uma maior complexidade em seus algoritmos em comparação ao método de Conjuntos Análogos, mas que vêm apresentando bons resultados nos mais variados cenários.

Bibliografia

[1] V. L. MACHADO, «Análise do impacto da previsão do tempo corrigida pelo método Model Output Calibration em modelos de doenças da maça.», Instituto de Ciências Exatas e Geociências – ICEG, rel. téc., 2017. URL: http://10.0.217.128:8080/jspui/handle/tede/32.

[2] G. Sampaio e P. L. d. S. Dias, «Evolução dos Modelos Climáticos e de Previsão de Tempo e Clima.», *Revista USP*, n.º 103, pp. 41–54, nov. de 2014. DOI: 10.11606/issn.2316-9036.v0i103p41-54. URL: http://www.revistas.usp.br/revusp/article/view/99179.

[3] C. Abbe, «THE PHYSICAL BASIS OF LONG-RANGE WEATHER FORECASTS», *Monthly Weather Review*, vol. 29, n.º 12, pp. 551–561, dez. de 1901. DOI: 10.1175/1520-0493(1901)29[551c:TPBOLW]2.0.CO;2. URL: https://journals.ametsoc.org/view/journals/mwre/29/12/1520-0493_1901_29_551c_tpbolw_2_0_co_2.xml.

[4] V. Bjerknes, «The problem of weather prediction, considered from the viewpoints of mechanics and physics», *Meteorologische Zeitschrift*, vol. 18, n.º 6, pp. 663–667, dez. de 2009. DOI: 10.1127/0941-2948/2009/416. URL: http://dx.doi.org/10.1127/0941-2948/2009/416.

[5] P. Lynch, «The origins of computer weather prediction and climate modeling», *Journal of Computational Physics*, vol. 227, n.º 7, pp. 3431–3444, 2008, Predicting weather, climate and extreme events, ISSN: 0021-9991. DOI: https://doi.org/10.

1016/j.jcp.2007.02.034. URL: http://www.sciencedirect.com/science/article/pii/S0021999107000952.

[6] T. G. OLIVEIRA G. S.; FLORENZANO, «Satélites e o meio ambiente.», vol. 6, pp. 1–8, mai. de 2006. URL: http://mtc-m12.sid.inpe.br/col/sid.inpe.br/mtc-m12@80/2006/08.04.16.53/doc/Boletim_Da_Terra_ao_Espaco_20-03-2006_PGM_4.pdf.

[7] N. A. PHILLIPS, «Energy Transformations and Meridional Circulations associated with simple Baroclinic Waves in a two-level, Quasi-geostrophic Model1», *Tellus*, vol. 6, n.º 3, pp. 273–286, 1954. DOI: https://doi.org/10.1111/j.2153-3490.1954.tb01123.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.2153-3490.1954.tb01123.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.2153-3490.1954.tb01123.x.

[8] E. S. Epstein, «Stochastic dynamic prediction», *Tellus*, vol. 21, n.º 6, pp. 739–759, 1969. DOI: 10.3402/tellusa.v21i6.10143. eprint: https://doi.org/10.3402/tellusa.v21i6.10143. URL: https://doi.org/10.3402/tellusa.v21i6.10143.

[9] C. E. Leith, «Theoretical Skill of Monte Carlo Forecasts», *Monthly Weather Review*, vol. 102, n.º 6, pp. 409–418, jun. de 1974. DOI: 10.1175/1520-0493(1974)102<0409:TSOMCF>2.0.CO;2. URL: https://journals.ametsoc.org/view/journals/mwre/102/6/1520-0493_1974_102_0409_tsomcf_2_0_co_2.xml.

[10] E. N. Lorenz, «Atmospheric Predictability as Revealed by Naturally Occurring Analogues», *Journal of the Atmospheric Sciences*, vol. 26, n.º 4, pp. 636–646, 1969. DOI: 10.1175/1520-0469(1969)26<636:aparbn>2.0.co;2.

[11] K. Ruosteenoja, «Factors Affecting the Occurrence and Lifetime of 500 mb Height Analogues: A Study Based on a Large Amount of Data», *Monthly Weather Review*, vol. 116, n.º 2, pp. 368–376, fev. de 1988. DOI: 10.1175/1520-0493(1988)116<0368:FATOAL>2.0.CO;2. URL: https://journals.ametsoc.org/view/journals/mwre/116/2/1520-0493_1988_116_0368_fatoal_2_0_co_2.xml.

[12] H. M. V. D. Dool, «A New Look at Weather Forecasting through Analogues», *Monthly Weather Review*, vol. 117, n.º 10, pp. 2230–2247, 1989. DOI: 10.1175/1520-0493(1989)117<2230:anlawf>2.0.co;2.

[13] L. D. Monache, T. Nipen, Y. Liu, G. Roux e R. Stull, «Kalman Filter and Analog Schemes to Postprocess Numerical Weather Predictions», *Monthly Weather Review*, vol. 139, n.º 11, pp. 3554–3570, 2011. DOI: 10.1175/2011mwr3653.1.

[14] R. E. Kalman, «A New Approach to Linear Filtering and Prediction Problems», *Journal of Basic Engineering*, vol. 82, n.º 1, pp. 35–45, mar. de 1960, ISSN: 0021-9223. DOI: 10.1115/1.3662552. eprint: https://asmedigitalcollection.asme.org/fluidsengineering/article-pdf/82/1/35/5518977/35_1.pdf. URL: https://doi.org/10.1115/1.3662552.

[15] L. D. Monache, F. A. Eckel, D. L. Rife, B. Nagarajan e K. Searight, «Probabilistic Weather Prediction with an Analog Ensemble», *Monthly Weather Review*, vol. 141, n.º 10, pp. 3498–3516, 2013. DOI: 10.1175/mwr-d-12-00281.1.

[16] E. N. Lorenz, «Deterministic Nonperiodic Flow», *Journal of Atmospheric Sciences*, vol. 20, n.º 2, pp. 130–141, mar. de 1963. DOI: 10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2. URL: https://journals.ametsoc.org/view/journals/atsc/20/2/1520-0469_1963_020_0130_dnf_2_0_co_2.xml.

[17] S. Alessandrini, L. D. Monache, S. Sperati e J. N. Nissen, «A novel application of an analog ensemble for short-term wind power forecasting», *Renewable Energy*, vol. 76, pp. 768–781, 2015. DOI: 10.1016/j.renene.2014.11.061.

[18] S. Alessandrini, L. D. Monache, S. Sperati e G. Cervone, «An analog ensemble for short-term probabilistic solar power forecast», *Applied Energy*, vol. 157, pp. 95–110, 2015. DOI: 10.1016/j.apenergy.2015.08.011.

[19] G. Cervone, L. Clemente-Harding, S. Alessandrini e L. Delle Monache, «Short-term photovoltaic power forecasting using Artificial Neural Networks and an Analog Ensemble», *Renewable Energy*, vol. 108, pp. 274–286, 2017, ISSN: 0960-1481. DOI:

https://doi.org/10.1016/j.renene.2017.02.052. URL: https://www.sciencedirect.com/science/article/pii/S0960148117301386.

[20] E. Wu, M. Z. Zapata, L. Delle Monache e J. Kleissl, «Observation-Based Analog Ensemble Solar Forecast in Coastal California», em *2019 IEEE 46th Photovoltaic Specialists Conference (PVSC)*, 2019, pp. 2440–2444. DOI: 10.1109/PVSC40753.2019.8980546.

[21] S. Alessandrini, S. Sperati e L. D. Monache, «Improving the Analog Ensemble Wind Speed Forecasts for Rare Events», *Monthly Weather Review*, vol. 147, n.º 7, pp. 2677–2692, jul. de 2019. DOI: 10.1175/MWR-D-19-0006.1. URL: https://journals.ametsoc.org/view/journals/mwre/147/7/mwr-d-19-0006.1.xml.

[22] M. Shahriari, G. Cervone, L. Clemente-Harding e L. Delle Monache, «Using the analog ensemble method as a proxy measurement for wind power predictability», *Renewable Energy*, vol. 146, pp. 789–801, 2020, ISSN: 0960-1481. DOI: https://doi.org/10.1016/j.renene.2019.06.132. URL: https://www.sciencedirect.com/science/article/pii/S0960148119309668.

[23] C. Balsa, C. V. Rodrigues, I. Lopes e J. Rufino, «Using Analog Ensembles with Alternative Metrics for Hindcasting with Multistations», *ParadigmPlus*, vol. 1, n.º 2, pp. 1–17, jun. de 2020. URL: https://journals.itiud.org/index.php/paradigmplus/article/view/11.

[24] S. Ghosh e S. Kumar, «Comparative Analysis of K-Means and Fuzzy C-Means Algorithms», *International Journal of Advanced Computer Science and Applications*, vol. 4, mai. de 2013. DOI: 10.14569/IJACSA.2013.040406.

[25] S. Panda, S. Sahu, P. Jena e S. Chattopadhyay, «Comparing Fuzzy-C Means and K-Means Clustering Techniques: A Comprehensive Study», em *Advances in Computer Science, Engineering & Applications*, D. C. Wyld, J. Zizka e D. Nagamalai, eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 451–460, ISBN: 978-3-642-30157-5.

[26] T. Chai e R. R. Draxler, «Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature», *Geoscientific Model Development*, vol. 7, n.º 3, pp. 1247–1250, 2014. DOI: 10.5194/gmd-7-1247-2014.

[27] A. Chesneau, C. Balsa, C. V. Rodrigues e I. M. Lopes, «Hindcasting with multistations using analog ensembles.», em *CEUR Workshop Proceedings*, 2019, pp. 215–229.

[28] K. E. Taylor, «Summarizing multiple aspects of model performance in a single diagram», *Journal of Geophysical Research: Atmospheres*, vol. 106, n.º D7, pp. 7183–7192, 2001. DOI: 10.1029/2000jd900719.

[29] *National Data Buoy Center*, https://www.ndbc.noaa.gov/.

[30] R. Rew e G. Davis, «NetCDF: an interface for scientific data access», *IEEE Computer Graphics and Applications*, vol. 10, n.º 4, pp. 76–82, 1990. DOI: 10.1109/38.56302.

[31] M. M. d. Santos, «Reconstitution of weather time series with an analog ensemble model. Master's thesis in Industrial Engineering.», tese de mestrado, Instituto Politécnico de Bragança, 2019. URL: http://hdl.handle.net/10198/19762.

Apêndice A

Proposta Original da Dissertação

INSTITUTO POLITÉCNICO DE BRAGANÇA
Escola Superior de Tecnologia e Gestão

Mestrado em Sistemas de Informação

Proposta de Trabalho

Ano letivo de 2019/2020

Título: **Implementação eficiente em R do método dos conjuntos análogos para reconstrução de séries temporais meteorológicas**

Tipo de Trabalho	Dissertação
Orientador:	Carlos Balsa, José Rufino
Coorientador(es)	Helyane Bronoski Borges
Data da proposta	03/2020
Observações	Proposta em desenvolvimento pelo aluno Leonardo Oliveira Guth de Araujo (a46677)

Objetivo

O objetivo principal do trabalho é realizar uma implementação na linguagem R, eficiente e de elevado desempenho computacional, do algoritmo dos Conjuntos Análogos, utilizado em previsões meteorológicas. Tipicamente, este algoritmo lida com uma grande quantidade de dados históricos, tendo elevados requisitos de memória RAM e de capacidade de processamento. Desta forma, pretende-se tirar partido das facilidades de processamento paralelo disponíveis no R, a fim de responder, de forma eficiente, aos desafios computacionais colocados pelo algoritmo em foco.

Descrição

Existem vários métodos para reconstruir ou prever séries temporais que utilizam técnicas de cariz estatístico aplicadas a um histórico de observações passadas. Neste âmbito, o estado da arte inclui redes neuronais artificiais e um conjunto de técnicas denominadas de Analog Ensembles (AnEn) ou Conjuntos Análogos. Estes últimos são de simples implementação e foram inicialmente concebidos para estimar a incerteza de modelos de previsão meteorológicos.

O algoritmo atual (implementação do AnEn) necessita de processar séries temporais multivariáveis de grandes dimensões, resultantes do registo por estações meteorológicas de variáveis climáticas ao longo de muitos anos. O processamento em tempo útil de tão grande quantidade de dados obriga a uma

eficiente implementação do algoritmo. Esta implementação deverá ser capaz de distribuir as operações a realizar (e a memória associada) pelos vários núcleos de processamento disponíveis na arquitetura usada, de forma a acelerar a velocidade de execução.

Para o efeito será utilizado o software livre R e respetiva linguagem de programação de alto nível, a qual inclui uma grande variedade de funções especialmente concebidas para o tratamento e análise de dados. Possui também facilidades destinadas à melhoria da eficiência computacional como, por exemplo, funções de aceleração dos cálculos para dados organizados em matrizes e vetores, e funções de paralelização automática do processamento.

A implementação realizada será testada com conjuntos de dados reais provenientes de estações meteorológicas, com o objetivo de reconstruir os valores de uma estação utilizando os valores de estações geograficamente próximas (vizinhas).

Metodologia de trabalho / Cronograma de Atividades

- Pesquisa bibliográfica e familiarização com a linguagem R e a estrutura de dados netcdf, utilizada em meteorologia. (1 mês)

- Melhoramento da eficiência do algoritmo dos conjuntos análogos sequencial através da vetorização dos cálculos que mais tempo consomem como, por exemplo, os ciclos. (1 mês)

- Estudo e implementação das técnicas de computação paralelas/distribuídas disponíveis na software R. (2 mês)

- Testes de eficiência computacional com vista a melhorar a eficiência do algoritmo na realização de previsões meteorológicas. Para tal serão testados os efeitos da variação da dimensão do conjunto dos análogos, da janela de tempo do análogo, etc. (2 mês)

- Aplicações do algoritmo no contexto meteorológico para previsões ou reconstituição de períodos de tempo em que não haja informações disponíveis. (2 mês)

- Em simultâneo com as etapas anteriores, será redigida a dissertação de mestrado, sendo contudo reservado o último mês em exclusivo para esta tarefa. (1 mês)

Pré-requisitos

- Domínio de linguagens de programação de alto nível (desejavelmente Python ou R)

- Familiarização com ambientes de trabalho (GUI e CLI) em Linux

Infraestruturas e Recursos Necessários

- Posto de trabalho individual (fornecido, se necessário, pelo Lab. de Computação Avançada)

- Cluster HPC do IPB (adstrito ao Lab. de Computação Avançada)

Apêndice B

Publicação Científica

Parametric Study of the Analog Ensembles Algorithm with Clustering Methods for Hindcasting with Multistations

Leonardo Araújo[1], Carlos Balsa[2], C. Veiga Rodrigues[3], and José Rufino[2⋆]

[1] Universidade Tecnológica Federal do Paraná,
Campus de Ponta Grossa, 84017-220 Ponta Grossa,
leonardoa.2016@alunos.utfpr.edu.br
[2] Research Centre in Digitalization and Intelligent Robotics (CeDRI), Instituto Politécnico de Bragança, Campus de Santa Apolónia, 5300-253 Bragança, Portugal
balsa@ipb.pt,rufino@ipb.pt
[3] Vestas Wind Systems A/S - Design Center Porto, Portugal,
carlos.rodrigues@fe.up.pt

Abstract. Weather prediction for locations without or scarce meteorological data available can be attempted by taking meteorological datasets from nearby stations. This hindcasting problem can be successfully solved using the Analog Ensemble (AnEn) method. This paper presents a parametric analysis of the AnEn method, and two variations (based on K-means and fuzzy C-means clustering methods), when used to search for analog ensembles in a historical dataset. The study allowed to identify the parameter combinations that yield the best prediction accuracy, improving 13% on the systematic error and 5% on the random error of the previous results obtained with the same dataset. In addition, important performance gains were achieved at the computational level.

Keywords: Analog Ensembles, Clustering, Time series, Meteorological data, Hindcasting.

1 Introduction

The Analog Ensemble (AnEn) method was originally used to estimate the uncertainty in a *deterministic forecast* produced by a Numerical Weather Prediction (NWP) model [1], having later developed into a data-based prediction method [2]. Alongside this forecast, *historical forecasts* are also available (a record of forecasts from NWP at past dates), as well as *historical observations* (a record of real meteorological data observed in a given place). To improve forecasting accuracy, a new forecast is compared to historical forecasts. The most similar past forecasts (analogs) to the new forecast are then mapped into corresponding observations and these are used to refine the new forecast.

⋆ This work has been supported by FCT – Fundação para a Ciência e Tecnologia within the Project Scope: UIDB/05757/2020.

As such, the AnEn method is mostly used as a post-processing procedure to improve the forecast. A deterministic NWP forecast model reports on a possible state of the atmosphere among many possible states. As the predicted state does not always match reality (due the model limitations and its inputs), it is useful to have a probability distribution function (PDF) of the possible states that quantifies the uncertainty associated with the prediction. The forecast PDF can be estimated using a set of N_a past validation observations corresponding to the N_a best analogs (past model forecasts) to the current NWP model forecast [2].

Significant contributions to the use of the Analog Ensemble method in meteorology were made by Monache [3, 2] and applied to a variety of operational scenarios [4–6], where its accuracy and usefulness were clearly demonstrated.

In another work [7], the AnEn method was adapted to a hindcasting problem concerning the weather prediction for a location with no data available. The results clearly showed the potential of the AnEn method, especially with two predictor stations. The need to study the influence of some of the algorithm parameters was also evidenced (like the width and number of analogs used in predictions), as well as the need to compare the K-means method (used to obtain the set of analogues) with alternative clustering methods and to evaluate the effect of the number of clusters used to decompose the data history.

The present work thus seeks to improve the performance of the AnEn algorithm in the context of multistation hindcasting, by exploring the new research paths proposed in [7]. Tests are carried out to determine the effect of the main parameters of the AnEn algorithm, namely the number of analogues, the window width and the number of clusters. Two alternative clustering methods are also included to determine the set of analogues (K-means and fuzzy C-means).

This paper is structured as follows. Section 2 presents the AnEn algorithm, along with different methods used to determine the analog ensembles; it also defines the measures used for similarity, prediction and errors. Section 3 presents a parametric study of the main variables of the AnEn algorithm that influence the prediction results, and a comparison between two ways of clustering the analog ensembles. Section 4 ends with the final considerations of this work.

2 Analog Ensemble Method and Variants

This section provides a short description of the Analog Ensemble (AnEn) approach and of the K-means and fuzzy C-means clustering variants used in this work to determine the analog ensembles. It also tackles similarity assessment between analogs and predictor, prediction methods, and error assessment metrics.

2.1 Analog Ensemble Method

The original AnEn method [2, 4] uses two time series with meteorological data: *historical data*, from a Numerical Weather Prediction (NWP) model, and *observed data*, with real meteorological records. To improve the accuracy of the NWP model for a certain prediction period, a new forecast in that period is

compared to historical forecasts of a training period, in order to find the past forecasts most similar to new forecast; these past forecasts form an *analogs ensemble*; then, the actual meteorological values observed at the same points in time as the analogs, are used to improve the accuracy of the new forecast.

The AnEn methodology may also be adapted to reconstruct absent meteorological data of a predicted weather station, by resorting to real meteorological data from nearby predictor stations. In this scenario, the *historical data* corresponds to past data from the predictor stations, and the *observed data* is past meteorological data from the predicted station. This methodology is illustrated in Figure 1 for the case of a single predictor station, and follows 3 mains steps.

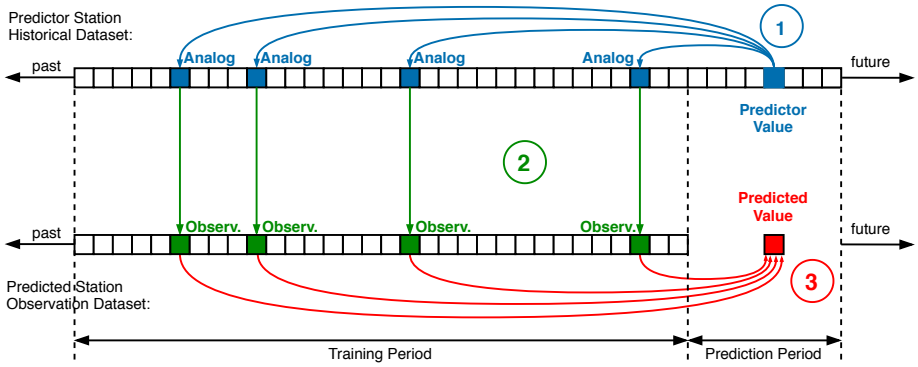


Fig. 1. Hindcasting with the Analog Ensemble method.

In step 1, an ensemble of analogs to the predictor is selected in the historical data set; each analog is a vector of $2k+1$ elements, where each element represents the value of a meteorological variable at a given instant and k represents half of the time window covered by the analog; analogs are selected because of their similarity to the predictor, with several similarity metrics being available.

In step 2, the analogs selected in the predictor station(s) are matched with the corresponding observations taken at the same time in the predicted station. Then, in step 3, these past observations are used to estimate (hindcast) the predicted value missing at the same instant of the known predictor.

This work uses the methodology just described, with a modification: real (observed) values corresponding to the predicted values are indeed available (not absent); thus, it is possible to assess the error of a predicted value in relation to the matching real value; as such, this study is about finding the combination of AnEn parameters that minimizes such errors; in addition, the effect of K-means and fuzzy C-means analogs clustering on the errors is also investigated.

2.2 Similarity Assessment

The choice of the analogs depends on their similarity to the predictor. The similarity metric originally used by Monache [2] is shown in Equation 1 (see also Table 1). This metric considers several (N_v) meteorological variables for comparing forecasts, and looks not only for a similar weather pattern, but also for similar numerical values to the various variables used in the forecasts (thus the inclusion of the standard deviation σ_{fi} in the metric). In addition, a weight w_i is assigned to each variable used. Also, F and A represent the forecast given a time t, in the prediction period and in the training period respectively. For a complementary description of this metric, including some variations, see [7].

$$\sum_{i=1}^{N_v} \frac{w_i}{\sigma_{fi}} \sqrt{\sum_{j=-k}^{k} (F_{i,t+j} - A_{i,t'+j})^2}. \tag{1}$$

In this work, only one meteorological variable is considered, and so a simplified similarity metric could have been used – see Equation 2.

$$\sqrt{\sum_{j=-k}^{k} (F_{i,t+j} - A_{i,t'+j})^2}. \tag{2}$$

However, two predictor stations are used instead of one, and so Equation 2 must be adapted to consider values from both – see Equation 3 (see also Table 1).

$$\sqrt{\sum_{j=-k}^{k} [(F^{s1}_{i,t+j} - A^{s1}_{i,t'+j})^2 + (F^{s2}_{i,t+j} - A^{s2}_{i,t'+j})^2]}. \tag{3}$$

Equation 3 corresponds to a *dependent station* approach [7]. It considers the additional predictor station to be nothing more than an additional predictor variable. As such, analogs are chosen together, across the two stations.

Table 1. Metrics parameters.

F_t	Forecast at given time t in the prediction period.
$A_{t'}$	Analog forecast at a time t' in the training period.
N_v	Number of meteorological variables considered when comparing forecasts.
w_i	Weight given to variable i.
σ_{fi}	Standard deviation of variable i in the historical dataset.
k	Half the width of the time window over which the metric is computed.
$F_{i,t+j}$	Forecast value at time $t+j$ in the time window for variable i.
$A_{i,t'+j}$	Analog value at time $t'+j$ in the time window for variable i.
$F^{s1}_{i,t+j}$	Forecast value at time $t+j$ in the time window from predictor station 1.
$A^{s1}_{i,t'+j}$	Analog value at time $t'+j$ in the time window from predictor station 1.
$F^{s2}_{i,t+j}$	Forecast value at time $t+j$ in the time window from predictor station 2.
$A^{s2}_{i,t'+j}$	Analog value at time $t'+j$ in the time window from predictor station 2.

2.3 Analog Clustering

Searching for the best analogs in the historical dataset may be very time consuming (even if the search is parallelized), due to the exhaustive nature of the search, that must cover all historical data points (and respective time windows centered on those points). An alternative to decrease this effort, is to use *clustering*. Under this approach, all historical data points (all possible analogs) are previously grouped into clusters, whereby related points (accordingly with the clustering algorithm) fall in the same cluster. The centroid of a cluster, generated from the average of all member points, can then be compared with the predictor. Doing this for all clusters would allow to determine the centroid most similar to the predictor. The cluster corresponding to this centroid, in whole or in part, will determine the set of observations used in the forecast.

Consequently, using clustering, the number of comparisons needed for similarity assessment is drastically reduced, once only the formed clusters centroids are compared with the predictor in order to chose the analogs.

This work explores two clustering methods: K-means and fuzzy C-means (henceforth simply C-means). The most relevant difference between both is that while K-means places each data point into a single cluster, C-means assigns each point to all clusters, with a specific membership weight per cluster. However, when using C-means, a data point ends up considered to belong to the cluster where it has the highest membership weight; these weights may be used in the step 3 of the AnEn algorithm, when calculating the predicted value. For further details on these two clustering algorithms refer to [8].

2.4 Prediction Method

With an ensemble of N_a analogs chosen, it becomes then possible to select the corresponding observations O_{t_i} and determine the predicted value P_t. This may be done by a simple arithmetic average of the observations:

$$P_t = \frac{1}{N_a} \sum_{i=1}^{N_a} O_{t_i} \tag{4}$$

Alternatively, when using C-means clustering, each observation O_{t_i} may be assigned a weight $w(O_{t_i})$ derived from the membership weight $w(A_{t_i})$ of the corresponding analog in its cluster. Basically, $w(O_{t_i})$ is $w(A_{t_i})$ normalized by considering all analogs of the best C-means cluster. Such weights may then be used to calculate the prediction as a weighted average:

$$P_t = \sum_{i=1}^{N_a} [w(O_{t_i}).O_{t_i}] \tag{5}$$

In this work, historical data from two different stations is used to determine the observations O_{t_i} that are going to be used to predict the value of P_t.

2.5 Error Assessment

To assess the accuracy of predictions, they are compared with the real values (once these are available in this work). That comparison resorts to several error measures. Previous research [9] showed the relevance of using together several of these measures. In this study, the error assessment metrics used are: Bias, Root Mean-Squared Error (RMSE), Standard Deviation of the Error (SDE) and Mean Absolute Error (MAE). In this section, only a brief explanation is provided for each metric. For the respective formulae and a similar usage context see [7].

The Bias measures the average error of the values obtained in comparison to the real values, verifying how close to the truth the results are. It is a very useful measure; however, this measure alone is not enough to assess the accuracy of the predictions. It only shows the systematic error of the predictions.

The RMSE is widely used as a standard comparison measure in meteorology. It is very sensitive to non-standard values, that is, when a value deviates a lot from the expected one it causes a great impact on the final value of the RMSE.

In addition to Bias and RMSE, the Standard Deviation of the Error (SDE) is also used in this study. Considering the Bias as a basic indicator of the *systematic error* in a prediction, the SDE is the equivalent indicator of the *random error*.

Finally, the MAE measure represents the absolute average distance from the truth. Despite having a formula similar to the Bias, it provides a measure a little closer to the truth as it does not allow positive and negative errors to cancel out.

3 Computational Experiments

This section presents the meteorological datasets used, the tests performed, the results obtained, and the computational enhancements that allowed for a faster execution of the hindcasting. Three parameters were selected to be studied: the Number of Clusters (N_c), the Number of Analogs (N_a) and the Windows Size (W_s). These parameters were chosen because they do not have a defined ideal value. Moreover, they are of great importance, as they directly influence the organization of the data that will be used for calculating the forecasts.

3.1 Meteorological Datasets

The meteorological data used for this study comes from the United States National Data Buoy Center [10]. The selected weather stations are located on the east coast of the United States, more specifically in the state of Virginia. They have weather data available since 2011 and such data can be used free of charge.

The stations used are shown in Figure 2: *ykr* at 37°15'5" N 76°20'33" W, *ykt* at 37°13'36" N 76°28'43" W, and *dom* at 36°57'44" N 76°25'27" W. They collect six meteorological variables: pressure (PRES), air temperature (ATMP), water temperature (WTMP), wind speed (WSPD), gust speed (GST) and wind direction (WDIR). Records are made mostly every 6 minutes, but data is not always complete (there is the possibility of periods without any data).

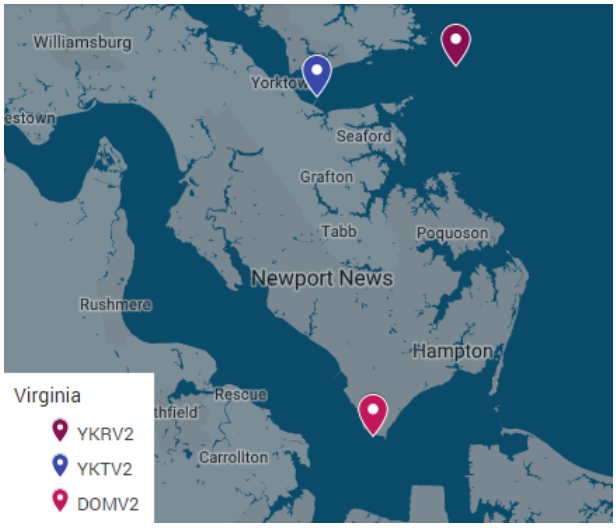


Fig. 2. Geolocation of the NDBC meteorological stations in Virginia [10]

A period of 9 years is used for the study, from 2011 to 2019. The data is separated into two groups: i) data from 2011 to 2018, used in the training period; ii) data from 2018 to 2019, used in the prediction period.

Data was not always originally collected with the same frequency and so it needed to be reorganized, in our case to match to a sampling period of 6 minutes. Having data for a period of 9 years, organized in samples taken every 6 minutes, implies a large volume of information to be dealt with, making the computational procedures to take a long time to run. For that reason, the data was filtered: only values between 10 am and noon were used.

In a previous study [7] the gust speed (GST) data for the station *ykt* was reconstructed, for the period corresponding to the years 2017 and 2018, using the value of GST at three different pairs of stations. The results induced that the best reconstruction is obtained using the predictive stations pair <*dom,ykr*> with the *dependent method*. In this work, we again consider the predictive stations pair <*dom,ykr*> as *dependent stations*, for rebuilding the GST over the years 2018 and 2019. We tried to find out if it is possible to improve the reconstruction by changing some of the parameters of the AnEn algorithm and using clustering.

3.2 Variation of the Number of Clusters

The number of clusters (N_c) is a variable relevant when using the K-means and C-means methods. It must be set *a priori* and indicates how many classes the historical dataset will be decomposed into. Clustering is done only with the training period data from the predictor station(s); this amount of data varies according to the limits defined for the training period and prediction period.

The influence of the variation of N_c is best understood through an example. Consider, for instance, a training period with 40,000 historical data elements and $N_c = 200$; in a real scenario, each cluster would have a varying number of elements but, for this example, lets assume that each cluster will receive the same average number of elements and that such number would also be 200; only one cluster is selected as an analog ensemble and, once chosen, a total of 200 analogs belonging to that cluster can be used to calculate the final forecast.

Increasing or decreasing N_c will have a direct influence in the number of analogs that can be used in calculating the prediction. In this example, increasing N_c to 2000 would make the number of elements per cluster to be reduced to 20 (on average), and such small number may not be enough to make a good prediction. On the other hand, reducing N_c to 20 would cause each cluster to have 2000 elements (on average), and this can worsen the accuracy of the prediction given the excessive amount of analogs that generate a lot of numerical noise.

To assess the impact of the variation of N_c in our scenario, the K-means variant of the AnEn method was tested, with values of N_c between 30 an 1000.

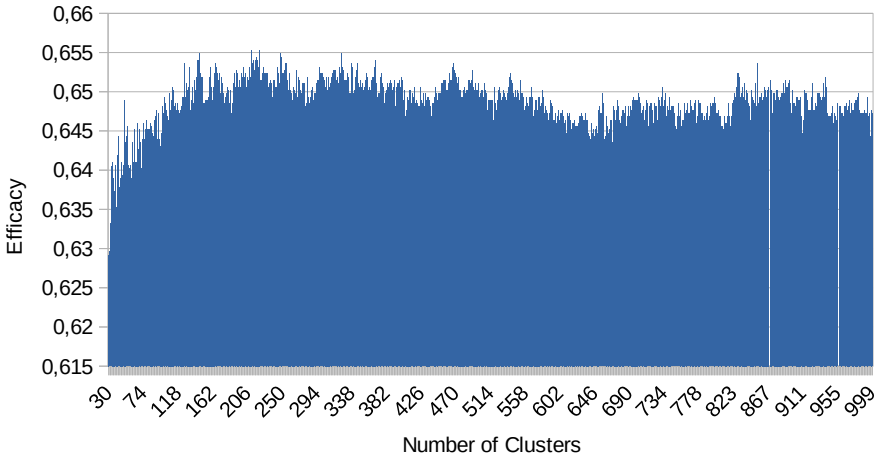


Fig. 3. Efficacy with different values of N_c for the K-means approach.

Figure 3 shows the results of the tests performed, expressed in an *efficacy* metric, given by 1/SDE. A high efficacy indicates a small value of the SDE along all the reconstruction period. The efficacy is defined here based on the SDE that, as mentioned in Section 2.5, is an indicator of the random error. SDE is given by $\sqrt{RMSE^2 - BIAS^2}$ and is particularly appropriate for this comparison.

The combination of stations considered in this study creates a training period of approximately 45,000 data elements. It is possible to notice in Figure 3 that the best results, in terms of efficacy, are not associated with very high or very low values of N_c. In the range studied, we found that the best efficacy is achieved for N_c close to 212 – the square root of the total number of data elements.

Despite this apparent relation, it is not possible to say that it always points to the N_c that ensures the best predictions; it is only possible to state that small and very large values of N_c, in relation to the square root of the dataset size, have less chance of getting the best results. Thus, the choice of the square root of the total number of data elements, for N_c, seems to be an appropriate heuristic.

3.3 Variation of the Number of Analogs

Like N_c, the number of analogs (N_a) of the analog ensemble is an essential parameter of the AnEn method. It indicates how many analogs will be used to select the observations that will feed the forecast. The way in which N_a is defined depends on the AnEn variant pursued (Monache, K-means or C-means).

The Monache approach involves choosing the N_a analogs with the lowest values of the metric given by Equation 2. With the clustering variants, the size of the analog ensemble has an upper limit given by the number of elements of the clusters formed, being possible to set N_a up to the size of the best cluster.

Table 2 shows the results produced (in terms of the four error metrics previously established) with different values of N_a when applying the Monache, K-means and C-means methods. For the clustering methods, N_c was set to 300.

The results obtained are expected to follow the logic presented above in the discussion of the N_c variation. A small number of N_a is not enough to make a good prediction due to the loss of information. In Table 2, it can be seen that for small values of N_a, regardless of the method, the largest of the errors are obtained. The Monache method has a decrease in error until N_a is equal to 100, remained constant until 200 and after that the errors grow again constantly.

Clustering methods have the amount of N_a, which makes sense to be used, limited by the size of the clusters. In these tests, the largest cluster formed by the K-means method had 357 elements and the smallest had only 6; in the C-means method the largest formed cluster had 862 elements and the smallest had only 1; this difference comes from the way the clusters are calculated.

That said, although the best results appear when a specific N_a is defined, results obtained when N_a is not specified (indicated by —), are very close to the best. Not specifying the N_a value means using all elements of the selected cluster as analogs. Thus, when the size of the analog cluster is not too large, it is preferable to define N_a as being equal to that dimension.

The results obtained with the K-means and C-means methods are very similar to each other and better than the ones obtained with the Monache method.

3.4 Variation of the Window Size

Previously, an analog was defined as a vector of $2k$+1 meteorological values, where k is half of the time window covered by the vector (see section 2.1). To say that k has a value of 10 means that the vector covers a time span of two hours (60 minutes before the central data sample and 60 minutes after it).

Tests were carried out with different values of k, seeking to understand their impact on the predictions. Table 3 shows the results for $k = 1, 2, ..., 10, 15$. In

Table 2. Predicting GST at the *ykt* station with a different number of analogs

N_a	Method	Bias	RMSE	MAE	SDE
—	K-means	0.264	1.556	1.171	1.534
—	C-means	0.270	1.559	1.166	1.536
20	Monache	0.802	2.311	1.749	2.167
20	K-means	0.367	1.607	1.215	1.564
20	C-means	0.367	1.598	1.207	1.556
40	Monache	0.798	2.297	1.737	2.154
40	K-means	0.361	1.581	1.190	1.539
40	C-means	0.361	1.582	1.191	1.540
60	Monache	0.794	2.291	1.732	2.149
60	K-means	0.337	1.579	1.191	1.542
60	C-means	0.343	1.571	1.179	1.533
80	Monache	0.790	2.289	1.731	2.148
80	K-means	0.331	1.571	1.183	1.536
80	C-means	0.337	1.563	1.173	1.527
100	**Monache**	**0.786**	**2.288**	**1.729**	**2.148**
100	K-means	0.314	1.566	1.180	1.534
100	C-means	0.335	1.566	1.173	1.530
150	Monache	0.778	2.288	1.729	2.151
150	K-means	0.296	1.570	1.180	1.542
150	C-means	0.318	1.562	1.169	1.529
200	Monache	0.772	2.288	1.729	2.154
200	K-means	0.269	1.546	1.168	1.523
200	C-means	0.302	1.562	1.170	1.532
250	Monache	0.767	2.290	1.730	2.158
250	K-means	0.257	1.550	1.169	1.528
250	**C-means**	**0.288**	**1.556**	**1.165**	**1.529**
300	Monache	0.762	2.291	1.731	2.161
300	**K-means**	**0.256**	**1.555**	**1.169**	**1.534**
300	C-means	0.289	1.562	1.168	1.535

the Monach method the number of analogs was set to 100, which correspond to the best performance of the method in previous analysis (see Table 2). For the clustering methods, N_c is set to 300 and N_a is set to the dimension of the analog cluster in agreement with the results obtained in the previous sections.

As may be observed in Table 3, for the three methods, increasing k (which means increasing the time span of the analog vector) above 10, is directly linked to the increase in the error values. Also notice that values of k below 10 (even the smallest values) lead to results that can be considered satisfactory.

With the Monach method the best result is found with k=2. For the K-means and C-means methods, the best results are attained with k=4 and k=7, respectively. It can be concluded that, for any method, values of k <10 are better choices, since in all cases tested values of k >10 always present the worst results.

Table 3. Predicting GST at the *ykt* station with different window sizes

k	Method	Bias	RMSE	MAE	SDE
1	Monache	0.359	2.180	1.687	2.150
1	K-means	0.261	1.561	1.173	1.539
1	C-means	0.306	1.570	1.178	1.539
2	**Monache**	**0.484**	**2.177**	**1.673**	**2.123**
2	K-means	0.265	1.558	1.172	1.536
2	C-means	0.316	1.563	1.173	1.530
3	Monache	0.592	2.203	1.683	2.122
3	K-means	0.261	1.553	1.167	1.531
3	C-means	0.327	1.568	1.180	1.534
4	Monache	0.670	2.230	1.697	2.126
4	**K-means**	**0.262**	**1.546**	**1.163**	**1.523**
4	C-means	0.333	1.564	1.178	1.528
5	Monache	0.731	2.254	1.710	2.133
5	K-means	0.257	1.550	1.169	1.528
5	C-means	0.337	1.563	1.173	1.527
6	Monache	0.777	2.280	1.727	2.143
6	K-means	0.261	1.560	1.177	1.538
6	C-means	0.337	1.570	1.174	1.533
7	Monache	0.798	2.297	1.737	2.154
7	K-means	0.271	1.558	1.172	1.534
7	**C-means**	**0.336**	**1.560**	**1.172**	**1.524**
8	Monache	0.823	2.310	1.745	2.159
8	K-means	0.261	1.562	1.181	1.540
8	C-means	0.346	1.578	1.183	1.540
9	Monache	0.843	2.321	1.749	2.163
9	K-means	0.265	1.563	1.180	1.540
9	C-means	0.338	1.578	1.183	1.541
10	Monache	0.857	2.329	1.751	2.166
10	K-means	0.260	1.569	1.185	1.547
10	C-means	0.349	1.585	1.189	1.547
15	Monache	0.889	2.354	1.762	2.180
15	K-means	0.276	1.597	1.203	1.573
15	C-means	0.334	1.591	1.196	1.556

3.5 Variation of the Weight Membership

A final study was made using only the C-means clustering method. As mentioned in section 2.4, a weight w_i can be associated with the observations used in the predictions. Table 4 presents test results where the weights, obtained from the membership variable, are inserted in the calculation of prediction. Additionally, the values obtained without weight are also included, in which the forecast is made through the arithmetic mean of all observations. The study was done considering $N_c = 300$ and N_a equal to the size of the analog cluster.

It is possible to understand from these results that, assigning the weights to the forecast calculation did not influence significantly the results, due to the low

Table 4. Results of tests with membership weight C-means

Weight	k	BIAS	RMSE	MAE	SDE
yes	1	0.303	1.571	1.179	1.542
no	1	0.306	1.570	1.178	1.539
yes	2	0.325	1.568	1.177	1.534
no	2	0.316	1.563	1.173	1.530
yes	3	0.324	1.562	1.176	1.528
no	3	0.327	1.568	1.180	1.534
yes	**4**	**0.327**	**1.562**	**1.174**	**1.527**
no	4	0.333	1.564	1.178	1.528
yes	5	0.338	1.568	1.178	1.531
no	5	0.337	1.563	1.173	1.527
yes	6	0.386	1.603	1.192	1.556
no	6	0.337	1.570	1.174	1.533
yes	7	0.373	1.588	1.190	1.543
no	**7**	**0.336**	**1.560**	**1.172**	**1.524**
yes	8	0.384	1.596	1.195	1.549
no	8	0.346	1.578	1.183	1.540
yes	9	0.378	1.592	1.193	1.546
no	9	0.338	1.578	1.183	1.541
yes	10	0.344	1.580	1.183	1.542
no	10	0.349	1.585	1.189	1.547
yes	15	0.336	1.596	1.195	1.560
no	15	0.334	1.591	1.196	1.556

variation of the error values. Moreover, for most of these comparisons, inserting weights into the prediction worsened the result obtained. However, the inclusion of weights causes the optimal k to be reduced from 7 to 4.

3.6 Results Comparison

This section presents a comparison of the best results obtained with each method.

Figure 4 is a Taylor diagram that supports a visual comparison for the studied methods. It shows that the K-means and C-means methods have results very close to each other and that the Monache method is not far from their results. K-means and C-means are closer to the *truth* (represented by the green ellipse on the bottom of the diagram), due to high correlation coefficients, low SDE value and larger standard deviation. However, their standard deviations are farther from the *truth*, meaning they do not accurately follow variations in real values.

Table 5 presents the values of the different errors corresponding to best error values of the three methods. These values are also compared with the best results obtained in previous work [7] (identified as "old"). As shown in Table 5, the values are better than those obtained previously with the Monach and K-means methods. This shows the importance of choosing the correct values for the parameters of the AnEn method. On the other hand, it is clear that there are no

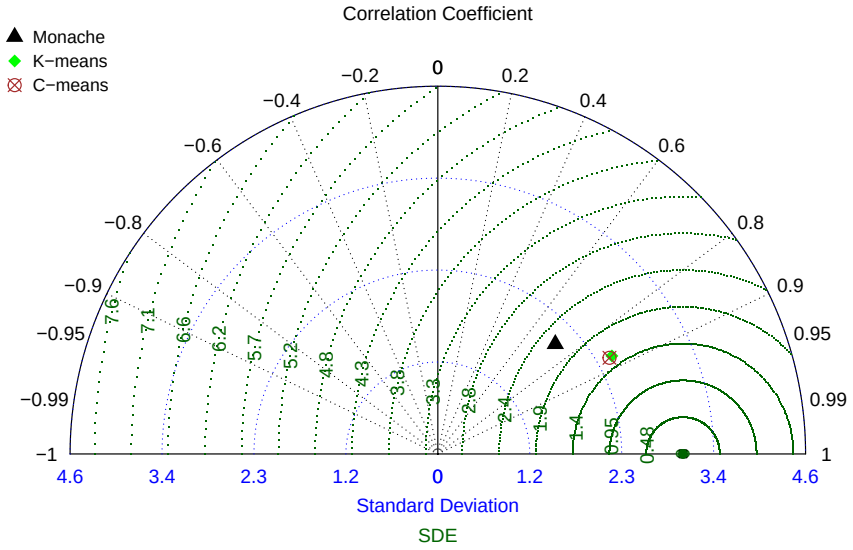


Fig. 4. Taylor diagram of the best result of each method.

Table 5. Comparison between the best results.

	Method	BIAS	RMSE	MAE	SDE
old	Monache	0.950	2.340	1.750	2,138
new	Monache	0.484	2.177	1.673	2.123
old	K-means	**0.300**	1.630	1.220	**1.600**
new	K-means	**0.262**	1.546	1.163	**1.523**
new	C-means	0.336	1.560	1.172	1.524

advantages in using the C-means method instead of K-means in the formation of candidate clusters for analogs, even though the C-means method is more robust.

Comparing the new results for K-means with ones from previous investigation [7], there is an improvement of ≈13% ((0.300-0.262)/0.300) on the systematic error (Bias) and ≈5% ((1.600-1.523)/1.600) on the random error (SDE).

3.7 Computational Enhancements

Improving the accuracy of the predictions was not the only objective achieved by this work. Important gains were also introduced in the computational codes that were used, both in data handling and execution time.

Before the changes introduced, the code used for this line of investigation [7] spent considerable time for the loading of the meteorological data and its organization in a way that it could be used easily during the analogs search: moreover, this procedure was performed per each execution of the algorithms.

For instance, for the K-means algorithm, the file loading time corresponded to approximately 65% of the total code execution time; now, this time has been

reduced to only 7% of that time. This was possible by using pre-organized files (the data used often repeats a specific set of variables, which makes this pre-organization possible). On the other hand, it takes a certain amount of additional storage to hold these new files; however, it ends up being worthwhile due to the time saved when performing many executions of the same algorithm.

In addition, the algorithms that perform clustering have changed from just one run at a time to several simultaneous runs, which speeds up the processing of a large set of tests by almost 400% when using the K-means algorithm.

The code is written in R [11] and all algorithm implementations have code that runs in parallel, taking advantage or R facilities for loop parallelization. Also, at no time was the memory consumption a limiting factor for the performance of the codes; what limits the performance in this case is the speed of the processor and its number of cores. The tests were performed on a Linux system running in a machine with a 16-core AMD EPYC 7351 CPU and 64 GB of RAM.

4 Conclusion

In this work, a parametric analysis of some important variables of the Analog Ensemble algorithm was performed, considering three different methods for the determination of the analogs: Monache, K-means, and C-means.

For the K-means variant, the definition of the number of clusters is essential. Though *a priori* estimation is difficult, the results observed show that choosing the square root of the time series dimension is a reasonable heuristic.

The C-means method does not seem to bring any benefits in comparison to the classical K-means clusterization, meaning that different weights for the analogs based on the membership variable did not result in improvements.

The right choice of the number of analogs varies with the method chosen to determine the analogs. In clustering methods (K-means and C-means), it is generally preferable to choose all elements belonging to the analog cluster. For the Monache method, the number of analogs with the best results is obtained for values near 100. This value can change, however, for tests with other variables.

Considering the size of the windows, the right choice seems to be values for k <10 for all tested methods. But this observation is related to the gust speed (GST) variable, using data from two specific forecast stations. Tests with other variables and stations are needed to infer if such values are still the most suitable.

References

[1] H. M. van den Dool, "A new look at weather forecasting through analogues," *Monthly Weather Review*, vol. 117, no. 10, pp. 2230 – 2247, 01 Oct. 1989.

[2] L. D. Monache, F. A. Eckel, D. L. Rife, B. Nagarajan, and K. Searight, "Probabilistic Weather Prediction with an Analog Ensemble," *Monthly Weather Review*, vol. 141, no. 10, pp. 3498–3516, 2013.

[3] L. D. Monache, T. Nipen, Y. Liu, G. Roux, and R. Stull, "Kalman filter and analog schemes to postprocess numerical weather predictions," *Monthly Weather Review*, vol. 139, no. 11, pp. 3554–3570, 2011.

[4] S. Alessandrini, L. D. Monache, S. Sperati, and J. N. Nissen, "A novel application of an analog ensemble for short-term wind power forecasting," *Renewable Energy*, vol. 76, pp. 768–781, 2015.

[5] S. Alessandrini, L. D. Monache, S. Sperati, and G. Cervone, "An analog ensemble for short-term probabilistic solar power forecast," *Applied Energy*, vol. 157, pp. 95–110, 2015.

[6] S. Alessandrini, L. D. Monache, C. M. Rozoff, and W. E. Lewis, "Probabilistic prediction of tropical cyclone intensity with an analog ensemble," *Monthly Weather Review*, vol. 146, no. 6, pp. 1723–1744, 2018.

[7] C. Balsa, C. V. Rodrigues, I. Lopes, and J. Rufino, "Using analog ensembles with alternative metrics for hindcasting with multistations," *ParadigmPlus*, vol. 1, pp. 1–17, Jun. 2020.

[8] S. Panda, S. Sahu, P. Jena, and S. Chattopadhyay, "Comparing fuzzy-c means and k-means clustering techniques: A comprehensive study," in *Advances in Computer Science, Engineering & Applications* (D. C. Wyld, J. Zizka, and D. Nagamalai, eds.), (Berlin, Heidelberg), pp. 451–460, Springer Berlin Heidelberg, 2012.

[9] T. Chai and R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature," *Geoscientific Model Development*, vol. 7, no. 3, pp. 1247–1250, 2014.

[10] "National Data Buoy Center." https://www.nature.com/nature/ (visited on 2020-10-20).

[11] R Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013.

Apêndice C

Elementos Complementares

C.1 Efeito da Variação do Número de Clusters

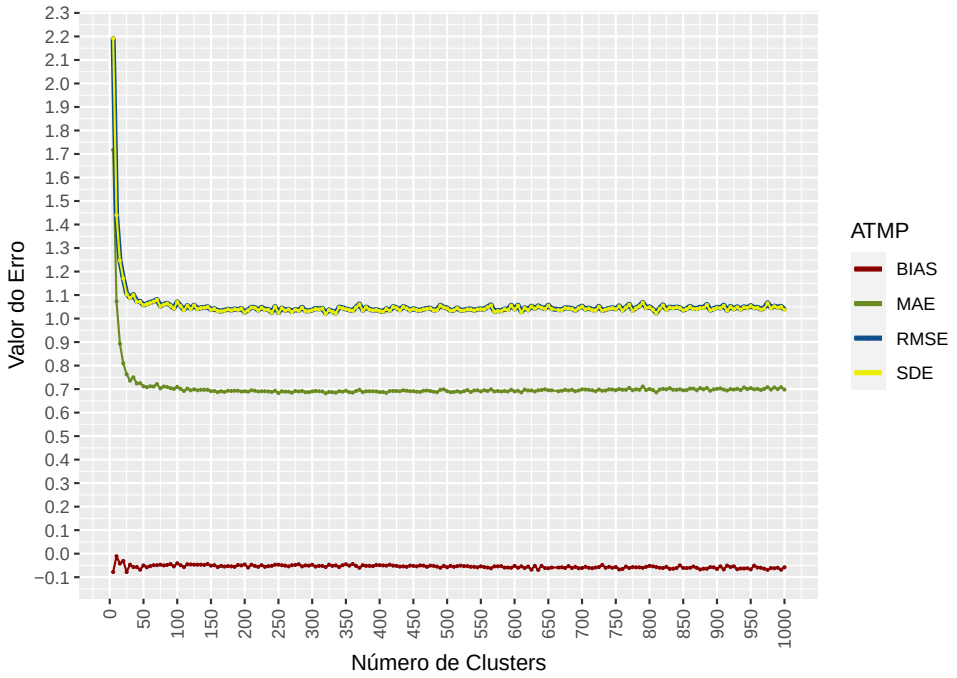


Figura C.1: Erros das previsões da variável ATMP com diferentes N_c - método C-means

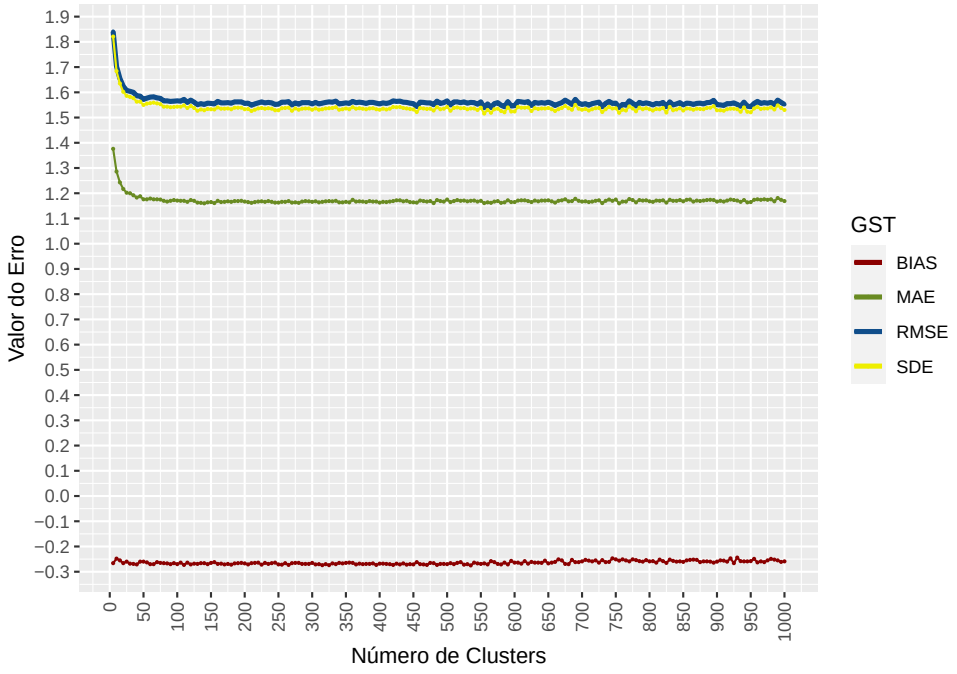


Figura C.2: Erros das previsões da variável GST com diferentes N_c - método C-means

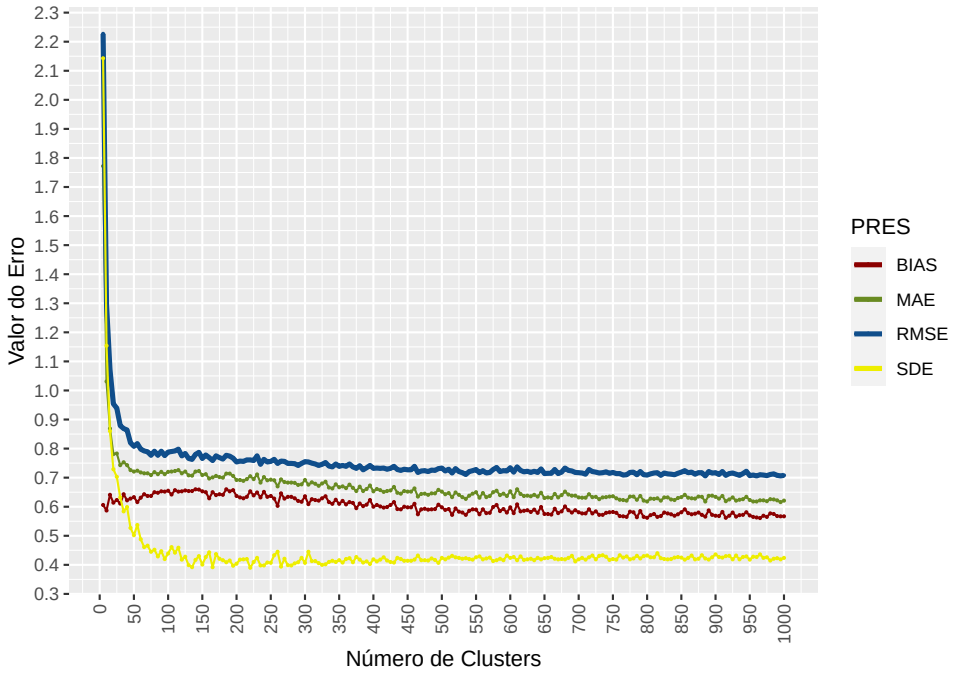


Figura C.3: Erros das previsões da variável PRES para diferentes N_c - método C-means

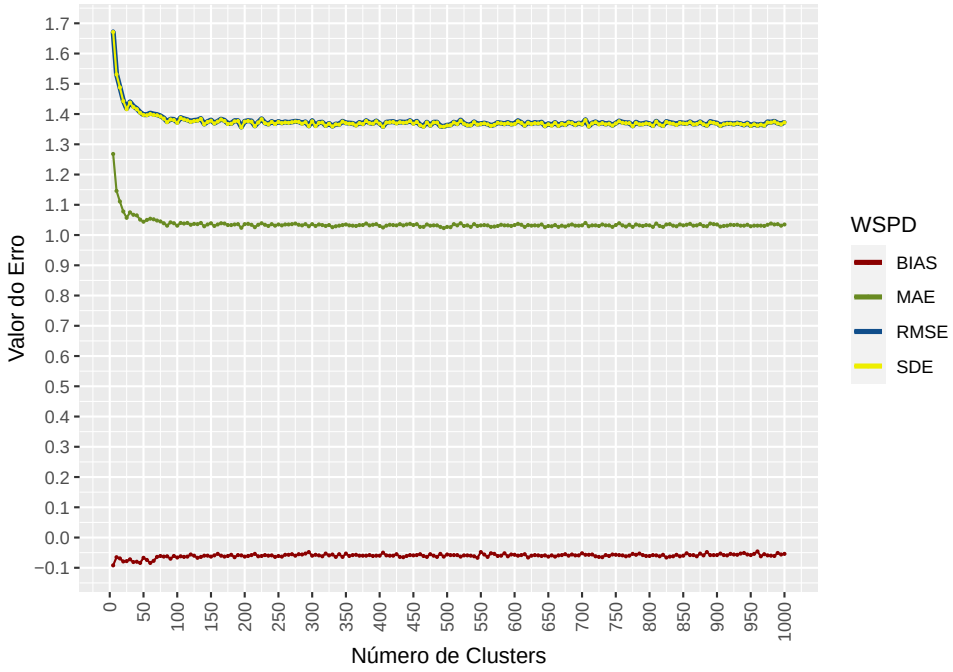


Figura C.4: Erros das previsões da variável WSPD para diferentes N_c - método C-means

C.2 Efeito da Variação do Número de Análogos

NA	BIAS	RMSE	MAE	SDE
—	1,15%	0,39%	-0,26%	0,39%
50	31,15%	1,67%	1,36%	0,72%
100	21,92%	0,71%	0,34%	0,07%
150	9,23%	1,16%	0,77%	0,91%
200	3,46%	0,64%	0,26%	0,52%
250	0,00%	0,06%	-0,09%	0,07%
300	-1,15%	0,52%	0,09%	0,59%
350	0,38%	0,26%	-0,26%	0,26%
400	**0,00%**	**0,00%**	**0,00%**	**0,00%**

(a) Método K-means.

NA	BIAS	RMSE	MAE	SDE
—	-8,45%	0,52%	0,43%	0,79%
50	23,59%	1,29%	1,64%	0,39%
100	15,85%	1,03%	1,03%	0,46%
150	8,45%	0,84%	0,95%	0,52%
200	2,82%	0,71%	0,52%	0,59%
250	3,52%	0,77%	0,86%	0,66%
300	**0,00%**	**0,00%**	**0,00%**	**0,00%**
350	-3,17%	0,39%	0,43%	0,46%
400	-2,46%	0,39%	0,52%	0,46%

(b) Método C-means.

NA	BIAS	RMSE	MAE	SDE
50	**0,00%**	**0,00%**	**0,00**	**0,00**
100	0,00%	0,18%	0,12%	0,19%
150	-0,27%	0,35%	0,23%	0,47%
200	-0,41%	0,58%	0,47%	0,70%
250	-0,68%	0,71%	0,58%	0,89%
300	-1,09%	0,80%	0,64%	1,08%
350	-1,50%	0,89%	0,76%	1,22%
400	-1,77%	0,98%	0,82%	1,36%

(c) Método Monache.

Tabela C.1: Diferença (%) entre erros na previsão de GST

NA	BIAS	RMSE	MAE	SDE
—	-28,85%	1,27%	0,58%	1,27%
50	7,69%	2,14%	1,74%	2,05%
100	5,77%	2,04%	0,72%	1,95%
150	11,54%	1,95%	0,72%	1,85%
200	-5,77%	0,29%	0,00%	0,29%
250	15,38%	2,04%	0,72%	1,95%
300	**0,00%**	**0,00%**	**0,00%**	**0,00%**
350	-3,85%	0,29%	-0,43%	0,19%
400	1,92%	1,36%	0,87%	1,36%

(a) Método K-means.

NA	BIAS	RMSE	MAE	SDE
—	-0,00%	0,19%	-0,15%	0,19%
50	-15,69%	1,65%	1,89%	1,65%
100	-25,49%	0,87%	0,58%	0,87%
150	-17,65%	1,07%	0,15%	1,07%
200	**0,00%**	**0,00%**	**0,00%**	**0,00%**
250	11,76%	1,16%	0,87%	1,07%
300	9,80%	1,45%	0,44%	1,36%
350	11,76%	1,26%	0,58%	1,16%
400	0,00%	0,29%	0,58%	0,19%

(b) Método C-means.

NA	BIAS	RMSE	MAE	SDE
50	-71,43%	1,13%	1,79%	1,13%
100	-85,71%	0,47%	0,55%	0,47%
150	-57,14%	0,09%	0,14%	0,09%
200	-28,57%	0,09%	0,14%	0,09%
250	**0,00%**	**0,00%**	**0,00%**	**0,00%**
300	14,29%	0,09%	0,00%	0,09%
350	42,86%	0,19%	0,14%	0,19%
400	71,43%	0,28%	0,14%	0,28%

(c) Método Monache.

Tabela C.2: Diferença (%) entre erros na previsão de ATMP

NA	BIAS	RMSE	MAE	SDE
—	6,25%	4,86%	6,55%	2,81%
50	**0,00%**	**0,00%**	**0,00%**	**0,00%**
100	6,44%	3,39%	5,69%	-1,64%
150	5,87%	3,53%	5,86%	0,00%
200	6,25%	4,71%	6,38%	2,11%
250	5,11%	3,98%	5,52%	2,11%
300	5,49%	4,57%	5,86%	3,04%
350	4,73%	3,83%	5,17%	2,11%
400	7,01%	3,98%	6,55%	-0,94%

(a) Método K-means.

NA	BIAS	RMSE	MAE	SDE
—	16,30%	10,56%	16,16%	-0,49%
50	**0,00%**	**0,00%**	**0,00%**	**0,00%**
100	8,97%	5,43%	8,92%	-0,98%
150	11,36%	6,89%	10,77%	-1,23%
200	12,64%	8,80%	13,30%	1,72%
250	11,36%	8,06%	11,95%	2,21%
300	9,89%	8,21%	11,28%	5,15%
350	12,64%	8,65%	12,12%	1,47%
400	14,47%	8,80%	13,47%	-2,21%

(b) Método C-means.

NA	BIAS	RMSE	MAE	SDE
50	1,83%	1,21%	0,26%	1,22%
100	0,37%	0,20%	-0,26%	0,24%
150	**0,00%**	**0,00%**	**0,00%**	**0,00%**
200	0,00%	0,20%	0,00%	0,49%
250	0,00%	0,81%	0,26%	1,46%
300	0,00%	1,42%	0,77%	2,43%
350	0,00%	2,23%	1,03%	3,41%
400	-0,37%	3,24%	1,28%	4,87%

(c) Método Monache.

Tabela C.3: Diferença (%) entre erros na previsão de PRES

NA	BIAS	RMSE	MAE	SDE
—	-1,67%	0,22%	0,49%	0,22%
50	101,67%	1,85%	1,28%	1,55%
100	61,67%	1,04%	0,59%	0,81%
150	25,00%	0,74%	0,69%	0,67%
200	11,67%	0,37%	0,49%	0,30%
250	13,33%	0,96%	0,79%	0,96%
300	**0,00%**	**0,00%**	**0,00%**	**0,00%**
350	0,00%	0,44%	0,49%	0,44%
400	-5,00%	0,89%	0,88%	0,81%

(a) Método K-means.

NA	BIAS	RMSE	MAE	SDE
—	-27,91%	0,59%	0,49%	0,74%
50	36,05%	1,40%	1,27%	1,25%
100	15,12%	0,96%	0,58%	0,96%
150	**0,00%**	**0,00%**	**0,00%**	**0,00%**
200	-11,63%	0,51%	0,29%	0,59%
250	-13,95%	1,25%	1,07%	1,33%
300	-18,60%	1,32%	1,07%	1,40%
350	-27,91%	0,73%	0,78%	0,81%
400	-19,77%	0,96%	0,78%	1,03%

(b) Método C-means.

NA	BIAS	RMSE	MAE	SDE
50	**0,00%**	**0,00%**	**0,00%**	**0,00%**
100	-2,29%	0,09%	0,13%	0,14%
150	-4,29%	0,09%	0,19%	0,24%
200	-5,43%	0,28%	0,38%	0,43%
250	-6,29%	0,38%	0,51%	0,53%
300	-7,14%	0,47%	0,63%	0,67%
350	-8,00%	0,62%	0,76%	0,82%
400	-8,57%	0,71%	0,82%	0,91%

(c) Método Monache.

Tabela C.4: Diferença (%) entre erros na previsão de WSPD

C.3 Eficiência Computacional

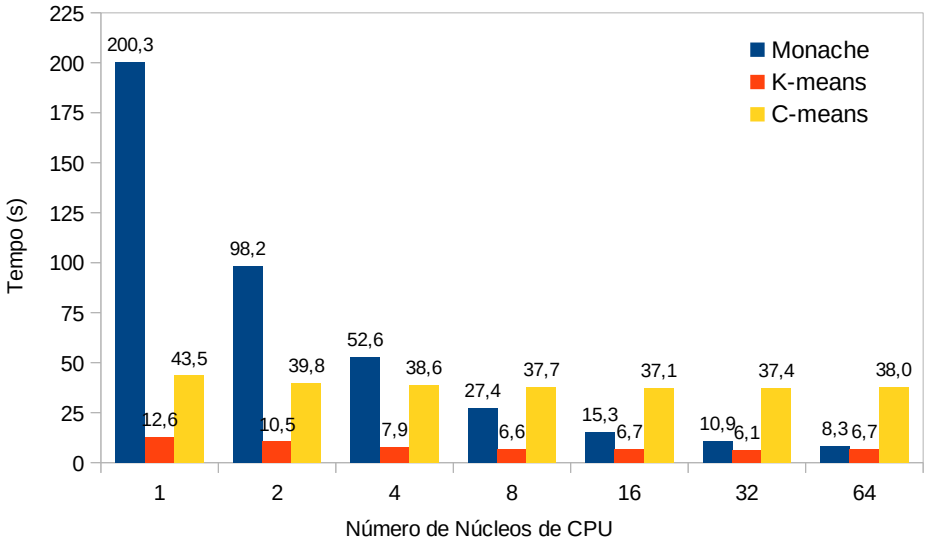


Figura C.5: Tempos (s) da previsão da variável GST em função do número de núcleos de CPU usados

		Número de Núcleos de CPU						
		1	2	4	8	16	32	64
Monache	**Speedup**	1,00	2,04	3,81	7,31	13,09	18,38	24,13
	Eficiência	100,0%	102,0%	95,2%	91,4%	81,8%	57,4%	37,7%
K-means	**Speedup**	1,00	1,20	1,59	1,91	1,88	2,07	1,88
	Eficiência	100,0%	60,0%	39,9%	23,9%	11,8%	6,5%	2,9%
C-means	**Speedup**	1,00	1,09	1,13	1,15	1,17	1,16	1,14
	Eficiência	100,00%	54,65%	28,17%	14,42%	7,33%	3,63%	1,79%

Tabela C.5: Speedup e Eficiência da previsão da variável GST em função do número de núcleos de CPU usados

Tabela C.6: Erros na previsão da variável v=GST

	Erros individuais (e_j)				Erros combinados
Métodos (m_i)	BIAS	RMSE	MAE	SDE	$\sum_j \lvert e_j \rvert$ s/ BIAS
Monache	-0,53	2,197	1,685	2,132	6,014
K-means	-0,257	1,554	1,171	1,533	**4,258**
C-means	-0,272	1,559	1,166	1,535	4,260

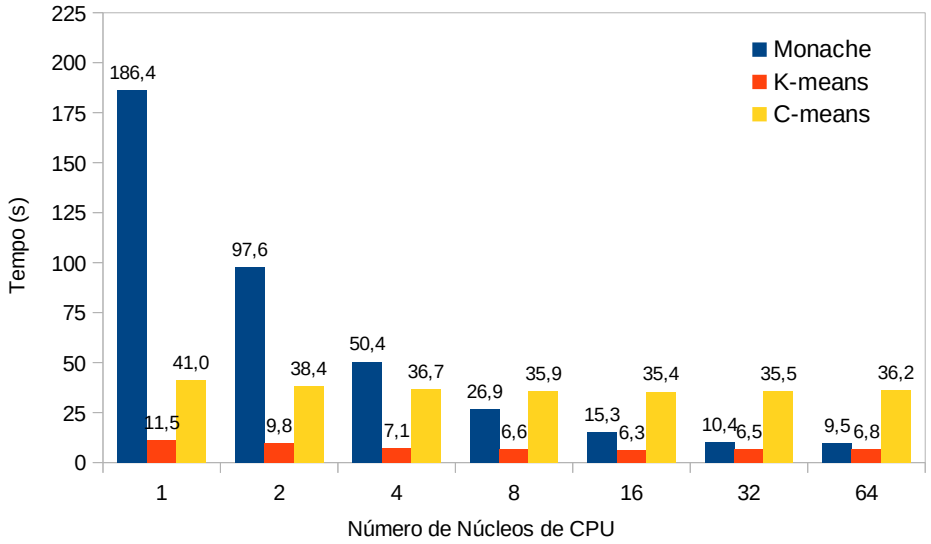


Figura C.6: Tempos (s) da previsão da variável WSPD em função do número de núcleos de CPU usados

		Número de Núcleos de CPU						
		1	2	4	8	16	32	64
Monache	**Speedup**	1,00	1,91	3,70	6,93	12,18	17,92	19,62
	Eficiência	100,0%	95,5%	92,5%	86,6%	76,1%	56,0%	30,7%
K-means	**Speedup**	1,00	1,17	1,62	1,74	1,83	1,77	1,69
	Eficiência	100,0%	58,7%	40,5%	21,8%	11,4%	5,5%	2,6%
C-means	**Speedup**	1,00	1,07	1,12	1,14	1,16	1,15	1,13
	Eficiência	100,00%	53,39%	27,93%	14,28%	7,24%	3,61%	1,77%

Tabela C.7: Speedup e Eficiência da previsão da variável WSPD em função do número de núcleos de CPU usados

Tabela C.8: Erros na previsão da variável v=WSPD

	Erros individuais (e_j)				Erros combinados
Métodos (m_i)	BIAS	RMSE	MAE	SDE	$\sum_j \lvert e_j \rvert$ s/ BIAS
Monache	-0,206	2,075	1,578	2,064	5,717
K-means	-0,064	1,355	1,021	1,353	**3,729**
C-means	-0,055	1,363	1,028	1,361	3,752