INSTITUTO POLITÉCNICO DE BRAGANÇA Escola Superior de Tecnologia e Gestão

Kuban State Agrarian University

Mathematical Statistics vs Machine Learning: Towards an intelligent modeling framework for soil and plant growth processes

Valentina Labintseva – a42652

Thesis presented to the School of Technology and Management in the scope of the Master in Information Systems.

.

Supervisors:

João Paulo Pais de Almeida

Darya Zamotajlova

Bragança

2019-2020

Dedication

To my graduate supervisors, friends and colleagues who helped me in this work.

Acknowledgment

This work was supported by Kuban State Agrarian University behalf I.T. Trubilin (KubSAU), Krasnodar, Russia, and Instituto Politécnico de Bragança (IPB), Bragança, Portugal, as well as National Agency Erasmus+Education and Training.

Resumo

O trabalho descrito nesta dissertação versa sobre métodos e técnicas no âmbito da Estatística Matemática e de *ML* usados para efeitos de previsão de colheitas e tratamento de solos em agricultura de precisão.

O objetivo do trabalho é investigar esses métodos em sua aplicação prática a um conjunto específico de dados.

No decorrer do trabalho, foram realizadas as seguintes tarefas: investigou-se a situação atual no campo da agricultura de precisão, investigaram-se os fundamentos teóricos dos métodos e técnicas da estatística matemática e de *ML*. Estes métodos e técnicas foram submetidos a testes práticos em um conjunto específico de dados. Foram tiradas conclusões sobre as vantagens e desvantagens desses métodos: Uma seslção de trabalhos científicos relacionados com a investigação sobre a introdução de um conjunto específico de nutrientes no solo foram também investigados.

As contribuições mais importantes para este trabalho são a aplicação prática de vários métodos de análise, bem como o projeto de uma ferramenta de apoio à decisão projetada para ajudar os agricultores a integrar a agricultura de precisão nas suas propriedades agrícolas.

Palavras Chave: Estatística matemática, *ML*, regressão linear, *clustering*, agricultura de precisão, agricultura inteligente, modelação, solo, plantas, nutrientes, humidade, ferramenta de apoio à decisão.

Abstract

The work described in this dissertation focuses on the methods for analyzing MS and ML that are used in PF.

The purpose of the work is to investigate these methods on their practical application to a specific set of data.

In the course of the work, the following tasks were completed: the current state of affairs in the field of PF was investigated, the theoretical foundations of the methods of MS and ML were investigated, which were subjected to practical tests on a specific set of data. Conclusions were drawn about the advantages and disadvantages of these methods. A selection of works of scientists engaged in research on the introduction of a specific set of nutrients into the soil was also investigated.

The most important contributions to this work are the practical application of various methods of analysis, as well as the design of a DST designed to help farmers integrate PF into their pilot training farms.

Key words: mathematical statistics, machine learning, linear regression, clustering, precision farming, intelligent farming, modeling, soil, plants, nutrients, moisture, decision support tool.

Contents

List of Figures

Table Index

Acronyms

ANN Artificial neural networks
DM Data Mining
DST Decision Support Tool
DT Decision Trees
GNSS Global Navigation Satellite System
GPS Global Positioning System
HC Hierarchical Clustering
LR Linear Regression
MA Multivariate Analysis
ML Machine Learning
MLR Multiple Linear Regression
MS Mathematical Statistics
NN Neural Networks
PF Precision Farming
TS Traditional Statistics

Chapter 1

Introduction

Agriculture, as a sphere of human life, began to progress no later than 10 thousand years ago. At this stage of progress of modern societies, the agricultural industry plays a key role in the economic condition of any country. The ability of a state to independently provide its citizens with food is one of the indicators of the country's independence.

Portugal's agriculture is represented by many industries, namely: the production of canned fish, processing of cork oak bark (leading place in the world), as well as viticulture, fruit growing, olive and chestnut trees growing. The population is engaged in both irrigated and rainfed farming, as the country's territory mainly extends to the mountain ranges (central and northern regions).

Thus, methods and techniques for growing fruit and field crops are a very important issue for the agricultural industry. Since the timely provision of plants and trees with the necessary nutrition components (water, minerals, fertilizers, etc.) is often a determining factor in the growth and evolution of plants, and also greatly affects the level of crop productivity, it is very important to thoroughly analyze the soil, samples from plantations (both young and old), as well as to monitor the amount of rainfall in the regions. This is

the only way to evaluate existing methods and techniques for growing crops in order to modernize and improve them.

It is possible to analyze the indicators of soil composition, plant stems, tree bark and other important components using modern scientific tools based on both MS methods (correlation-regression, variance, multicriteria analysis) and ML, both supervised and unsupervised learning, approaches.

Depending on the datasets and the type of task at hand, some methods may perform better than others. Therefore, it is often required to take several methods and combine their work into a kind of methodological symbiosis, which can give a more accurate result than applying each of the methods separately. Such DSTs can evolve into entire analytical systems.

At present, intelligent modeling of various processes is capable of not only processing large amounts of data, but also displaying non-trivial dependencies. Such intelligent modeling systems have extensive databases and knowledge bases that are used both to support decision-making and to modernize the system itself.

1.1 About the structure of the thesis

The objective of this dissertation is a reasoned answer to the question of the target suitability of using certain methods of MS and of ML in data analysis tasks to model soil and plant growth processes, which have a strong connection with the field of PA.

The organizational structure is implemented consistently in accordance with the main goals of this work:

- explore the current state of affairs in the field of IF;
- study the works of researchers engaged in research on related issues;
- study the theoretical foundations methods of MS and of ML and compare the advantages and disadvantages of these methods;

- to carry out practical implementation on the example of a certain set of data and draw conclusions from the results of the analyzes;
- design a DST based on the findings on the target suitability of these methods, which will be able to help farmers in integrating some elements of PF into their experimental training farms.

Chapter 2

State of the Art in Precision Farming

In order to improve agricultural efficiency and increase soil productivity and fertility, PF systems are being constantly developed. The concept of PF System arose back in the 1980s, when the United States began to make the first maps for differentiated fertilizer application based on soil analyzes.

The concept of PF has gained widespread acceptance over the past years due to the development of mobile technology, high-speed internet, Internet of Things, cloud computing and accurate monitoring data through remote sensing (satellite) [15]. PF systems, which are meant to contribute to a more sustainable and productive agriculture, are built upon the use of modern technology at all stages of the process. Different sections of the field are heterogeneous in their characteristics, and technologies allow to define these zones and take into account their features when planning work. Thanks to this, farmers more efficiently spend seeds, fertilizers and pesticides, while collecting higher yields. Intuition and luck mean less - technologies allow to make decisions based on accurate data and the rational resources using helps preserve the environment.

2.1 The main components affecting the condition of the soil

To make farming effective, the farmer needs to collect and analyze data on the field state at each work stage. Such data in the field of PF include the status of the following components:

- the soil: to determine heterogeneous patches of soil, farmers conduct an agrochemical survey of the soil. As a rule, such an analysis is done every four years. Soil samples are taken either manually or using special equipment, and then sent to the laboratory for research. According to its results, farmers make digital maps of the properties of their fields - they are used to create tasks for the technique of applying seeds and fertilizers. The soil is examined in more than 30 parameters, the main ones are acidity, the content of phosphorus, potassium and humus. Acidity (pH) is the easiest to find out. Its value can be found out both as a result of laboratory research, and when measuring soil with a field sensor. At the same time, acidity is one of the important factors of productivity - for each crop there is that pH value at which it grows best. According to the humus content, farmers assess the fertility of different parts of the field and calculate the norms of seeds and fertilizers. Phosphorus and potassium are necessary for the growth and development of plants, so it is important to know their content to calculate the exact rate of fertilizer [19];

- the yield: yield data is one of the most valuable in PF. On-board computers on technology record the amount of harvested with reference to the coordinates. Based on this information, farmers make digital maps that identify problem areas on the field. The reason for the low fertility can be determined by comparing the yield map with the relief map, distribution of nutrients, or other field indicators. For example, a farmer collected data on the yield of his field for two years and found plots with the worst indicators. He measured the acidity of the soil - it turned out that its level in these areas is very low. To bring acidity to the optimum level and increase the yield, the farmer deposited chalk in these areas. The more yield data, the better. Information for several years allows you to draw up accurate tasks for the differential application of fertilizers and seeds for the next season. Moreover, data for 5 years will help to save several times more than data for 1

year. An example of the distribution of productivity in a field is presented in Figure 2.1 [31];

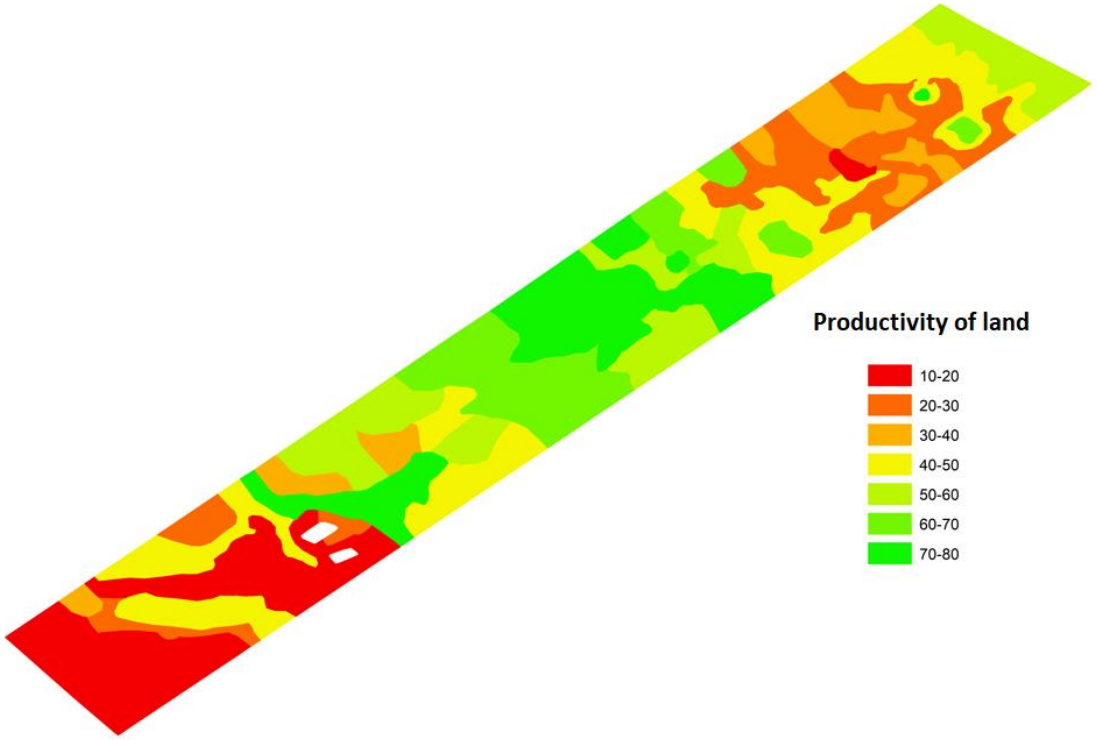

Figure 2.1: Map yield of field

- the work of equipment: with poor-quality field work, farmers incur large costs. If, when sowing seeds, the tractor made an inaccurate technological strip, then an overlap is formed on the field. Fertilizing and pesticides will pass in the same lane, and as a result, the double rate of expensive chemicals will be spent on the floors. For example, a tractor without an on-board computer sows a field of 100 hectares with rapeseed over a width of 6 meters. With each turn of the technique, an overlap of about 25 centimeters is formed. This increases the farmer's expenses for seeds, fertilizers and pesticides by 5%, or $ 1,500. If the equipment is equipped with on-board computers and GPS-navigators, then this problem is solved. Overlappings and skipped areas can either be completely avoided by accurate navigation, or quickly corrected after analyzing the work done. Also, the data from the equipment helps to assess the norms of fertilizer and pesticide application, the speed of the equipment in the field and control the operation of machine operators;

- the plants: seedlings are evaluated by multispectral satellite imagery. One of the most popular methods is measuring the vegetation index NDVI. The field is divided into sections with different indices, which allows you to quickly determine the state of seedlings even in hard-to-reach areas. Also, based on the vegetation index, maps are made for applying fertilizers and pesticides by zone;

- the weather: weather stations and sensors allow to remotely monitor the weather in the fields. This is especially important for farms that grow vegetables and fruits. Sensors allow to prepare in time for a critical temperature change and calculate irrigation. Also, weather data allows to predict the appearance of diseases and parasites on plants. This makes it possible to determine when it is best to apply pesticides - and whether they should be applied at all. For example, a farmer twice a year contributes protective equipment to a field with potatoes so that the crop does not die. On this he spends $ 20 thousand per season. He does not know for sure whether parasites will appear or not, he uses chemicals just in case. Using sensors in the field of this overspending can be avoided.;

- the relief: this factor affects the distribution of water and nutrients in the soil, which determine the fertility of the plots. Crop and topography correlate with each other: high yields are more often in the lowlands, and medium and low yields are in higher elevations. In most countries of Europe and the USA, a farmer can obtain data on the relief of his field in state cartographic services. Another option is to order mapping of the relief in a private company. The terrain model is built by shooting from a drone, according to the lidar survey, or according to the results of driving around fields on an ATV with special equipment. Based on this information, maps are made for applying seeds and fertilizers by zone, which allows the farmer to spend money and time efficiently. An example of a height distribution and associated field productivity is shown in Figure 2.2 [31];

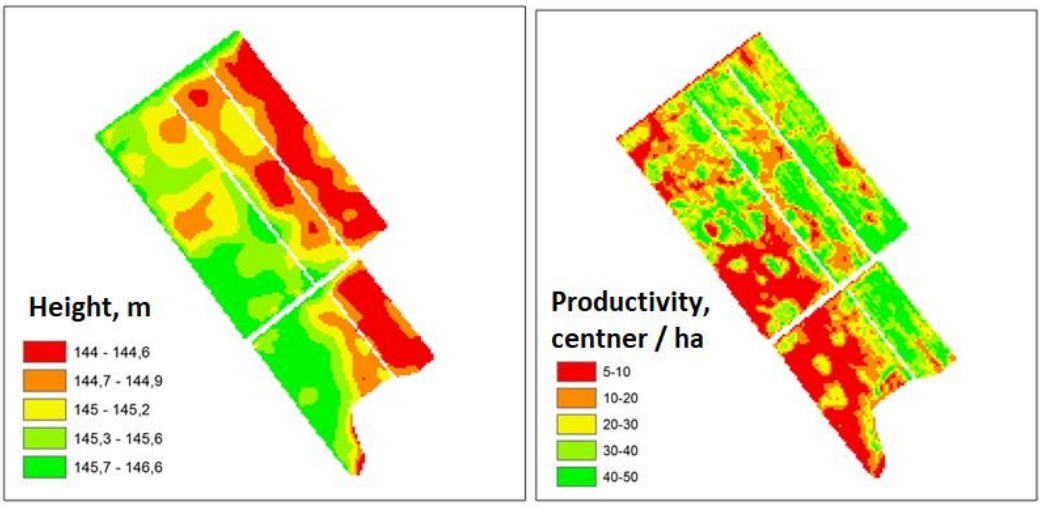

Figure 2.2: Map of heights and field yields

2.2 Technologies and technical means of precision farming

In order to collect the above data, certain technologies and hardware and software tools, such as [16]:

1) GPS/GNSS

Almost immediately, as agriculture gained access to GPS in the 1990s, operators and manufacturers found ways to use this technology to simplify fieldwork. As a result, GPS trackers appeared that help track the operation of machine operators, and antennas, thanks to which it is possible to more accurately process fields.

In turn, GNSS makes this technology universal. It covers all existing satellite positioning systems: GPS, Galileo and GLONASS.

2) Mobile devices

The next major innovation of the last 20 years is the development of mobile devices. “Without a cell phone, we probably would still be sitting at the barn and waiting for someone to come to us and fix our things,” says Illinois farmer John Reifsteck. Now phones have evolved technologically into smartphones and tablets [20].

Thanks to mobile devices, one can use various applications that will help in the agribusiness: trace elements calculators, weather forecasts, field maps and GPS navigation.

3) The robotics

Robots solve various problems in agriculture, such as planting crops, monitoring the condition of crops and even cutting weeds in the garden. For example, Knize engineers created an autonomous grain basket system, thanks to which the cart follows the harvester across the field at a safe distance. Another interesting project is Fendt MARS, the essence of which is the development of small and light robots for sowing corn, which consume little energy, are controlled from a tablet and transfer data to the cloud storage. The Fendt MARS Benefit Diagram is shown in Figure 2.3 [32].

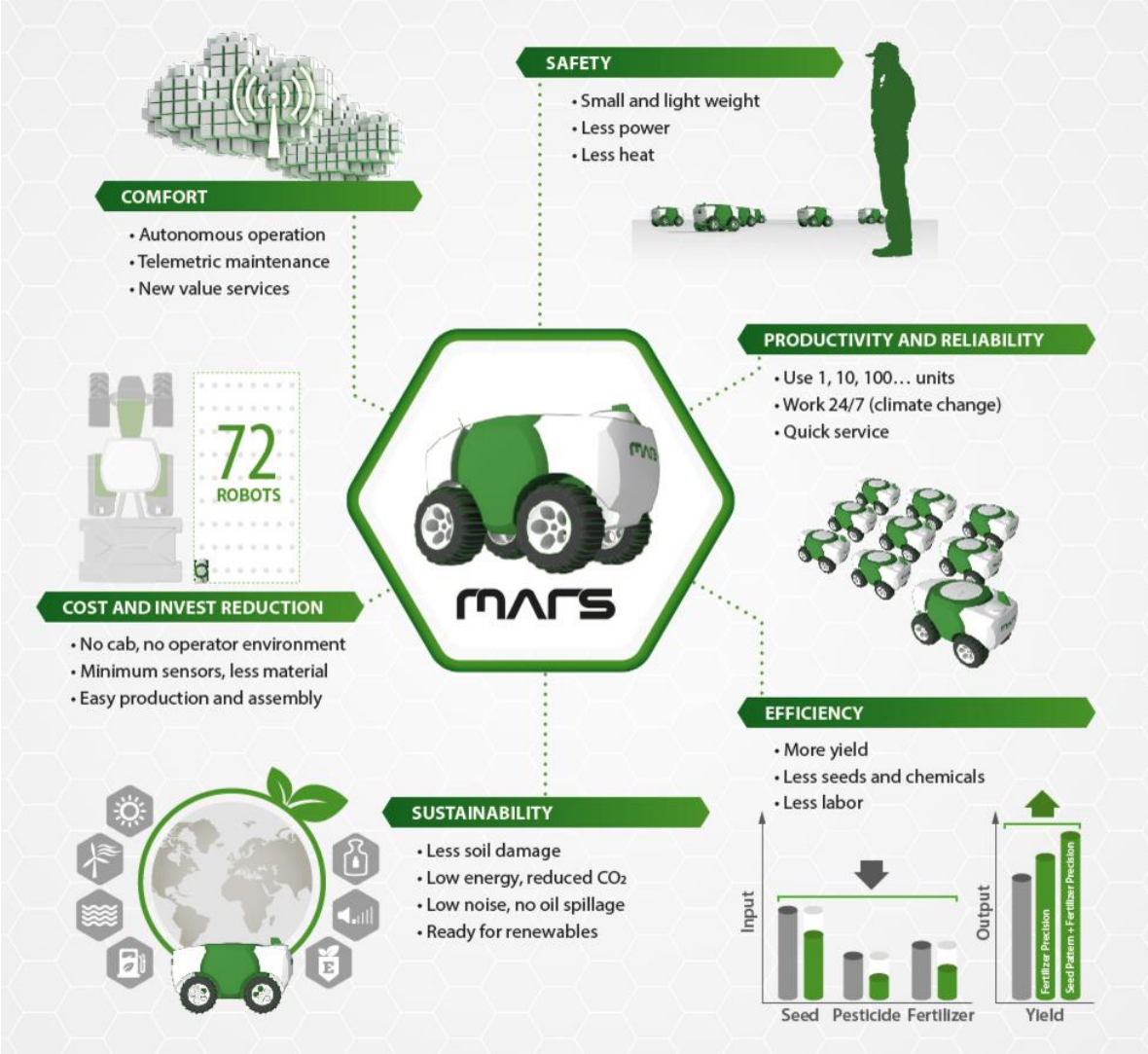

Figure 2.3: Fendt MARS benefits

4) Irrigation systems

Irrigation innovations are growing rapidly because farmers often face water shortages due to drought. Modern technological solutions allow you to remotely monitor and control almost every aspect of the irrigation process. They save water, time, fuel and vehicle wear.

Separately, a VRI system can be distinguished, thanks to which soil moisture is monitored, a weather forecast is made, after which it is possible to effectively distribute the required amount of water over the fields [18].

5) The Internet of Things

The concept of connecting devices to each other using a network. This technology is used, for example, in smart homes. This approach is also applied in agriculture, when data from field sensors and satellite monitoring are sent to the same source.

6) Sensors

Wireless sensors are used in PF to collect data on moisture, compaction, soil fertility, climate change and other important aspects. This information helps to better allocate farmer resources. By the way, there is even a sensor for determining the level of CO2 in the air, which was developed in Ukraine.

7) The differentiated seeding rate

Thanks to lengthy field studies, it was recorded that due to an increase in the seeding rate in productive areas of the fields and their decrease in less fertile zones, the yield of crops increases.

To achieve the best results using the technology of interchangeable seeding rates, it is necessary to determine zones that differ in the type of soil, topography, and irrigation, as well as the history of yield over the past few years.

8) Monitoring of weather changes

Perhaps one of the most unpredictable variables in agriculture is the weather. But now you can prepare for its changes using weather modeling. For example, the ClearAg system

is an agricultural platform that predicts and models possible weather conditions and then indicates how it will affect irrigation, soil quality, and yield. The ClearAg platform user interface is shown in Figure 2.4 [33].

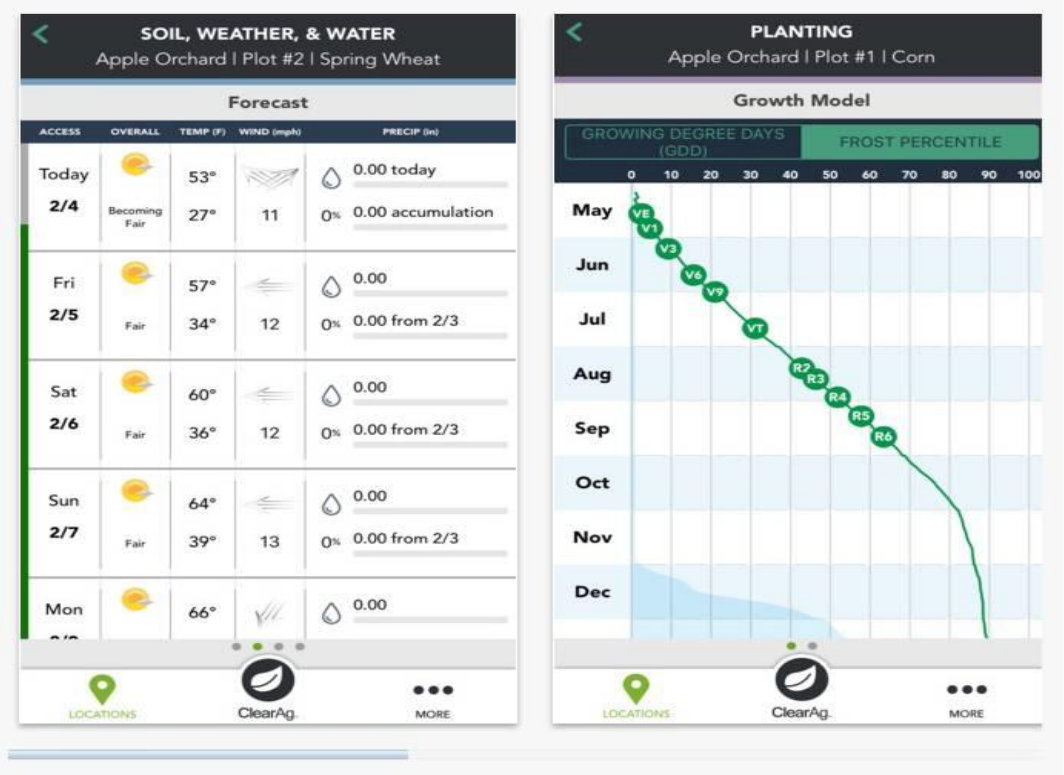

Figure 2.4: ClearAg user interface

It must be kept in mind that weather forecast applications are the most popular among German farmers.

9) Monitoring of the amount of nitrogen in the soil

Recently, applications for calculating and monitoring the number of trace elements in soil are gaining popularity. In particular, the so-called nitrogen calculators are of particular interest to farmers, because the nitrogen cycle is very complicated and managing it is not an easy task. An example of such an application is Adapt-N, thanks to which it is possible to trace the level of nitrogen in the soil and form a nitrogen map of the field (Figure 2.5).

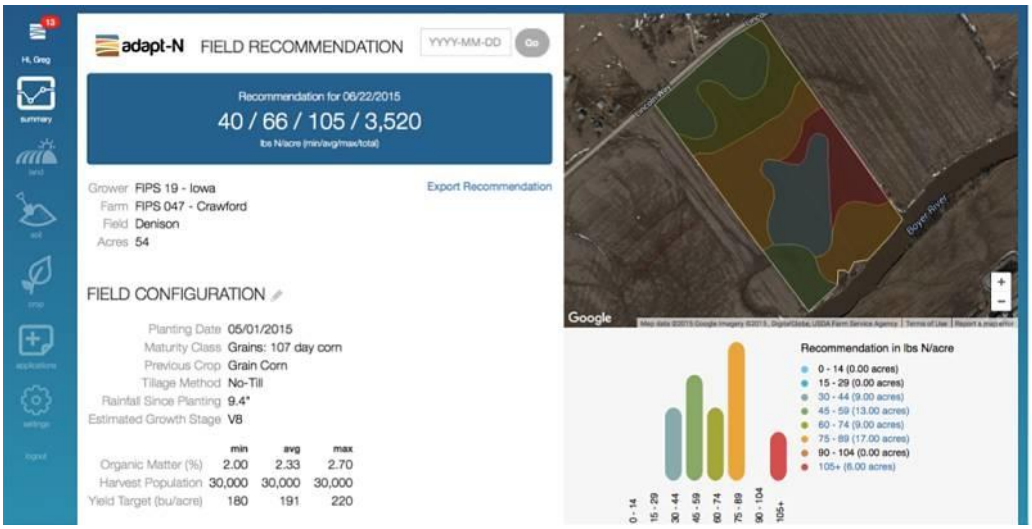

Figure 2.5: Adapt-N user interface

10) The standardization

The issue of compatibility of components from various equipment manufacturers is relevant now. The first step towards solving this problem was the formation of the Agricultural Industry Electronics Foundation. It includes more than 170 companies and organizations that are actively collaborating to bring standards into action. The most successful AEF project was ISOBUS - a protocol for compatibility of tractors and outboard equipment from different manufacturers.

Thus, using these technologies separately, the effect of their implementation will be insignificant. But if you combine these technological solutions, then you can confidently call yourself an “exact farmer”, as well as increase the yield of each field several times. The principle of ISOBUS operation is shown in Figure 2.6 [35].

Figure 2.6: ISOBUS working principle

2.3 Social events as an engine of scientific progress

In order for progress in the field of PF not to stand still, it is necessary to support research and development of new systems. The Agri-Innovation Summit 2017 (AIS 2017) took place in Oeiras, Portugal on 11 and 12 October 2017. The event gathered a diversified audience of over 500 people, including farmers, advisers, researchers, NGOs, RDP managing authorities, NRNs, LEADER Local Action Groups (LAGs), rural businesses and start-ups. It was aimed to:

- promote cross-fertilisation between multi-actor innovation initiatives;
- raise awareness about innovation and digitisation opportunities for agriculture;
- provide input for EU innovation policies for agriculture and rural areas after 2020.

In two days, 119 innovative projects were presented in three sections:

1) Resource use (adaptation and mitigation)
    a. resource efficiency / water and energy / circular economy / fertilization
    b. agro-environment and climate change
    c. genetic resources / climate change adaptation
    d. plant protection / animal health and welfare

2) Management of farming, food and forestry systems & valorization of the territory
    a. agriculture products and food processing
    b. forest management and fire prevention
    c. valorization of the territory and use of endogenous resources
    d. sustainable management
3) Agriculture 4.0 and rural development
    a. digital entrepreneurship in rural areas
    b. PF
    c. management tools to support farmers' decision-making
    d. robotics / mechanization

Table 2.1 presents data on projects considered under the PF section.

The use of PF can definitely be called the future of the agricultural industry [17]. However, such systems often require farmers to invest heavily, which is not always available on small farms. Also, buying a system is not all. You must either invest in the training of your worker, who will subsequently work with this system and act as the operator of the necessary equipment, or hire a specialist to work with this equipment and software. However, there are more accessible instruments for the complexity of mastering and the cost of investing resources that will help farmers monitor the status of their farms.

Table 2.1: The Agri-Innovation Summit 2017 projects

Project title	**Practical problem**	**Parameters project**			
		Objectives	**Expected results**	**Results so far/first lessons**	**Who will benefit**
Control of additional water use in crop production - situational, site-specific and automated (Precision Irrigation) [18]	In the federal state of Brandenburg (Germany), irrigation of arable land is a measure to maintain agricultural value despite decreasing summer rainfalls. To avoid over-using the available water resources, however, a precise irrigation control needs to be developed and tested under local conditions.	Development of an economic solution for site-specific irrigation, which takes into account the actual water need of the crops. The potential of infrared thermography for precision irrigation control is evaluated in addition to traditional soil based approaches.	An existing model for steering irrigation is adapted to site-specific irrigation control. The model results are automatically transferred to the steering unit of centre pivots to help save labour resources. Since the steering approach is applied at farm scale and evaluated in cost-benefit analyses, we shall be able to develop a practical solution for precision irrigation for local farmers.	Two existing center pivots were modified to enable the site-specific application of irrigation water. We derived soil-based irrigation management zones and controlled the timing and amount of irrigation water with an offline prototype of our steering model. Moreover, we acquired aerial images at the infrared spectrum to derive crop canopy temperatures and to calculate crop water stress indices.	Farmers, governmental and non-governmental institutions, scientists.
Data assimilation from soil-crop-climate sensor network in IRRINET DSS	There is an increasing interest in the adoption of sensors to monitor the soilplant- water system from growers and producer organizations. Nevertheless, data integration and accessibility, as well as a real benefit for farmers in	Integration of soil, crop and environmental sensors within the IRRINET regional DSS for irrigation management, which will allow farmers to benefit	Integrated environmental data from private sensors and weather stations to the IRRINET DSS. Creation of links between IRRINET and weather and soil sensors located in pilot farms. Validation of the IRRINET	Six farms with private sensors network already integrated into IRRINET dss. Protocols for data integration and validation is in testing and calibration phase. First year of field trials almost completed.	Farmers with irrigated crops in Emilia-Romagna Region.

Project title	Practical problem	Parameters project			
		Objectives	**Expected results**	**Results so far/first lessons**	**Who will benefit**
	terms of water savings are still missing.	from an increased reliability of the monitored data and to automatize data integration and interactions in the IRRINET portal.	irrigation scheduling advices based on the irrigation needs identified in farms. Protocols for validation and integration in IRRINET of sensors data.		
Evaluation of innovative agronomic strategy to improve precision in managing biotic and abiotic stress in fruit orchard	Variability in fruit orchard is often very high. Extensive fruit farms operators are often trained to use high level of chemical inputs to correct trees deficiencies, without considering the orchard variability. Such management has high impact in costs and environment pollution.	The objective is to better understand how to evaluate the potential yield in an orchard and to map it with a geo-statistical significance in order to connect it to biotic and/or abiotic trees stress, thus enabling to plan an agronomic strategy to avoid these stresses.	We expect to improve production levels, while reducing chemical inputs.	We are still working on geo-referenced data with geo-statistical analysis to understand which is best suited to be used to create prescription maps and plan targeted interventions next year.	Fruit growers.
Increasing the viability of sown biodiverse pastures	Most Portuguese pastures are poor grasslands on degraded soils. Some farmers invest in improved	Optimize the use of fertilizers in sown biodiverse pastures by	Technological method for obtaining high-resolution phosphate fertilization	The first activities will be selection of experimental plots and obtaining satellite data. 3D terrain models will	Farmers will benefit from optimizing pasture fertilization,

Project title	Practical problem	Parameters project			
		Objectives	**Expected results**	**Results so far/first lessons**	**Who will benefit**
through optimization of phosphate fertilization	and fertilized grasslands, namely sown biodiverse pastures, however their economic viability is threatened by production costs, namely phosphate fertilizers.	using remote data sensing for evaluating pasture nutrient needs and using Variable Rate Technology for fertilizer distribution.	prescription maps. Obtain phosphate fertilization prescriptions in order to optimize pasture productivity and to reduce production costs. Establish a service to farmers in improving the economic viability of sown biodiverse pasture.	be obtained as well as soil measurements with optic sensors, capacitance sensors and electric conductivity sensors. Soil and plant samples will also be analysed. UAV flights will be performed, collecting multispectral images and correlating them with soil and vegetation measurements.	improving its productivity and decreasing production costs.
PARRA - Integrated platform for monitoring and evaluating vine health (automatic detection of flavescence dorée: work on cost optimisation of data collection etc.)	A vineyard, one of the most important crops in Portugal, currently faces the threat of deceases with strong economical impact, like Flavescence Dorée. The early detection of this quarantine disease in large areas of vineyards, within a close time frame, will contribute to strongly reduce its impact.	To develop a solution, exploring drones and other platforms, to collect vineyard data and develop automatic analysis algorithms to identify the presence of early stage symptoms of the Flavescence Dorée. The project will explore hyperspectral sensors and will evaluate its results in real scenarios.	To improve temporal responsiveness in disease detection and containment; The reduction of operational costs of inspection and verification of vineyard grubbing actions and the reduction of production losses and yield of the vine due to this type of disease; PARRA's approach also ensures the development of a scalable solution, both in terms of size and usage application by different stakeholders	The Project has collected multiple vineyard samples, performed laboratorial analysis and collected hyperspectral data in order to characterise the symptoms; laboratory observation established the minimum number of samples; Symptoms variability due to climate and location of the disease; A samples collection protocol was developed to ensure proper handling of the bio and hyperspectral data.	Vineyards producers; Wine makers; Agricultural enterprises; Government agencies and laboratories; Inspection agents

Project title	Practical problem	Parameters project			
		Objectives	**Expected results**	**Results so far/first lessons**	**Who will benefit**
SMARTCROP - Sustainable competitiveness	The promoter, a maize producer in Vale do Tejo, seeks to implement a Smart agricultural production process based on the collection, compilation, treatment and data analysis, improving competitiveness with an agricultural intervention at the right time, in the right place, with the right amount.	Improve decision making process on farming management. Test and fit new technologies and equipment. Improve inputs and electric power efficiency. Integrate irrigation management with power meters. Identify critical points of gases emissions. Evaluate the cost/benefit of the new management system.	Application of new farm management tools to increase yields, reduce inputs and carbon emissions. Develop an online platform to support irrigation management and power usage.	Integration of Precision Agriculture methods and technologies. Implementation of data communication system between different tractors terminal systems and head office computers. Creation of a new sustainability indicator dashboard. Improvement of Irristrat™ online platform and integration of a new module for the efficient use of energy	Farmers seeking to improve and rethink farm management in a more conscious way to apply inputs, water and power usage.
SMARTFARMING - Precision integrated system for irrigated farming efficiency and sustainability	PF is getting common among the farmers, and they have now precise and valuable information about their crops (soil, crops and applied water/fertilizing) in each point of the field. How could we use this information, and low cost technology, on a precise irrigation of a pivot?	Gain competences on Variable Rate Irrigation, with clear benefits in the efficient use of resources, especially irrigation water, soil conservation and energy, regarding the maximum crop yield, ecosystem sustainability	Based on the integration of the different data collected from wide range of sources, it will be created a high-value precision output in each moment of the season. This way irrigation precision system will result on a decision support system controlled by a skilled specialist, uploaded to Variable	From our field experience on the last 4 years, we realized that the pivots irrigation is not efficient at all, due to its homogeneous water displacement on heterogeneous fields. There is starting to appear low-cost technology to technically solve the problem, but the farmers need to join the electronics	Farmers that are already using irrigation pivots and the ones that will be reconverted to irrigation and install new pivots.

Project title	Practical problem	Parameters project			
		Objectives	**Expected results**	**Results so far/first lessons**	**Who will benefit**
		and competitiveness of agricultural sector.	Rate equipment in the field (implemented with minor investment on farmer's irrigation equipment).	to the agronomics, to know "how to" do it each moment of the season.	
WATER4EVER - Optimization of irrigation to conserve soil and aquatic resources	Agriculture is the largest consumer of water and a key source of diffuse pollution, promoting eutrophication of water bodies, with associated biodiversity loss. Regulated Deficit Irrigation is part of the solution by decreasing water and nutrient surpluses, thus improving management practices.	To establish a direct link between water quality and specific agricultural practices by combining EO, in-situ measuring, hydrological and crop models to develop tools for (i) supporting regulated deficit irrigation, and (ii) assessing the benefits for hydrological resources at the catchment scale.	The following results are expected: (1) to minimize diffuse agriculture pollution through improved irrigation management techniques; (2) to develop low cost sensors and new remote sensing approaches for plot scale monitoring; and (3) to develop models as interdisciplinary tools to optimize irrigation and fertilization practices and to link spatial and temporal scales	The project is still in its initial phase. Consortium members already have all sensors and models necessary to set up the experiments, which will now be improved based on the partners' experiences and following a multidisciplinary approach.	Farmers, Agronomists, Water Agencies

2.4 A literature review on the agricultural modeling

In order to understand what methods of analysis are used by scientists at the moment, it is necessary to review the works devoted to studies of crops growing in Portugal (mainly olive and chestnut), as well as the conditions for the growth and evolution of these crops (soil type, rainfall, temperature, application of minerals and fertilizers, etc.). Next, an analysis will be made of ten articles by scientists who conducted similar studies from the following organizations:

- Mountain Research Centre, Polytechnic Institute of Bragança, Bragança, Portugal;
- Centre for the Research and Technology of Agro-Environmental and Biological Sciences, University of Trás-os-Montes e Alto Douro, Vila Real, Portugal;
- Regional Department of Agriculture and Fisheries, Mirandela, Portugal;
- Universidade Tecnológica Federal do Paraná, Campus Dois Vizinhos, Paraná, Brazil;
- Atlanlusi Europe, Lda., Batalha, Portugal;
- ARBOREA, Associação Agro-Florestal e Ambiental da Terra Fria Transmontana, Vinhais, Portugal.

I want to thank them for their research.

The studies under analysis were conducted in the vicinity of Braganca County. It should be noted that the type of soil in this region is almost always acidic (leptosol derived from schist [4]; Eutric Cambisols [5]; Eutric Regosol [6]; loam texture dystric Regosols [7]; Leptosols originating from a bed rock of schist [8]; Umbric Cambisol derived from granite and Umbric Leptosol derived from basic rocks [9]). A more detailed composition for each study is presented below:

- the soil is loamy textured (54% sand, 25% silt, and 21% clay), slightly acid (pH = 6.0) and low in organic carbon (0.8 g kg−1, Walkley-Black). Phosphorus and

potassium levels are respectively high and very high as determined by Egner-Rhiem method. Both exchangeable calcium and magnesium are high, respectively, 11.8 and 6.4 cmolc kg−1 as determined by the ammonium acetate method [5];

- the texture is sand, with 88.4% sand, 9.8% silt, and 1.8% clay [9];
- the soil is a loamy-sand textured with 72.4% sand, 21.0% silt, and 6.6% clay [9].

Chestnut plantations were investigated for the effect of liming, the use of nitrogen, phosphorus, potassium and boron on young seedlings. Different combinations of these fertilizers had different results [1]. The nutritional status of chestnut groves was diagnosed using soil and leaf analysis [2], and the reaction of chestnuts to organ mineral fertilizers and fertilizers with controlled release in dry farming was studied [9].

First of all, it is necessary to study the conclusions from the source [2] since this will help to draw up a more complete picture of the nutrients that are already in the soil (deficit, surplus), and to understand what fertilizers should be applied to this area of chestnut plantation.

The organic matter content of the soils where chestnut trees are grown in the Bragança district, determined from 1041 soil samples, was low for 82% of the samples (Fig. 1). The majority of these soils were very acid (13%) or acid (73%). Extractable PAL levels were very low (48%) or low (26%) and extractable KAL levels were medium (39%) to high (44%). Exchangeable Ca and Mg were at very low (48% and 37%, respectively) or low (44% and 39%) levels probably reflecting soil acidity. A more detailed study of the percentage of soil in Figure 2.7.

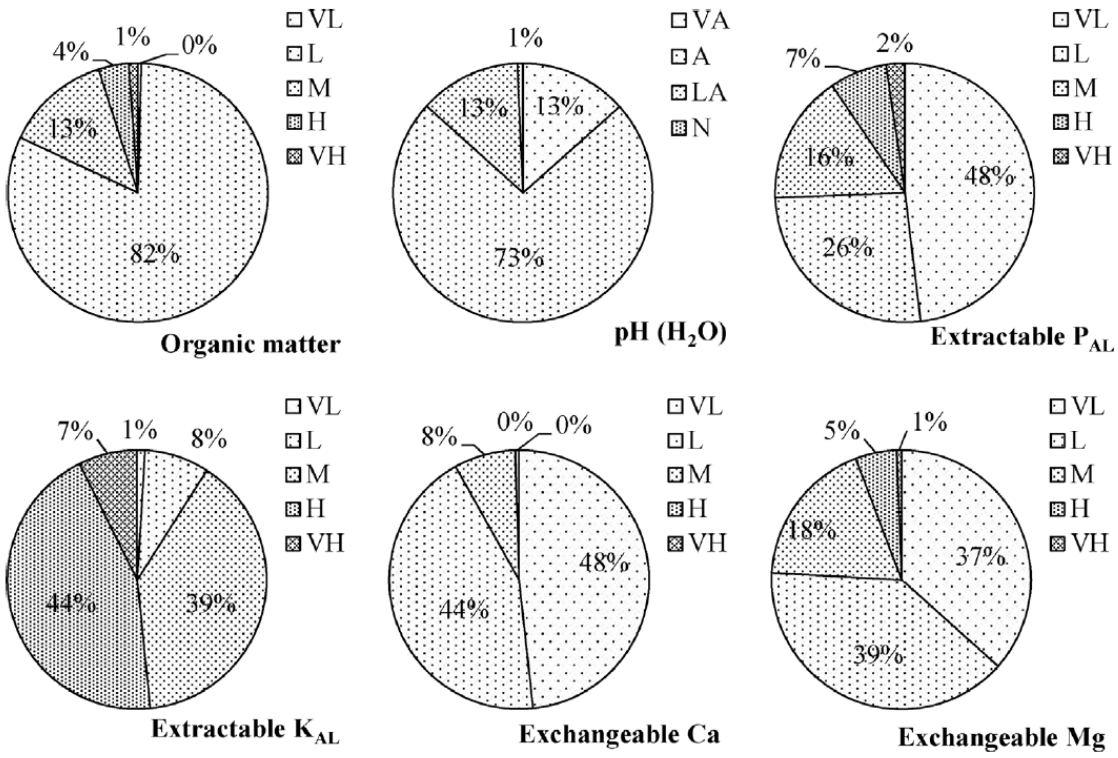

Figure 2.7: Relative frequency of soil fertility parameter levels: organic matter (VL, very low; L, low; M, medium; H, high; VH, very high), pH(H2O) (VA, very acid; A, acid; LA, low acid; N, neutral), extractable PAL and KAL and exchangeable Ca and Mg (VL, very low; L, low; M, medium; H, high; VH, very high) from a population of 1041 soil samples voluntary delivered by farmers between 2010 and 2016 from chestnut groves of Bragança district, NE Portugal.

Results from 198 leaf samples voluntarily delivered to the laboratory by farmers of Bragança district showed 63% of orchards with N levels below the sufficiency range established for chestnut (Figure 2.8). Leaf P and K concentrations below the adequate range were found respectively in 18% and 34% of the samples and leaf Ca and Mg concentrations below the lower limit of their sufficiency ranges were respectively 19% and 21%. Boron seems to be the most deficient element among micronutrients with 40% of the records below the respective sufficiency range. Records above the higher limits of the sufficiency ranges were residual for N, P and K. Records of leaf Ca and Mg concentrations above the higher limit of the sufficiency range were 8% for both elements. Among micronutrients, records above the sufficiency range varied between 3% for Cu and 12% for Mn.

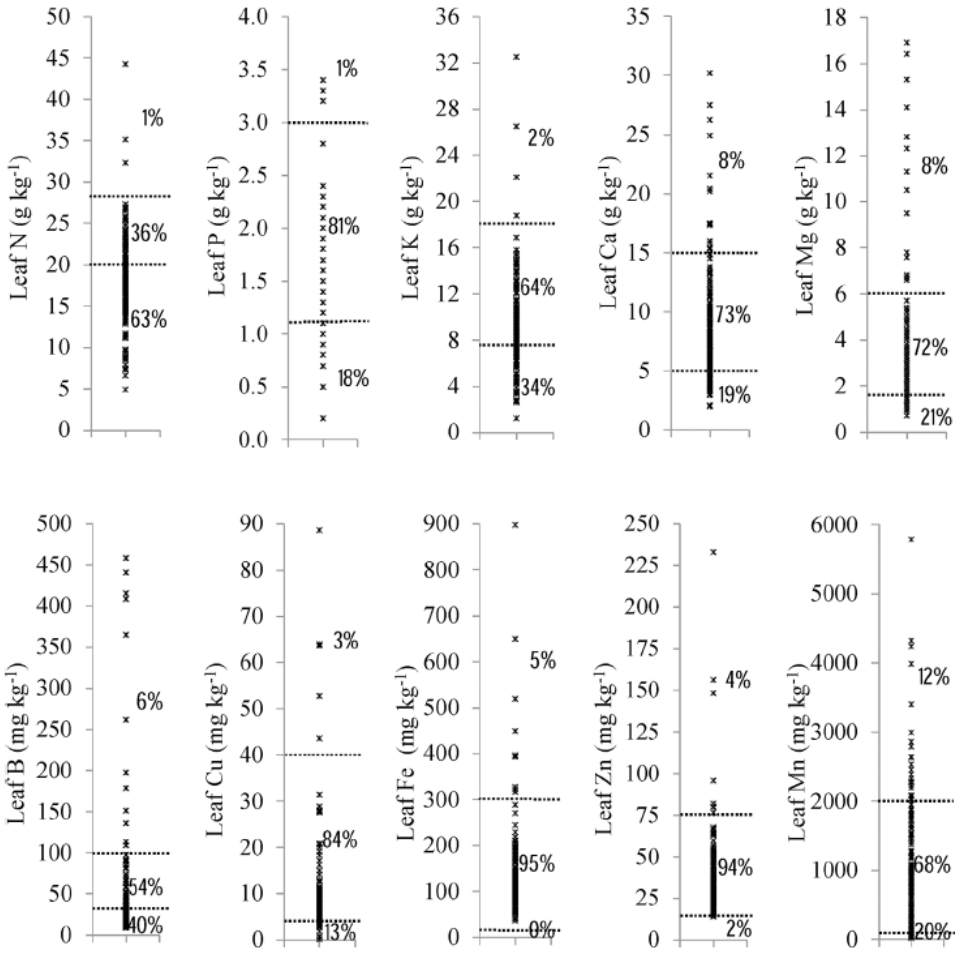

Figure 2.8: Classification of the concentration of nutrients in the leaves into three sufficiency ranges: excess (above the upper dashed line); adequate (between the two dashed lines); and deficient (below the bottom dashed line), from a population of 198 leaf samples taken between 2010 and 2016 from chestnut groves of Bragança district, NE Portugal.

In the course of further research, the following conclusions were made:

Nitrogen nutritional status of chestnut trees seems to be in a very unsatisfactory status and national and international data did not reveal this. It seems that N will have to be applied regularly in chestnut groves as in all other crops including fruit tree species.

Soil testing seems to underestimate the soil P availability, suggesting a greater shortage of the nutrient than indicated by leaf analysis. Chestnut tree fertilization strategy has been very focused on this nutrient, perhaps as the soil testing is currently more usual among farmers than leaf analyses, although it has emerged in this study as the least problematic among the macronutrients, judging from the results of leaf analysis.

Regarding K, soil testing underestimated the lack of the nutrient, incorrectly indicating a favourable situation of K levels in soils that leaf analysis did not corroborate. The situation is delicate in that it can lead farmers to apply less K than necessary. If the problem of K nutrition in chestnut is a difficulty in K uptake, due to the extreme dryness of the soil in the summer, it may be necessary to implement foliar fertilization strategies.

Soil acidity seems to have to be corrected with dolomitic limestone. The deficiencies of Ca and Mg seem to be frequent, although more accentuated in Ca. The application of limestone may lead to the need to pay more attention to K due to the increased risk of K deficiency by ionic antagonism with Ca and Mg. In contrast, pH increase may enhance soil P availability, further reducing the problem of P.

Among the micronutrients, B deserves greater attention, due to the frequency in which it appeared as deficient but also due to the risk of toxicity with its excessive use as a fertilizer.

In chestnut, as generally in fruit tree crops, soil analysis alone is not a satisfactory guide for fertilizer recommendations. As shown in this paper [2], it is difficult to establish relationships between soil properties, including soil nutrient levels, and plant nutritional status. It seems urgent in these orchards to establish regular plant analysis programs and to apply the nutrients which are deficient as fertilizers. It would also be a useful task for the scientific community to establish improved sufficiency ranges for this important crop, as there is little work reporting results of leaf analysis.

Now we can begin to study the effect of fertilizers (N, P, K, B) and liming on chestnut groves [1]. The experiment was arranged in a randomized block design with three replicates. The six fertilizer treatments were: 1) lime plus N, P, K, and B (LNPKB); 2) N, P, K, and B (NPKB); 3) P, K, and B (–NPKB); 4) N, K, and B (N–PKB); 5) N, P, and B (NP–KB); and 6) N, P, and K (NPK–B. Each experimental unit was composed of a group of 10 trees. The plants within each experimental unit were planted at 1.5 m in the row and 9 m between rows. For each treatment three replicates (three groups of 10 plants) were included. The total area of each experimental unit was 135 m2. However, the fertilizers were applied in a reduced area of 60 m2 per experimental unit (15 m of the row and 4 m laterally, 2 m on each side of the planting line). Lime was applied only once at a rate of 2000 kg ha–1 considering only the 60 m2 mentioned above, and before the establishment of the crop. A product with 80% CaCO3 and 5% MgCO3 was used. Phosphorus and K

were applied annually in early spring at the rate of 100 kg ha–1 (expressed as P2O5 and K2O), as 18% superphosphate (P2O5) and 60% potassium chloride (K2O). Nitrogen was also applied every year in early spring at a rate of 30 kg N ha–1 as ammonium nitrate 20.5% N (50% NH4 +, 50% NO3 –). Boron was also applied annually at a rate of 3 kg ha–1 of B as borax (11% B). In 2016 the rate of B was reduced to 2 kg ha–1). It is also necessary to take into account the temperature data presented in Figure 2.9 and the percentage indicators of soil parameters selected at the time the tests began (Table 2.2).

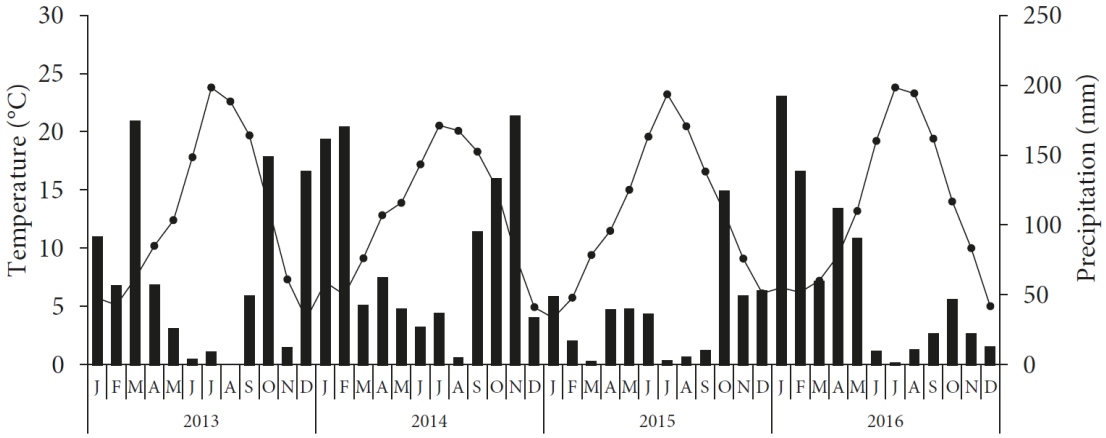

Figure 2.9: Average monthly temperature and accumulated precipitation recorded during the experimental period (2013–2016) at the weather station of Qta de Sta Apolónia, Bragança.

Table 2.2 Selected properties of the soil sampled shortly before the trials started at a depth of 0–20 cm.

Soil properties		Soil properties	
Clay (%)	11.8	Extractable P (mg kg^{-1})[d]	15.48
Silt (%)	26.3	Extractable K (mg kg^{-1})[d]	86.00
Sand (%)	61.9	Exchangeable bases[e]	
Texture (USDA)	Loam	Ca ($cmol_c$ kg^{-1})	5.69
pH_{H2O}	5.50	Mg ($cmol_c$ kg^{-1})	2.07
pH_{KCl}	4.30	K ($cmol_c$ kg^{-1})	0.23
Oxidizable C (g kg^{-1})[a]	0.87	Na ($cmol_c$ kg^{-1})	0.36
Total organic C (g kg^{-1})[b]	2.34	Exchangeable acidity ($cmol_c$ kg^{-1})	0.10
Extractable B (mg kg^{-1})[c]	0.25	Exchangeable Al ($cmol_c$ kg^{-1})	0.10

a Walkley–Black; b incineration; c azomethine H; d ammonium-lactate; e ammonium-acetate, pH 7.

The findings of this study are given below:

Liming a soil with initial pH of 5.5 did not significantly influence tree crop growth, showing that chestnut might be a tolerant crop to moderate soil acidity. The treatments NP–KB and NPK–B gave shorter trees with thinner trunks, which made K and B the most limiting nutrients for chestnut growth under the conditions of this experiment. In these treatments, leaf K and B were respectively found in the deficiency range of each nutrient. The result was not expected since initial soil K levels were found to be at a medium level. Leaf N was found oscillating below and above the lower limit of the adequate range in the –NPKB treatment, giving the indication that N was the third most limiting nutrient in this experiment. SPAD readings and the NDVI were able to diagnose the low soil N availability in the –NPKB treatment but not the decreased growth of the trees associated with the NP–KB and NPK–B treatments. The dark adaptation protocols of FV/FM, FV/F0, and the advanced OJIP measurements failed to diagnose any of the nutrient stresses detected by leaf analysis and growth reduction.

Researchers working on the issue of the reaction of chestnuts to the introduction of organic fertilizers and fertilizers with controlled release in the conditions of rainfed farming [9], made the following conclusions:

1) the fertilizer most concentrated in N and B and the one with the highest percentage of N protected gave higher leaf N and B concentrations and nut yield, which highlights the importance of N and B in these agroecosystems;

2) fertilization increased soil organic matter particularly in the surface layer of the cover cropped orchard, an aspect of great relevance to the sustainability of these agroecosystems;

3) there was not enough experimental evidence to suggest that organic materials and fertilizers with mechanisms for controlled-release provided on their own major benefits to the trees in the four years of the trials;

4) the results suggest that long term studies should be performed to better clarify the benefits of using slow-release fertilizers in large trees such as chestnut. Meanwhile, farmers should be cautious about the use of fertilizers with mechanisms of nutrient protection if they are substantially more expensive than the conventional water-soluble fertilizers.

But research does not stop only at chestnut plantations. For comparison, the works in which the studied objects are olive trees were investigated. In the course of research, it was determined that the introduction of boron directly into the soil gives a greater remobilization of this substance in plant tissues than fertilizer by spraying the crowns of trees [6].

The effect of nitrogen fertilizers on the frequency of fruit flies, spotting of olive leaves and anthracnose in two varieties of olives grown under rainfed conditions was also studied. In this study the effect of nitrogen fertilization (0, 20, 40 and 120 kg N ha−1) on the incidence of olive fruit fly (Bactrocera oleae), olive leaf spot (Spilocaea oleaginea) and olive anthracnose (Colletotrichum gloeosporioides) was evaluated in two olive orchards of the cultivars ‘Cobrançosa’ and ‘Madural’. The study took place in 2017 and 2018 in Lombo, NE Portugal. Nitrogen fertilization significantly increased olive yield and nitrogen concentrations in plant tissues and induced a delay in fruit maturity. The incidence of the olive fruit fly was significantly lower in the more fertilized treatments of both cultivars. However, ‘Madural’ showed itself to be more susceptible than ‘Cobrançosa’ to olive fruit fly. The effect of nitrogen on the reduction of the incidence of the olive fruit fly was attributed to the delay caused in fruit maturation, which might have desynchronized the attractiveness of the fruits for insects to lay their eggs on their flight curve. In contrast, olive leaf spot and olive anthracnose were not influenced by nitrogen fertilization. As a result, nutrient management in olive groves must balance carefully the requirements of economic rationality with environmental preservation, particularly with regard to the harmful relationship between the use of excessive nitrogen rates and several adverse environmental effects. [8]

In gardening, it is always necessary to pay attention to such factor as pruning. It can significantly affect the productivity of plantations and the ability of trees to absorb fertilizers. For olive trees, pruning is considered as a mean of increasing yields. The olive trees showed a high plasticity or tolerance to pruning, since olive yield did not decrease in response to light or moderate pruning regimes. It seems that it is possible to carry out light or moderate pruning to achieve several objectives of orchard management without significant loss of production. The results also showed that if pruning is done under a hard regime it should only be performed after a heavy crop. In addition, if done under a light regime, pruning can also reduce the alternate-year bearing behavior of the olive tree.

The object of the next study, which studied the effect of the use of soil and foliar fertilizers with nitrogen and boron, was an almond plantation grown under rain-fed conditions [7]. Below are the conclusions obtained during the experiments:

Leaf sprays of N and B, including post-harvest applications, failed to increase crop nutritional status and kernel yield. In these low-input farming systems, the application of fertilizers to the soil seems to be simpler and a more effective way of managing the nutritional status of the orchard.

To control the processes taking place in the garden or in the field, it is very important to observe the time limits of the stages of sowing, fertilizing, processing, etc. In scientific work [5], a quantitative assessment of oilseed rape yield losses caused by a delay in sowing in the Mediterranean environment, and the possibility of preventing losses when nitrogen was introduced into the soil as a fertilizer were studied.

The experiment was arranged in a split-plot design with the dates of sowing as the main-plots and N rates the sub-plots. Rapeseed recovered 128 to 212 kg N hm-2 before top-dress N application in late winter if sown before the last week of September. Seed yield was very dependent on the date of sowing, varying from 3.4 to 6.2 Mg hm-2 on the first sowing date in September to 0.3 to 1.0 Mg hm-2 on the last sowing date in November. The daily loss in seed production was 68.9 kg hm-2 (or 482.3 kg hm-2 per week) or 1.53 % (or 10.7 % per week). N rate significantly increased seed yield within each sowing date but did not allow latesowed plants to regain the productivity levels of those sown earlier. Apparent N recovery and agronomic N efficiency were particularly dependent on the growing conditions associated to different sowing dates.

A very important factor in the management of the garden when applying fertilizers are the sufficiency ranges for a particular species. On the example of peppermint, the ranges of sufficiency and removal of nutrients obtained in field and pot experiments with fertilizers were studied [5].

Peppermint is an important aromatic and medicinal plant used across the world in pharmaceutical, cosmetic and food industries. However, there is a lack of agronomic research on this crop which hinders the implementation of best agricultural practice at farm level. Plant analysis, for instance, cannot be used as a tool to implement a suitable

fertilizer recommendation program, since sufficiency ranges and crop nutrient removals have not yet been established.

In the course of the experiments, it was possible to evaluate the reaction of peppermint to various levels of nitrogen (N), phosphorus (P), potassium (K) and boron (B) in order to establish the ranges of sufficiency from macro, micronutrients and SPAD reading, as well as to establish the absorption of nutrients crops in aerial biomass. Field trials and pot experiments were conducted from 2013 to 2015 in a wide range of conditions involving 12 N, P, K or B fertilizer trials and a total of 48 cuts of biomass. Nitrogen fertilization increased dry matter yield of peppermint on the vast majority of sampling dates. In contrast, P, K, or B did not produce a significant effect on dry matter yield in any of the experiments. The sufficiency ranges set for macronutrients N, P, K, Ca and Mg are respectively 32.0 – 42.0, 1.2 – 4.5, 10.0 – 30.0, 7.0 – 23.0, and 4.0 – 10.0 g kg−1. Those for micronutrients B, copper (Cu), iron (Fe), zinc (Zn), and manganese (Mn) are respectively 20 – 200, 5 – 25, 100 – 600, 25 – 300, and 30 – 200mg kg−1. Sufficiency range for SPAD-readings is 45 – 50 SPAD units. All these ranges were established for the commercial harvesting date. The amounts of N, P, K, calcium (Ca), and magnesium (Mg) removed in aboveground biomass are respectively 22.7, 1.6, 26.4, 16.4 and 4.8 kg Mg−1 of dry biomass.

2.5 Chapter Summary

Having analyzed 9 research papers, we can draw the following conclusions:

1) The following theoretical data analysis methods were used:
    a. the definition of NDVI;
    b. LR and correlation;
    c. Multiple Correspondence Analysis, MCA;
    d. Tukey–Kramer honestly significant difference test, Tukey–Kramer HSD test;
    e. Dispersion analysis ANOVA;

f. Shapiro-Wilk test;

g. Bartlett's test.

2) Instruments and apparatus used in experiments: SPAD-502, Field Scout CM 1000, Minolta SPAD-502 plus.

3) Software with which various types of analysis were carried out, – JMP (statistical software).

4) Test substances in soil, plant and tree tissues: Boron (B), Nitrogen (N), Phosphorus (P), Potassium (K), Calcium (Ca), Cooper (Cu), Iron (Fe), Zinc (Zn), Manganese (Mn), Magnesium (Mg).

Summing up the intermediate results after the analysis of scientific articles, it should be noted that:

- more productive processing of orchards and fields requires more data for analysis, which will establish more adequate ranges of sufficiency of necessary substances for the growth and development of trees and plants;
- it is also necessary to provide data from different territories of plantation cultivation (territories of rainfed lands, gardens with systemic irrigation, etc.);
- data on trimming of trees in a larger volume is required, preferably in combination with the previous two paragraphs.

Building a DST based on data analysis results:

- quantitative and percentage content in the soil and tissues of plants and trees of substances necessary for growth and evolution;
- territorial location of gardens and fields;
- about pruning carried out at a certain frequency, etc.

will allow farming and gardening of a new generation, which will be able to bring richer crops.

Chapter 3

Theoretical methods of Classical Statistics and Machine Learning at the core of intelligent farming

In this chapter we intend to describe and discuss, in some detail, some of the methods from MS and ML that are used in PF for modelling purposes.

In the first section we focus on the MA of variance – ANOVA – and MLR while in section 3.2 we look over ML techniques, discussing but a few of the most common.

3.1 Methods from Mathematical Statistics

3.1.1 Analysis of Variance - ANOVA

The Analysis of Variance – commonly known as ANOVA - is a statistical method for studying the relationship. It is used to study the influence of one or several qualitative variables on one dependent quantitative variable. The analysis of variance is based on the assumption that some variables can be considered as causes (independent variables) and others as consequences (dependent variables or responses). Independent variables in the analysis of variance are called

factors, since during the experiment the researcher can change their values of the dependent quantitative variable.

The main goal of the analysis of variance is to investigate the significance of the difference between the average values of the dependent quantitative variable for the groups of factors. This is achieved by decomposing the total variance of the dependent variable into components: the variance due to the division into groups (intergroup dispersion) and the variance due to other factors (intragroup dispersion). By analyzing these components of the variance, we can estimate the proportion of the influence of each factor on the dependent variable. A separate task of the analysis of variance is to identify due to which particular groups there is a difference in the average values of the dependent variable.

The essence of analysis of variance is to study the influence of one or more independent variables, usually referred to as factors, on the dependent variable. Dependent variables are represented by absolute scales (ratio scale). Independent variables are nominative (scale of names), that is, reflect group affiliation, and can have two or more values (type, gradation or level). Examples of the independent variable Xi with two values are gender (female: X1, male: X2) or the type of experimental group (control: X1, experimental: X2). The gradations corresponding to independent samples of objects are called intergroup, and the gradations corresponding to dependent samples are called intragroup.

Depending on the type and number of variables distinguish:

- one-factor and MA of variance (one or more independent variables);
- one-dimensional and multidimensional analysis of variance (one or more dependent variables);
- analysis of variance with repeated measurements (for dependent samples);
- analysis of variance with constant factors, random factors, and mixed models with factors of both types.

The procedure of analysis of variance consists in determining the ratio of systematic (intergroup) variance to random (intragroup) variance in the measured data. As an indicator of variability, the sum of the squared deviations of the parameter values from the mean is used: *SS* (*Sum of Squares*). It can be shown that the total sum of squares of SS_{total} is decomposed into the intergroup sum of squares of SS_{bg} and the intragroup sum of squares SS_{wg}: $SS_{total} = SS_{bg} + SS_{wg}$.

Degrees of freedom are laid out in a similar way:

$$df_{total} = df_{bg} + df_{wg}, \text{ where}$$

$$df_{total} = N - 1,$$

$$df_{bg} = J - 1,$$

$$df_{wg} = N - J,$$

and *N* is full sample size, and *J* – count of groups.

Then the variance of each part, referred to in the analysis of variance as "Mean Square", or *MS* is the ratio of the sum of squares to the number of their degrees of freedom:

$$MS_{total} = \frac{SS_{total}}{N-1},$$

$$MS_{bg} = \frac{SS_{bg}}{J-1},$$

$$MS_{wg} = \frac{SS_{wg}}{N-J}.$$

The ratio of intergroup and intragroup variances has an F-distribution (Fisher distribution) and is determined using (Fisher's F-test):

$$F_{df_{bg}, df_{wg}} = \frac{MS_{bg}}{MS_{wg}}.$$

The starting points of the analysis of variance are:

- normal distribution of values of the studied trait in the general population;
- equality of variances in the compared populations;
- random and independent sampling.

The null hypothesis in the analysis of variance is the statement about the equality of mean values:

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_j.$$

If the null hypothesis is rejected, an alternative hypothesis is accepted that not all means are equal, that is, there are at least two groups that differ in average values:

$$H_1: \mu_1 \neq \mu_2 \neq \cdots \neq \mu_j.$$

If there are three or more groups, post-hoc t-tests or the contrast method are used to determine the differences between the means.

The simplest case of analysis of variance is a one-dimensional one-factor analysis for two or more independent groups, when all groups are combined according to one characteristic. During the analysis, the null hypothesis of equality of means is tested. In the analysis of two groups, the analysis of variance is identical to the two-sample student t-criterion for independent samples, and the value of the F-statistics is equal to the square of the corresponding t-statistics.

The Levene's test is usually used to confirm the equality of variances clause. In the case of rejection of the hypothesis of equality of variances, the main analysis is not applicable. If the variances are equal, then the Fisher F-test is used to assess the ratio of intergroup and intragroup variability.

If the F-statistic exceeds a critical value, then the null hypothesis cannot be accepted (rejected) and a conclusion is drawn about the inequality of the means. When analyzing the middle of two groups, the results can be interpreted immediately after applying the Fisher criterion.

If there are three or more groups, pairwise comparison of means is required to identify statistically significant differences between them. A priori analysis involves the method of contrasts, in which the intergroup sum of squares is divided into the sum of the squares of the individual contrasts:

$$SS_{bg}: SS_{\varphi_1} + SS_{\varphi_2} + \cdots + SS_{\varphi_n},$$

where φ is the contrast between the means of the two groups, and then using the Fisher test checks the ratio of the average square for each contrast to the intragroup mean square:

$$F_{1,df_{wg}} = \frac{MS_{\varphi_i}}{MS_{wg}}.$$

Post hoc analysis includes post-hoc t-tests using the Bonferroni or Scheffe methods, as well as a comparison of the differences of the averages according to the Tukey method. A feature of post-hoc tests is the use of the intragroup mean square MSwg to evaluate any pair of means. Bonferroni and Scheffe tests are the most conservative because they use the smallest critical area for a given significance level α.

In addition to estimating averages, analysis of variance includes determining the coefficient of determination R^2, which shows how much of the total variability this factor explains:

$$R^2 = \frac{SS_{bg}}{SS_{total}}.$$

MA allows you to check the influence of several factors on the dependent variable. Unlike the one-factor model, where there is one intergroup sum of squares, the MA model includes the sum of squares for each factor separately and the sum of squares of all interactions between them. So,

in a two-factor model, the intergroup sum of squares is decomposed into the sum of squares of factor A, the sum of squares of factor B and the sum of squares of the interaction of factors A and B:

$$SS_{total} = SS_A + SS_B + SS_{AB} + SS_{wg}.$$

Accordingly, the three-factor model includes the sum of the squares of factor A, the sum of the squares of factor B, the sum of the squares of factor C and the sum of the squares of the interactions of factors A and B, B and C, A and C, as well as the interactions of all three factors A, B, C:

$$SS_{total} = SS_A + SS_B + SS_C + SS_{AB} + SS_{BC} + SS_{AC} + SS_{ABC} + SS_{wg}.$$

Degrees of freedom are laid out in a similar way:

$$df_{total} = df_A + df_B + df_{AB} + df_{wg}, \text{ where}$$

$$df_{total} = N - 1,$$

$$df_A = J - 1,$$

$$df_B = K - 1,$$

$$df_{AB} = (J - 1)(K - 1),$$

$$df_{wg} = N - JK,$$

and N is full sample size, J – the count of levels (groups) of the factor A, and K — the count of levels (groups) of the factor B. The analysis tests several null hypotheses:

- hypothesis of equality of averages under the influence of a factor A:

$$H_0: \mu_{1,*} = \mu_{2,*} = \cdots = \mu_{j,*};$$

- hypothesis of equality of averages under the influence of a factor B:

$$H_0: \mu_{*,1} = \mu_{*,2} = \cdots = \mu_{*,k};$$

- hypothesis of the absence of interaction of factors A и B:

$$H_0: (ab)_{j,k} = 0 \text{ for every } j \text{ and } k.$$

Each hypothesis is tested using the Fisher test.:

$$F_{df_A, df_{wg}} = \frac{MS_A}{MS_{wg}};$$

$$F_{df_B, df_{wg}} = \frac{MS_B}{MS_{wg}};$$

$$F_{df_{AB}, df_{wg}} = \frac{MS_{AB}}{MS_{wg}}.$$

When rejecting the null hypothesis of the influence of a single factor, it is accepted that the main effect of factor A (B, etc.) is present. When the null hypothesis on the interaction of factors is rejected, the assertion is accepted that the influence of factor A manifests itself differently at different levels of factor B. Usually, in this case, the results of the general analysis are recognized as invalid and the effect of factor A is checked separately at each level of factor B using one-way analysis of variance or t-test.

3.1.2 Multi Linear regression

The next method of MS that will be considered is LR. In MS, LR is a method of approximating the relationships between input and output variables based on a linear model. It is part of a broader statistical technique called regression analysis.

In statistical modeling and ML, the input variable is the value on which the change in the output variable depends, and the purpose of building the model is to approximate this dependence.

Input variables are usually associated with some features that describe the state of the studied objects or business processes. Since there can be several such features, the input variables can form a vector of input variables (more simply called an input vector).

In mathematical and statistical modeling, the output is called the model variable, which depends on the input variables and random factors affecting the simulated process or object.

The output variable represents the results of the model. It changes (varies) under the influence of changes in the input variable and random factors. The study of the variability of the output variable when the input is changed is the purpose of the simulation.

It is generally accepted to denote the input variable *X*, the output *Y*, and the simulated relationship between them by *f*. An additional term is added, which reflects part of the variability of the output variable due to random factors, which is denoted by ε.

In regression analysis, input (independent) variables are also called predictor variables or regressors, and dependent variables are called criterion variables.

If we consider the relationship between one input and one output variable, then there is a simple LR. For this, the regression equation $y = ax + b$ is determined and the corresponding straight line, known as the regression line, is constructed (Figure 3.1) [36].

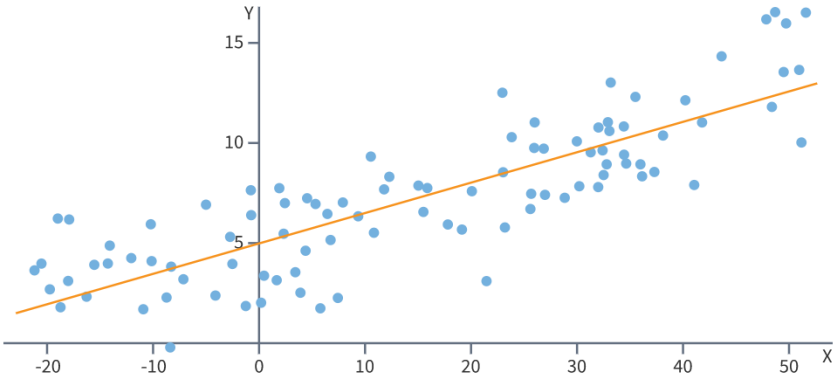

Figure 3.1: Regression line

The coefficients a and b, also called model parameters, are determined in such a way that the sum of the squared deviations of the points corresponding to real data observations from the regression line is minimal. Coefficients are usually estimated using the least squares method.

LR was the first type of regression analysis that was carefully studied and began to be widely used in practical applications. This is due to the fact that in linear models the estimation of parameters is simpler, and also because the statistical properties of the obtained estimates are easier to determine.

LR has many practical uses. Most applications fall into one of two broad categories:

- if the goal is forecasting, LR can be used to fit the model to the observed data set;
- if the goal is to explain the variability of the output variable, LR analysis can be used to quantify the strength of the relationship between the output and input variables.

A MLR model is a LR model where the number of independent variables is greater than one. The equation of MLR has the form:

$$Y = b_0 + b_1x_1 + b_2x_2 + \ldots + b_nx_n$$

As in simple LR, the parameters $b_0, b_1, b_2,\ldots,b_n$ of the model are calculated using the least squares method. The difference between simple and MLR is that instead of a regression line, it uses a hyperplane.

The advantage of MLR over simple is that the use of several input variables in the model allows you to increase the share of the explained variance of the output variable, and thus improve the model's fit to the data. Those. as each new variable is added to the model, the coefficient of determination increases.

However, in MLR, problems arise that are not characteristic of the simple model:

- multicollinearity may occur;
- it is necessary to choose the best model in which a minimal set of independent variables can explain the largest share of the variance of the dependent. For these purposes, the Akaike information criterion and its modifications, Bayes and Hannan-Quin information criteria are used.

Above we considered methods of MS, which are accurate and powerful tools that have proven their worth by many thousands of tests. But even such methods have disadvantages. However, improving the performance of these methods and increasing the accuracy of their results can be achieved by applying the ML approaches.

3.2 Machine Learning Techniques

ML is an extensive subsection of artificial intelligence, a mathematical discipline that uses sections of MS, numerical optimization methods, probability theory, discrete analysis, and extracts knowledge from data. When it comes to Analytics, it is necessary to include both DM / ML and traditional statistical methods in this concept. But do not forget that they differ both in the fields of application and in general philosophy.

ML is a subspecies of Artificial Intelligence (AI). The main task of AI is to understand and reproduce the human thought process and to reproduce it in a non-human (machine) environment. ML is aimed at the automatic detection of patterns in data using computational algorithms and their further structuring into new, but similar data. Hence, the main task is the study and creation of systems that are able to draw conclusions from data using training by examples. ML is not the same as “DM” or “predictive analytics” but it is also an integral part of both of them.

The basic concepts of ML began to be formulated in the 1950s, and in the late 80s and early 90s various successful startups appeared on this topic with a variety of application methods, such as

real-time tracking of fraud cases, character recognition and advisory services (first generation of ML systems). ML is also closely related to the “Pattern Recognition Theory” (TPO). True, while ML was created primarily by programmers and for programmers, SRW has its roots in engineering. But despite this, they are two sides of the same industry, whose task is DM. The growing interest in ML today signals the beginning of the next major wave of innovation. The details of ML tasks are shown in Figure 3.2 [37].

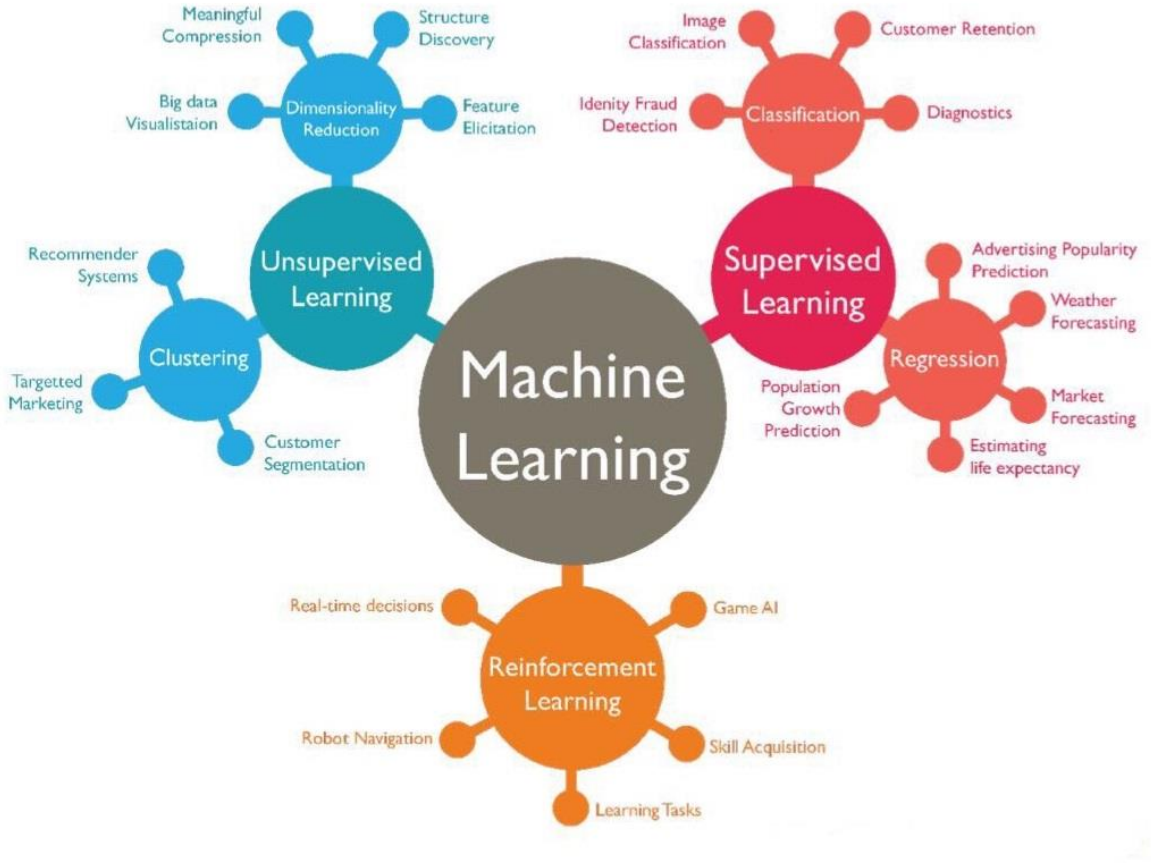

Figure 3.2: Types of ML

Areas of application for ML include:

- DM and predictive analytics;
- detecting cases of fraud, placing advertisements, evaluating borrowers, recommendations, creating medicines, trading stocks, improving consumer experience, etc.;
- word processing and analytics;
- Internet - search, spam filtering, tonality analysis and more;
- visualization of results;

- other: speech recognition, human genome, bioinformatics, optical character recognition (OCR), face recognition, self-driving cars, scene analysis, etc.

ML, as a rule, tries to reduce the number of initial assumptions and is freer to use various methods and means to solve the problem, often resorting to heuristics. The preferred teaching method in ML is inductive. In its extreme manifestations, there is much data available with inductive training, but there is practically no past experience, or it is not required for successful training. Another range of ML tool is analytical (or deductive) training, where there is either little data or it is preferable to work with their small sets. At the same time, there is also a good knowledge of the problem and related information. In the real world, ML usually alternates using both methods. TS, on the other hand, are extremely conservative in their approaches to the problem and often make too many assumptions, especially when it comes to data distribution. Table 3.1 illustrates the difference in approaches and ideologies of these two areas.

Learning can be achieved by writing a program that takes into account all possible combinations of data. This process will be extremely resource-intensive and impossible to perform in relation to real work situations. Such a program will also never be as good as a carefully written and verified learning algorithm. Similar algorithms automatically learn from examples (like people), and then summarize the data based on what they have learned (inductive learning). The ability to generalize and categorize is a key indicator of assessing the performance of a learning platform. The most popular training algorithms can be divided into controlled and uncontrolled, and then each category can be divided by capabilities and functions (also called DM functions).

Supervised learning methods include the following:

- classification - the definition to which category a particular concept belongs;
- regression - forecasting a specific indicator over a selected period of time;
- forecasting - assessment of macro variables (total);
- determination of the importance of the indicator – recognition of the most significant indicators in the calculation of other classification or regression.

Unsupervised leaning methods include:

- clustering - search for existing data groups;
- associative models - analysis of associative and interrelated observations, for example, the “consumer basket” (combinations of products that often get together).

In this paper, the following ML approaches will be described in detail:

- Clustering: *k*-means method;
- Hierarchical Clustering;
- Decision Trees;
- Artificial neural networks.

3.2.1 Clustering: k-means method

The *k*-means method is a method of cluster analysis, the purpose of which is to divide m observations (from space R^n) into *k* clusters, with each observation referring to the cluster to the center (centroid) of which it is closest [10].

Euclidean distance is used as a measure of proximity:

$$\rho(x, y) = \|x - y\| = \sqrt{\sum_{\rho=1}^{n}(x_\rho - y_\rho)^2}, \text{ where}$$

So, let's look at a number of observations $\left(x^{(1)}, x^{(2)}, \cdots, x^{(m)}\right), x^{(j)} \in R^n$.

The k-means method divides m observations into k groups (or clusters) ($k \leq m$) $S = \{S_1, S_2, \cdots, S_k\}$ to minimize the total square deviation of cluster points from the centroids of these clusters:$\min\left[\sum_{i=1}^{k}\sum_{x^{(j)}\in S_i}\left\|x^{(j)} - \mu_i\right\|^2\right]$, where $x^j \in R^n, \mu_i \in R^n$

μ_i – centroid for the cluster S_i.

The algorithm of the method looks as follows: if the measure of proximity to the centroid is determined, then dividing objects into clusters is reduced to determining the centroids of these clusters. The number of clusters k is set by the researcher in advance [11].

Consider an initial set of k means (centroids) $\mu_1, \cdots, \mu_k$ in clusters $S_1, S_2, \cdots, S_k$. At the first stage, the centroids of the clusters are selected randomly or according to a certain rule (for example, choose the centroids that maximize the initial distances between the clusters).

We refer observations to those clusters whose mean (centroid) is closest to them. Each observation belongs to only one cluster, even if it can be assigned to two or more clusters.

Then the centroid of each *i*-th cluster is recalculated according to the following rule:

$$\mu_i = \frac{1}{S_j} \sum_{x^{(j)} \in S_j} x^{(j)}$$

Thus, the k-means algorithm consists in recalculating the centroid at each step for each cluster obtained in the previous step.

The algorithm stops when the values μ_i do not change: $\mu_i^{step\ \tau} = \mu_i^{step\ \tau+1}$

Important: An incorrect choice of the initial number of clusters k may lead to incorrect results. This is why it is important when using the k-means method to first check the appropriate number of clusters for a given dataset [12].

So, let's once again highlight some of the features of the k-means method:

1) Euclidean distance is used as a metric;
2) the number of clusters is not known in advance and is chosen by the researcher in advance;
3) the quality of clustering depends on the initial partition.

3.2.2 Hierarchical clustering

So, let's move on to the HC method. HC (also graph clustering algorithms and hierarchical cluster analysis) - a set of data ordering algorithms aimed at creating a hierarchy (tree) of nested clusters. HC algorithms are often referred to as taxonomy algorithms [13]. Two classes of HC methods are distinguished:

- agglomerative methods: new clusters are created by combining smaller clusters and, thus, a tree is created from leaves to the trunk;
- divisive or divisional methods: new clusters are created by dividing larger clusters into smaller ones and, thus, the tree is created from the trunk to the leaves.

HE algorithms suggest that the analyzed set of objects is characterized by a certain degree of connectivity. According to the number of signs, monothetic and polythetic classification methods are sometimes distinguished. Like most visual methods for representing dependencies, graphs quickly lose visibility as the number of clusters increases.

For a visual presentation of the results of clustering, a dendrogram is used - a tree constructed from a matrix of measures of proximity between clusters. In the nodes of the tree are subsets of objects from the training set. Moreover, on each tier of the tree, the set of objects from all nodes makes up

the initial set of objects. The union of nodes between tiers corresponds to the merger of two clusters. In this case, the edge length corresponds to the distance between the clusters. Figure 3.3 shows the Fisher iris clustering dendrogram — a data set for the classification problem, based on which Ronald Fisher demonstrated the work of the discriminant analysis method developed by him in 1936 [29]. Sometimes it is also called Anderson's irises, as the data was collected by the American botanist Edgar Anderson. This data set has already become classical, and is often used in the literature to illustrate the operation of various statistical algorithms.

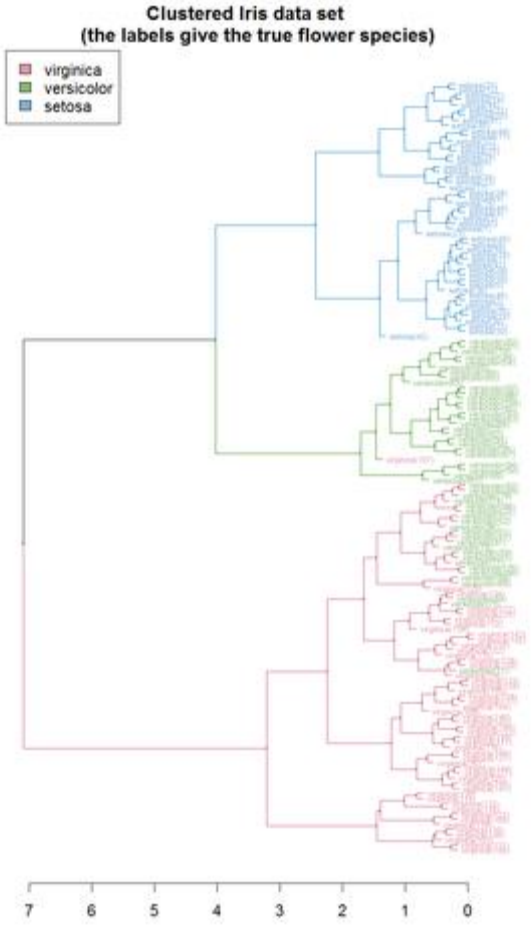

Figure 3.3: Fisher's iris clustering dendrogram

A tree is built from leaves to the root. At the initial time, each object is contained in its own cluster. Next, an iterative process of merging the two closest clusters occurs until all clusters are combined into one or the required number of clusters is found [14]. At each step, you must be able to calculate the distance between the clusters and recalculate the distance between the new clusters. The distance between singleton clusters is determined through the distance between objects: $R(\{x\}, \{y\}) = \rho(x, y)$. To calculate the distance $R(U, V)$ between clusters U and V in practice, various functions are used depending on the specifics of the task. There are several cluster distance functions:

- The single linkage method

$$R_{min}(U,V) = \min_{u \in U, v \in V} \rho(u,v)$$

- The complete linkage method

$$R_{max}(U,V) = \max_{u \in U, v \in V} \rho(u,v)$$

- UPGMA (Unweighted Pair Group Method with Arithmetic mean)

$$R_{avg}(U,V) = \frac{1}{|U| \cdot |V|} \sum_{u \in U} \sum_{v \in V} \rho(u,v)$$

- UPGMC (Unweighted Pair Group Method with Centroid average)

$$R_c(U,V) = \rho^2 \left(\sum_{u \in U} \frac{u}{|U|}, \sum_{v \in V} \frac{v}{|V|} \right)$$

- Ward's method

$$R_{ward}(U,V) = \frac{|U| \cdot |V|}{|U| + |V|} \rho^2 \left(\sum_{u \in U} \frac{u}{|U|}, \sum_{v \in V} \frac{v}{|V|} \right)$$

At each step, you must be able to quickly calculate the distance from the formed cluster $W = U \cup V$ to any other cluster S using the known distances from the previous steps. This is easily accomplished using the formula proposed by Lance and Williams in 1967:

$$R(W,S) = \alpha_U \cdot R(U,S) + \alpha_V \cdot R(V,S) + \beta \cdot R(U,V) + \gamma \cdot |R(U,S) - R(V,S)|,$$

where $\alpha_U, \alpha_V, \beta, \gamma$ – numeric parameters.

Each of the above distance functions satisfies the Lance-Williams formula with the following coefficients:

- The single linkage method

$$\alpha_U = \frac{1}{2}, \alpha_V = \frac{1}{2}, \beta = 0, \gamma = -\frac{1}{2}$$

- The complete linkage method

$$\alpha_U = \frac{1}{2}, \alpha_V = \frac{1}{2}, \beta = 0, \gamma = \frac{1}{2}$$

- UPGMA (Unweighted Pair Group Method with Arithmetic mean)

$$\alpha_U = \frac{|U|}{|W|}, \alpha_V = \frac{|V|}{|W|}, \beta = 0, \gamma = 0$$

- UPGMC (Unweighted Pair Group Method with Centroid average)

$$\alpha_U = \frac{|U|}{|W|}, \alpha_V = \frac{|V|}{|W|}, \beta = -\alpha_U \cdot \alpha_V, \gamma = 0$$

- Ward's method

$$\alpha_U = \frac{|S| + |U|}{|S| + |W|}, \alpha_V = \frac{|S| + |V|}{|S| + |W|}, \beta = \frac{-|S|}{|S| + |W|}, \gamma = 0$$

We introduce the notation R_t – the distance between the clusters selected in step t for combining.

The dendrogram allows you to represent the relationship between many objects with any number of specified characteristics on a two-dimensional graph, where all objects are plotted on one axis, and the distance R_t on the other. If no restrictions are imposed on this distance, then the dendrogram will have a large number of self-intersections and the image will cease to be visual. So that any cluster can be represented as a continuous segment on the axis of objects and the edges do not intersect, it is necessary to impose a monotonicity constraint on R_t.

By definition, the distance function R is monotonic if, at each subsequent step, the distance between the clusters does not decrease: $R_2 \leq R_3 \leq \cdots \leq R_m$.

The distance is monotonic if Milligan's theorem is valid for the coefficients in the Lance-Williams formula. Theorem: if the following three conditions are satisfied, then clustering is monotonic:

1) $\alpha_U \geq 0, \alpha_V \geq 0;$

2) $\alpha_U + \alpha_V + \beta \geq 1;$

3) $min\{\alpha_U, \alpha_V\} + \gamma \geq 0.$

Of the above distances, the theorem satisfies everything except the centroid.

To determine the number of clusters, the maximum length interval is found $|R_{t+1} - R_t|$. The resulting clusters are the clusters obtained in step t. The number of clusters is equal to $m - t + 1$. However, when the number of clusters is not known in advance and there are not very many objects in the sample, it is useful to study the entire dendrogram.

3.2.3 Decision Trees

Let's move on to the next method. So, DT are a family of models that allow you to restore nonlinear dependencies of arbitrary complexity. DT describe well the decision-making process in many situations.

The first work on the use of DT for data analysis appeared in the 60s, and since then a lot of attention has been given to them for several decades. Despite its interpretability and high expressive ability, trees are extremely difficult to optimize because of their discrete structure - the tree cannot be differentiated by parameters and at least a local optimum can be found using gradient descent. Moreover, even the number of parameters they have is not constant and can vary depending on the depth, the choice of crushing criteria and other details. Because of this, all the methods for constructing DT are greedy and heuristic.

To date, DT are not very often used as separate classification or regression methods. At the same time, as it turned out, they combine very well into compositions - decisive forests, which are one of the most powerful and versatile models.

Consider a binary tree in which:

- each inner vertex v is assigned a function (or a predicate) $\beta_v: \mathbb{X} \to \{0,1\}$;
- each leaf vertex v is assigned a prediction $c_v \in Y$ (in the case of classification, a probability vector can also be assigned to a leaf).

Consider now the algorithm $a(x)$, which starts from the root vertex v_0 and calculates the value of the function β_{v_0}. If it is equal to zero, then the algorithm goes to the left child vertex, otherwise to the right one, calculates the predicate value at the new vertex and makes a transition either to the left or to the right. The process continues until a leaf top is reached; the algorithm returns the class that is assigned to this vertex. Such an algorithm is called a binary decision tree.

In practice, in most cases, one-dimensional predicates β_v used, which compare the value of one of the attributes with a threshold:

$$\beta_v(x; j, t) = [x_j < t].$$

Multidimensional predicates exist, for example:

- linear $\beta_v(x) = [\langle w, x \rangle < t]$;

- metrical $\beta_v(x) = [\rho(x, x_v) < t]$, where the point x_v is one of the sample objects by any point of the attribute space.

Multidimensional predicates make it possible to construct even more complex dividing surfaces, but they are rarely used in practice, for example, because they reinforce the already outstanding ability of trees to retrain. Further we will talk only about one-dimensional predicates.

It is easy to make sure that for any sample, you can build a decision tree that does not allow any errors on it - even with simple one-dimensional predicates, you can create a tree with exactly one sample object in each sheet. Most likely, this tree will be retrained and will not be able to show good quality on new data. One could pose the problem of finding a tree that is minimal (in terms of the number of leaves) among all trees that do not allow for errors in learning - in this case, one could hope for the tree to have a generalizing ability. Unfortunately, this task is NP-complete, and therefore we have to confine ourselves to greedy tree-building algorithms.

DT can handle missing values - situations in which for some objects the values of one or more attributes are unknown. To do this, it is necessary to modify the sampling procedure at the vertex, which can be done in several ways.

After the tree is built, you can start pruning - removing some peaks in order to reduce complexity and increase the generalizing ability. There are several haircut approaches that we will mention a bit below.

Thus, the specific method for constructing the decision tree is determined:

- view predicates at the tops;
- quality features $Q(X, j, t)$;
- stop criterion;
- the method of processing missing values;
- the prunning method.

There may also be various extensions associated with taking into account the weights of objects, working with categorical signs, etc. Below we discuss the options for each of these items.

When constructing a tree, it is necessary to set the quality functional, on the basis of which the sampling is carried out at each step. We denote by R_m the set of objects that have reached the vertex to be split at this step, and by R_ℓ and R_r – we denote the objects that fall into the left and right subtrees, respectively, for a given predicate. We will use functionals of the following form:

$$Q(R_m, j, s) = H\left(R_m - \frac{|R_\ell|}{|R_m|}H(R_\ell) - \frac{|R_r|}{|R_m|}H(R_r)\right).$$

Where $H(R)$ – is impurity criterion, that evaluates the quality of the distribution of the target variable among objects of the set R. The smaller the diversity of the target variable, the less should be the value of the information content criterion - and, accordingly, we will try to minimize its value. In this case, we will maximize the quality functional $Q(R_m, j, s)$.

As discussed above, in each leaf the tree will produce a constant - a real number, probability or class. Based on this, it can be proposed to evaluate the quality of the set of objects R by how well their target variables are predicted by a constant (with an optimal choice of this constant):

$$H(R) = \min_{c \in \mathbb{Y}} \frac{1}{|R|} \sum_{(x_i, y_i) \in R} L(y_i, c),$$

where $L(y,c)$ – is a loss function. Next, we will discuss which specific informational criteria are often used in regression and classification problems.

As usual, in the regression we choose the squared deviation as a function of losses. In this case, the information content criterion will look like

$$H(R) = \min_{c \in \mathbb{Y}} \frac{1}{|R|} \sum_{(x_i, y_i) \in R} (y_i - c)^2.$$

The minimum in this expression will be achieved at the average value of the target variable. The criterion can be rewritten as follows:

$$H(R) = \frac{1}{|R|} \sum_{(x_i, y_i) \in R} \left(y_i - \frac{1}{|R|} \sum_{(x_i, y_i) \in R} y_i\right)^2.$$

We have found that the information content of a vertex is measured by its dispersion - the lower the spread of the target variable, the better the vertex. Of course, you can use other error functions L - for example, when choosing the absolute deviation, we get as the criterion the average absolute deviation from the median.

Consider the classification. Denote by p_k the fraction of class objects k ($k \in \{1, \ldots, K\}$), that reach the vertex R:

$$p_k = \frac{1}{|R|} \sum_{(x_i, y_i) \in R} [y_i = k].$$

Consider the error indicator as a function of losses:

$$H(R) = \min_{c \in \mathbb{Y}} \frac{1}{|R|} \sum_{(x_i, y_i) \in R} [y_i \neq c].$$

It is easy to see that the most popular class k_* will be the optimal prediction here, which means that the criterion will be equal to the following error rate:

$$H(R) = \frac{1}{|R|} \sum_{(x_i, y_i) \in R} [y_i \neq k_*] = 1 - p_{k_*}.$$

This criterion is rather crude, since it takes into account the frequency p_{k_*} of only one class.

Consider a situation in which we issue at the vertex not one class, but a distribution on all classes $c = (c_1, \cdots, c_K), \sum_k^K c_k = 1$. The quality of such a distribution can be measured, for example, using the Brier score:

$$H(R) = \min_{\sum_k c_k = 1} \frac{1}{|R|} \sum_{(x_i, y_i) \in R} \sum_{k=1}^{K} (c_k - [y_i = k])^2.$$

It can be shown that the optimal probability vector consists of fractions of classes p_k: $c_* = (p_1, \cdots, p_K)$

If we substitute these probabilities into the initial criterion of informativeness and carry out a series of transformations, then we obtain the Gini criterion:

$$H(R) = \sum_{k=1}^{K} p_k (1 - p_k).$$

Next, we consider the entropy criterion. We are already familiar with the more popular way of assessing the quality of probabilities - logarithmic losses, or the logarithm of likelihood:

$$H(R) = \min_{\sum_k c_k = 1} \left(-\frac{1}{|R|} \sum_{(x_i, y_i) \in R} \sum_{k=1}^{K} [y_i = k] \log c_k \right).$$

To derive the optimal values of c_k, recall that all values of c_k must be summed to unity. As is known from optimization methods, to account for this restriction, it is necessary to search for the minimum of the Lagrangian:

$$L(c,\lambda) = -\frac{1}{|R|}\sum_{(x_i,y_i)\in R}\sum_{k=1}^{K}[y_i = k]\log c_k + \lambda\sum_{k=1}^{K} c_k \to \min_{c_k}$$

Differentiating, we obtain:

$$\frac{\partial}{\partial c_k}L(c,\lambda) = -\frac{1}{|R|}\sum_{(x_i,y_i)\in R}[y_i = k]\frac{1}{c_k} + \lambda = -\frac{p_k}{c_k} + \lambda = 0,$$

whence we express $c_k = p_k/\lambda$. Summing up these equalities with respect to k, we obtain

$$1 = \sum_{k=1}^{K} c_k = \frac{1}{\lambda}\sum_{k=1}^{K} p_k = \frac{1}{\lambda},$$

whence $\lambda = 1$. Therefore, the minimum is reached at $c_k = p_k$, as in the previous case. Substituting these expressions into the criterion, we obtain that it will be the entropy of the distribution of classes:

$$H(R) = -\sum_{k=1}^{K} p_k \log p_k.$$

It is known from probability theory that entropy is bounded below by zero, and the minimum is achieved at degenerate distributions $(p_i = 1,\ p_j = 0\ for\ i \neq j\ )$. The entropy takes the maximum value for uniform distribution. This shows that the entropy criterion prefers more "degenerate" class distributions at the vertex.

Next, consider the breakdown criteria. You can come up with a lot of stopping criteria. We list some restrictions and criteria:

- the limitation of the maximum depth of the tree;
- the limitation of the minimum number of objects in the sheet;
- the limiting the maximum number of leaves in a tree;
- the stop if all objects in the sheet belong to the same class;
- the requirement that the quality functional during crushing improved by at least *s* percent.

Using the right choice of such criteria and their parameters can significantly affect the quality of the tree. However, this selection is time consuming and requires cross-validation.

Tree shearing is an alternative to the stopping criteria described above. When using a haircut, a retrained tree is first built (for example, until there is one object in each sheet), and then its structure is optimized in order to improve the generalizing ability. There are a number of studies showing that cutting allows you to achieve better quality compared to the early stop of building a tree based on various criteria.

However, at the moment, cutting methods are rarely used and are not implemented in most libraries for data analysis. The reason is that the trees themselves are weak algorithms and are not of great interest, and when used in compositions, they either have to be retrained (in random forests), or have a very small depth (in boosting), which is why there is no need for a haircut.

One of the cutting methods is cost-complexity pruning. We denote the tree obtained as a result of the greedy algorithm (is any algorithm that follows the problem-solving heuristic of making the locally optimal choice at each stage. In many problems, a greedy strategy does not usually produce an optimal solution, but nonetheless, a greedy heuristic may yield locally optimal solutions that approximate a globally optimal solution in a reasonable amount of time [30])by T_0. Since each of the leaves contains objects of only one class, the value of the functional $R(T)$ will be minimal on the tree T_0 itself (among all subtrees). However, this functional characterizes only the quality of the tree in the training set, and excessive fitting to it can lead to retraining. To overcome this problem, we introduce a new functional $R_\alpha(T)$, which is the sum of the original functional $R(T)$ and a fine for the size of the tree:

$$R_\alpha(T) = R(T) + \alpha|T|,$$

where $|T|$ is the number of leaves in subtree T, and $\alpha > 0$ is a parameter. This is one of the examples of regularized quality criteria that seek a balance between the quality of the classification of the training sample and the complexity of the constructed model.

It can be shown that there is a sequence of nested trees with the same roots:

$$T_K \subset T_{K-1} \subset \ldots \subset T_0,$$

(here T_K is a trivial tree consisting of the root of the tree T_0), in which each tree T_i minimizes the criterion $R_\alpha(T) = R(T) + \alpha|T|$ for α from the interval $\alpha \in [\alpha_i, \alpha_{i+1})$, and

$$0 = \alpha_0 < \alpha_1 < \cdot \cdot \cdot < \alpha_K < \infty.$$

This sequence can be quite effectively found by traversing the tree. Next, the optimal tree is selected from it according to the deferred selection or using cross-validation.

One of the main advantages of DT is the ability to work with missing values. Let's consider some options.

Suppose we need to calculate the quality functional for the predicate $\beta(x) = [x_j < t]$, but in the sample R for some objects the value of the sign j is not known – we denote them by V_j. In this case, when calculating the functional, you can simply ignore these objects, making an adjustment for the loss of information from this:

$$Q(R,j,s) \approx \frac{|R \setminus V_j|}{|R|} Q(R \setminus V_j, j, s).$$

Then, if this predicate turns out to be the best, put the objects from V_j both in the left and right subtrees. You can also assign them with the weight $|R_\ell|/|R|$ in the left subtree and $|R_r|/|R|$ in the right. In the future, weights can be taken into account by adding them as coefficients in front of the indicators $[y_i = k]$ in all formulas. At the stage of applying the tree, you must perform a similar trick. At the stage of applying the tree, you must perform a similar trick. If an object hits a vertex whose predicate cannot be calculated due to a skip, then forecasts for it are calculated in both subtrees, and then averaged with weights proportional to the number of training objects in these subtrees. In other words, if the probability forecast for the class k in the subtree R_m is denoted by $a_{mk}(x)$, then we obtain the formula:

$$a_{mk}(x) = \begin{cases} a_{\ell k}(x), & \beta_m(x) = 0; \\ a_{rk}(x), & \beta_m(x) = 1; \\ \dfrac{|R_\ell|}{|R_m|} a_{\ell k}(x) + \dfrac{|R_r|}{|R_m|} a_{rk}(x), & \beta_m(x) \text{cannot be calculated}. \end{cases}$$

Another approach is to construct surrogate predicates at each vertex. This is the name of a predicate that uses a different attribute, but at the same time gives a partition as close as possible to the given one.

Note that often much more simple methods are also shown by much simpler methods of processing gaps - for example, you can replace all gaps with zero. For trees, it would also be wise to replace the gaps in the trait with numbers that exceed any value of this trait. In this case, it will be possible to select such a partition according to this criterion in the tree that all objects with known values will go to the left subtree, and all objects with gaps will go to the right.

The most obvious way to handle categorical features is to split the vertex into as many subtrees as there are possible values for the feature (multi-way splits). This approach may show good results, but there is a risk of obtaining a tree with an extremely large number of leaves.

Let's take a closer look at another approach. Let the categorial criterion x_j have a set of values $Q = \{u_1, \ldots, u_q\}$, $|Q| = q$. We divide the set of values into two disjoint subsets: $Q = Q_1 \sqcup Q_2$, and define a predicate as an indicator of getting into the first subset: $\beta(x) = [x_j \in Q_1]$. Thus, the object will fall into the left subtree if the attribute x_j falls into the set Q_1, and in the first subtree otherwise. The main problem is that to construct the optimal predicate, it is necessary to sort out 2^{q-1}–1 partition option, which may not be entirely possible.

It turns out that you can do without exhaustive search in cases with binary classification and regression. Denote by $R_m(u)$ the set of objects that hit the vertex m and whose *j-th* attribute has the value u; by $N_m(u)$ we denote the number of such objects.

In the case of binary classification, we will arrange all the values of the categorical attribute based on how much of the class of objects with this value +1:

$$\frac{1}{N_m(u_{(1)})} \sum_{x_i \in R_m(u_{(1)})} [y_i = +1] \leq \cdots \leq \frac{1}{N_m(u_{(q)})} \sum_{x_i \in R_m(u_{(q)})} [y_i = +1],$$

after which we replace the category $u_{(i)}$ with the number i, and we look for a partition for a real attribute. It can be shown that if we look for the optimal partition by the Gini criterion or the entropy criterion, then we will get the same partition as when sorting through all possible 2^{q-1} –1 variant.

For the regression problem with the MSE functional, this will also be true if we order the attribute values according to the average response of objects with this value:

$$\frac{1}{N_m(u_{(1)})} \sum_{x_i \in R_m(u_{(1)})} y_i \leq \cdots \leq \frac{1}{N_m(u_{(q)})} \sum_{x_i \in R_m(u_{(q)})} y_i.$$

There are several popular methods for building trees:

- ID3: uses the entropy criterion. Builds a tree until each sheet contains objects of the same class, or until a vertex split gives a decrease in the entropy criterion;
- C4.5: uses the Gain Ratio criterion (normalized entropy criterion). Stop criterion - restriction on the number of objects in the sheet. Haircuts are made using the Error-Based Pruning method, which uses generalizing ability scores to make a decision about removing a vertex. Missing values are processed using a method that ignores objects with missing values when calculating branching criteria, and then transfers such objects to both subtrees with specific weights;

- CART: uses the Gini criterion. Shearing is done with cost-complexity pruning. The surrogate predicate method is used to handle gaps.

As follows from the definition, the decision tree $a(x)$ splits the entire attribute space into a certain number of disjoint subsets $\{J_1, \ldots, J_n\}$, and in each subset J_j gives a constant forecast w_j. Therefore, the corresponding algorithm can be written analytically:

$$a(x) = \sum_{j=1}^{n} w_i \left[x \in J_j\right].$$

It turns out that the decision tree with the help of a greedy algorithm selects the transformation of features for a given task, and then simply builds a linear model over these features.

3.2.4 Artificial neural networks

Next, we move on to an approach called “ANN” - a mathematical model, as well as its software or hardware embodiment, built on the principle of organization and functioning of biological NN - networks of nerve cells of a living organism.

ANN consists of artificial neurons, each of which is a simplified model of a biological neuron. All that an artificial neuron does is receive signals from many inputs, process them in a single way, and transmit the result to many other artificial neurons, i.e. does the same thing as a biological neuron. Biological neurons are interconnected by axons, the joints are called synapses. In synapses, an amplification or attenuation of an electrochemical signal occurs. Connections between artificial neurons are called synaptic, or simply synapses. The synapse has one parameter - the weight coefficient, depending on its value, one or another change in information occurs when it is transmitted from one neuron to another. Due to this, the input information is processed and converted into a result, and the training of a neural network is based on the experimental selection of such a weight coefficient for each synapse, which leads to the desired result.

The structure of the simplest neural network is shown in Figure 3.4 [38]. Green indicates the neurons of the input layer, blue - the neurons of the hidden layer, yellow - the neuron\s of the output layer.

Neurons of the input layer receive data from the outside (for example, from sensors of the face recognition system) and after processing them transmit signals through synapses to the neurons of the next layer. The neurons of the second layer (it is called hidden because it is not directly

connected to either the input or the output of the ANN) process the received signals and transmit them to the neurons of the output layer. Since we are talking about simulating neurons, each input level processor is associated with several processors of the hidden level, each of which, in turn, is associated with several processors of the output level.

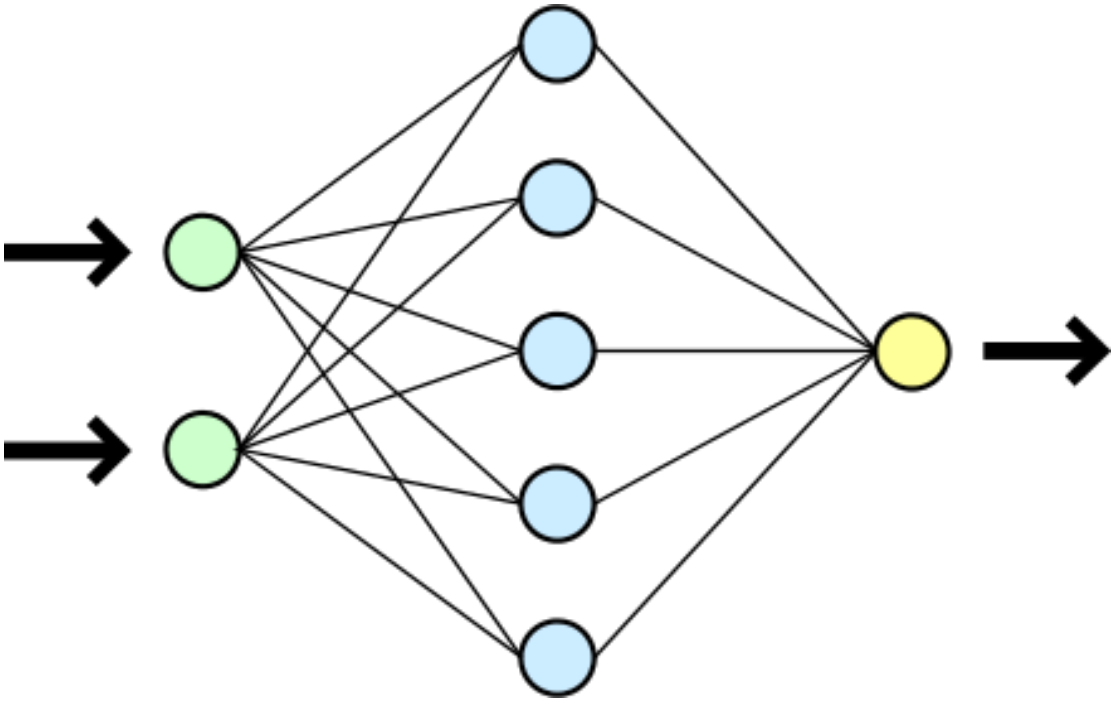

Figure 3.4: Scheme of the simplest neural network

This simplest ANN is capable of learning and can find simple relationships in data. An ANN, capable of finding not only simple relationships, but also the relationships between relationships, has a much more complex structure. It may contain several hidden layers of neurons, interspersed with layers that perform complex logical transformations. Each subsequent layer of the network looks for relationships in the previous one. Such ANNs are capable of deep (deep) learning. Thanks to the transition to the use of NN with deep training, Google was able to dramatically improve the quality of the work of its popular product “Translator”.

As of the beginning of 2019, the Wikipedia Internet resource totaled 26 types of NN. Of these, 12 were named after the names of their inventors, the rest had such names as chaotic, Siamese, oscillatory, adaptive resonance, etc. In order to somehow systematize existing and future NN, attempts are made to classify them, presented in Table 3.1.

Table 3.1: Classification of ANN

Classification sign	**The name of the artificial neural network or the approach of its training and construction**
The input type	analog (input real numbers)
	binary (binary numbers input)
	figurative (signs, hieroglyphs, symbols at the entrance)
The nature of training	the learning with a teacher (the output space of neural network solutions is known)
	the learning without a teacher (the output space of decisions is formed only on the basis of input influences; such networks are called self-organizing)
	the reinforcement training (the system of assigning fines and incentives received as a result of the interaction of the ANN with the environment is used)
The nature of the synapse settings	the networks with fixed connections (weights of the neural network are selected immediately, based on the conditions of the problem)
	the networks with dynamic connections (these networks configure synaptic connections during training)
Signal Transmission Time	the synchronous networks (transmission time for each synaptic connection is either zero or a fixed constant)
	the asynchronous networks (transmission time for each connection between the elements is different, but also constant)
The nature of ties	the direct distribution networks (all connections are directed strictly from input neurons to output ones)
	the recurrent networks (the signal from the output neurons or neurons of the hidden layer is partially transmitted back to the inputs of the neurons of the input layer)
	Hopfield recurrent network (filters the input data, returning to a stable state and, thus, allows solving the problems of data compression and construction of associative memory)
	the bidirectional networks (between layers there are connections both in the direction from the input layer to the output layer, and in the opposite)

In addition, radial basis networks (or RBF networks), self-organizing maps (in particular, Kohonen self-organizing map), and networks of other classes that are not yet fully formed, are used.

NN are used to solve complex problems that require analytical calculations similar to those made by the human brain. The most common tasks for which NN are used are:

- the pattern recognition - various objects can act as images: text symbols, images, sound samples, etc. Currently, this is the widest area of application of NN. In particular, this

ability of them is used in Google when you are looking for a photo, or in the smartphone's camera when it determines the position of your face and highlights it, and in many other applications;

- the classification - distribution of data by parameters. For example, a set of data about people is fed to the ANN input and you need to decide who you can give credit to and who not. This work can be done by a neural network, analyzing information such as age, solvency, credit history, etc.;
- the decision making and management – this task is close to the classification problem. Classifications are subject to situations whose characteristics are input to the neural network. The output of the network should result in a sign of the decision that it made. Moreover, various criteria of the state of the controlled system are used as input signals;
- the clustering - clustering refers to the partitioning of the set of input signals into classes, while neither the quantity nor the characteristics of the classes are known in advance. After training, such a network is able to determine which class the input signal belongs to. The network may also signal that the input signal does not belong to any of the selected classes - this is a sign of the appearance of new data that are not in the training set. Thus, such a network can identify new, previously unknown signal classes;
- the forecasting - the ability of a neural network to predict directly follows from its ability to generalize and highlight hidden relationships between input and output data. After training, the network is able to predict the future value of a certain sequence based on several previous values and (or) some currently existing factors;
- the approximation - a neural network is able to approximate any continuous function with a certain predetermined accuracy;
- the data compression and associative memory - the ability of NN to identify the relationships between different parameters makes it possible to present data more compactly if the data is closely related. The reverse process - restoring the original data set from a piece of information - is called (auto) associative memory. Associative memory also allows you to restore the original signal / image from noisy / damaged input data.

NN are also used for data analysis, solving optimization problems, finding patterns in large volumes of data, spatial orientation, etc.

The list of advantages and disadvantages of ANN is presented in Table 3.2.

Table 3.2: Advantages and disadvantages of ANN

Advantages		**Disadvantages**	
Title	**Description**	**Title**	**Description**
The problem solving in the face of uncertainty	Thanks to the ability to learn, the neural network allows you to solve problems with unknown patterns and relationships between input and output data, which allows you to work with incomplete data	The answer given by the ANN is always approximate	NN are not able to give accurate and unambiguous answers. But the tasks in which it is necessary to use ANN and at the same time get accurate answers are quite rare
The noise resistance in input	A neural network can independently detect parameters that are not informative for analysis and filter them out, and therefore there is no need for a preliminary analysis of the input data	The inability to make decisions in several stages	A neural network cannot solve tasks that require the sequential execution of several steps; she is able to solve the problem only "in one go". Therefore, a neural network cannot, for example, prove a mathematical theorem
The flexibility of the structure of NN	The components of neurocomputers - neurons and the connections between them - can be combined in various ways. Due to this, one neurocomputer can be used to solve various problems, often not related to each other. High performance. Input data is processed by many neurons simultaneously, due to which NN solve problems faster than most other algorithms	The complexity and duration of training	In order for the neural network to correctly solve the tasks, it is required to conduct its training on tens of millions of sets of input data. But various accelerated learning technologies have already been developed, modern video cards allow you to train NN hundreds of times faster, and recently, ready-made, pre-trained NN, in particular, recognizing images, have appeared. Based on such NN, you can create applications without having to do long training
The adapting to environmental change	Learning from data, NN are able to adapt to a changing environment (for example, to changes in the market situation, if the task of a neural network is to predict price fluctuations on the exchange). If it is necessary to solve a problem in a non-stationary environment, then NN can be created, retrained	The inability to solve computational problems	It is impossible to load, for example, a mathematical equation into the ANN and obtain its solutions for various parameters. But this is not the purpose of NN

Advantages		Disadvantages	
Title	**Description**	**Title**	**Description**
	in real time. The higher the adaptive ability of the system, the more stable will be its operation in an unsteady environment		
The fault tolerance of NN	The neural network reacts to an adverse change in conditions with only a slight decrease in productivity. This feature is explained by the distributed nature of information storage in the neural network, therefore only serious damage to the structure can significantly affect the performance of the neural network		

3.3 Mathematical Statistics vs. Machine Learning

In this section we describe what are the main differences between the two described approaches in what regards the analysis of data, the needed assumptions and the type of information it can provide (Table 3.3).

Table 3.3: Differences between ML and TS

ML	TS
Task: “study” of all data types	Task: data analysis and classification
No mandatory assumptions regarding task and data distribution	Big assumptions regarding the issue and data distribution
More freely in methods and approaches	Conservative in methods and approaches
The generalization is achieved empirically, by confirming the assumptions in practice in addition to the prepared tests	The generalization is pursued based on pre-prepared tests
Feel free to use a heuristic approach to find the best solution	Big assumptions in the data and the resolved issue, the search for the optimal solution is carried out within the given framework

ML	TS
A large number of functions are often useful. It is preferable to use algorithms capable of processing a large number of functions	Often requires independent functions. Use less of these features is preferred.
Reducing the analyzed data is not necessary. Promotes a culture of wealth: “the more data, the better”	Actively promotes analysis of smaller amounts of data (sampling, input reduction)
Solves more complex problems of learning, perception, reproduction of knowledge	Mostly focused on traditional analysis methods

3.4 Chapter Summary

This chapter explored the theoretical methods of Classical MS, which were presented by Analysis of variance - ANOVA and the MLR approach, as well as ML methods, presented by the k-Means Clustering method, HC, DT and ANN approach. The study showed that both the first group of methods and the second have both advantages and disadvantages. However, their combined use allows not only to neglect the negative aspects of some methods, but also synergistically multiply the positive result of such symbiosis.

Chapter 4

Practical implementation of theoretical methods

4.1 Preparing the dataset “Plant data” for further analysis

To get started building MLR models (classical MS) as well as HC (ML) and testing them directly, you need to make sure that the distribution of the variables matches the definition of normal. Several different approaches can be used for this.

The most visual and therefore the most popular graphical methods are the histogram of the frequency distribution of the variable values and Q-Q Plot (Quantile-Quantile Plot). Figure 4.1 shows an example of plotting a Q-Q Plot.

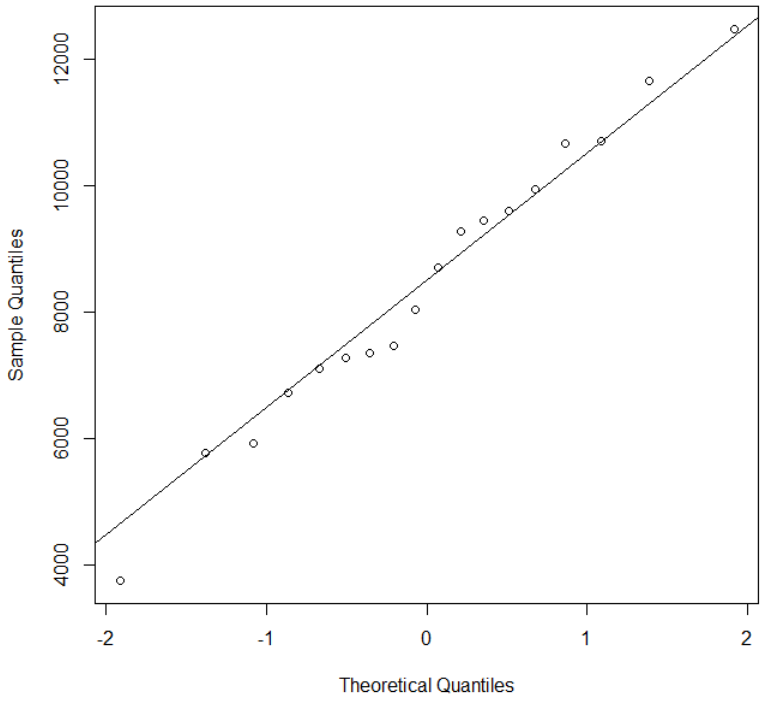

Figure 4.1: Variable “Height, cm” Q-Q Plot, dataset “Plant data”

In order to study the type of distribution of variables from the “Plant data” dataset, histograms of all available numerical values were taken. An example of one of the captured histograms is shown in Figure 4.2.

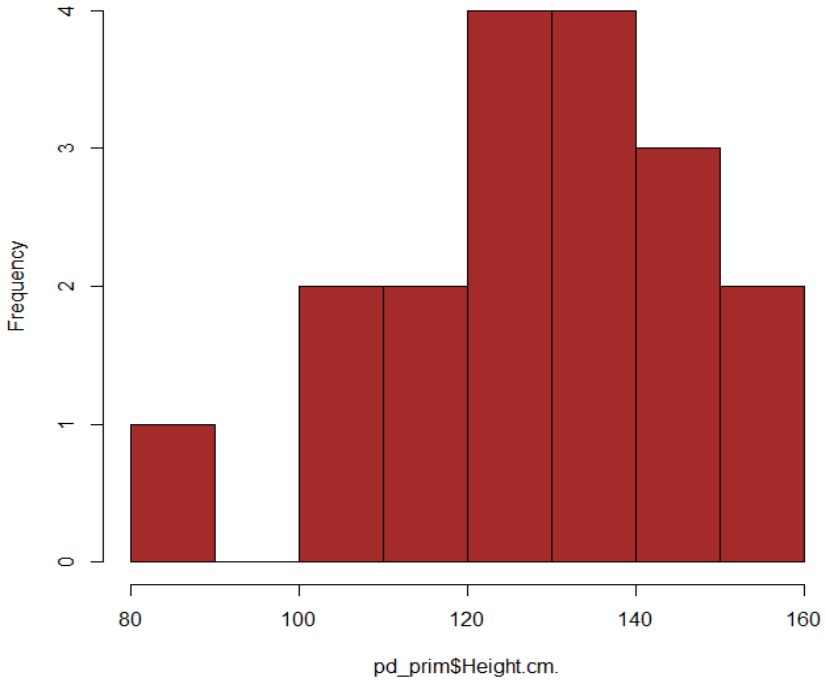

Figure 4.2: Variable “Height, cm” histogram, dataset “Plant data”

After a visual representation of the distributions of the variables, the Shapiro-Wilk and Anderson-Darling tests were taken to check the distributions for normality. The obtained test results are presented in Table 4.1.

Table 4.1: Test results for primary variables (unnormalized)

Primary data frame (pd_prim)				
Variable	**Value**		**P-value**	
	W (Shapiro–Wilk)	**A (Anderson–Darling)**	**W (Shapiro–Wilk)**	**A (Anderson–Darling)**
Height.cm.	0,97336	0,1973	0,8686	0,8686
Diameter.cm.	0,90754	0,62947	0,07786	0,08488
Prunning.firewood.mg.	0,96969	0,21512	0,7918	0,8199
N11	0,93248	0,42288	0,2147	0,2862
P11	0,95333	0,30244	0,4797	0,5389
K11	0,96261	0,27132	0,6527	0,6299
Ca11	0,97599	0,18874	0,8996	0,8872
Mg11	0,98809	0,1305	0,9964	0,9773
B12	0,91639	0,63933	0,1116	0,08003
Cu12	0,8248	1,3062	0,003478	0,001517
Fe12	0,9268	0,54929	0,1705	0,1347
Zn12	0,94695	0,31996	0,3792	0,5067
Mn12	0,92786	0,53458	0,1781	0,1474
N21	0,88823	0,97123	0,03599	0,01109
P21	0,96803	0,25129	0,7599	0,6997
K21	0,9591	0,23575	0,5843	0,7537
Ca21	0,9423	0,33365	0,3187	0,4732
Mg21	0,97606	0,20665	0,9006	0,8441
B22	0,85095	1,0791	0,008767	0,005844
Cu22	0,7941	1,6525	0,001255	0,0001949
Fe22	0,92149	0,62116	0,1374	0,08919

Primary data frame (pd_prim)				
Variable	**Value**		**P-value**	
	W (Shapiro–Wilk)	**A (Anderson–Darling)**	**W (Shapiro–Wilk)**	**A (Anderson–Darling)**
Zn22	0,97705	0,24469	0,9146	0,7229
Mn22	0,80272	1,3162	0,00166	0,00143
N31	0,94848	0,28499	0,4017	0,5859
P31	0,95594	0,43712	0,5255	0,2635
K31	0,90061	0,57494	0,05889	0,1175
Ca31	0,93995	0,47546	0,2893	0,2104
Mg31	0,93476	0,41128	0,2353	0,3059
B32	0,93333	0,49008	0,2221	0,1929
Cu32	0,89082	0,82112	0,03987	0,0271
Fe32	0,91209	0,5043	0,09364	0,1771
Zn32	0,92621	0,3708	0,1665	0,3849
Mn32	0,96494	0,28737	0,6989	0,5788
N41	0,92874	0,53873	0,1846	0,1437
P41	0,97424	0,20404	0,8725	0,8512
K41	0,85222	1,0666	0,00918	0,006294
Ca41	0,93599	0,39402	0,2471	0,3376
Mg41	0,96632	0,25585	0,7263	0,6835
B42	0,87287	0,98518	0,01985	0,01021
Cu42	0,81051	1,6221	0,002146	0,0002332
Fe42	0,93239	0,35464	0,2138	0,4213
Zn42	0,95273	0,27688	0,4695	0,6114
Mn42	0,8545	0,76962	0,009979	0,03683
SPAD.30.05.2014	0,94894	0,41403	0,4086	0,3011
SPAD.28.06.2014	0,80471	1,583	0,001772	0,0002941

Primary data frame (pd_prim)				
Variable	**Value**		**P-value**	
	W (Shapiro–Wilk)	**A (Anderson–Darling)**	**W (Shapiro–Wilk)**	**A (Anderson–Darling)**
SPAD.07.08.2014	0,95493	0,2686	0,5075	0,6391
SPAD.24.07.2015	0,95489	0,42657	0,5068	0,2801
SPAD.15.07.2016	0,96131	0,40239	0,6271	0,3219
SPAD.22.08.2017	0,95141	0,32945	0,4476	0,4842
NDVI.30.05.2014	0,87107	0,85364	0,01853	0,02233
NDVI.28.06.2014	0,79309	1,6869	0,001215	0,0001589
NDVI.07.08.2014	0,81479	1,6852	0,002476	0,0001606
NDVI.24.07.2015	0,93865	0,4876	0,2747	0,1957

According to the test results, 14 variables were identified ("Cu12", "N21", "B22", "Cu22", "Mn22", "Cu32", "K41", "B42", "Cu42", "Mn42", "SPAD, 28.06.2014", "NDVI, 30.05.2014", "NDVI, 28.06.2014", "NDVI, 07.08. 2014"), for which the hypothesis of the normal distribution of values was rejected. That is, out of 53 numerical variables of the "Plant data" dataset, only 39 had a distribution close to normal.

To normalize the distribution of variable values, the BoxCox method was applied. In reality, we often have to deal with statistical data that, for one reason or another, do not pass the normality test. In this situation, there are two ways out: either turn to nonparametric methods, or use special methods to convert the original "abnormal statistics" into "normal" ones. Among the many such transformation methods, one of the best (for an unknown distribution type) is the Box-Cox transformation.

For the original sequence $y = \{y_1, \cdots, y_n\}, y_i > 0, i = 1, \cdots, n$ the one-parameter Box-Cox transformation with the λ parameter is defined in the following way:

$$y_i^{\lambda} = \begin{cases} \frac{y_i^{\lambda}-1}{\lambda}, & \lambda \neq 0, \\ log(y_i), & \lambda = 0. \end{cases} \quad (1)$$

The λ parameter can be selected to maximize the log likelihood. Another way to find the optimal parameter value is based on finding the maximum value of the correlation

coefficient between the quantiles of the normal distribution function and the sorted transformed sequence.

Since the original method assumes working only with positive values, several modifications were proposed, taking into account zero and negative values.

The most obvious option is to shift all values by the constant α so that the condition $(y_i + \alpha) > 0, i = 1, \cdots, n$ is satisfied. After that, the transformation looks like this:

$$y_i^{\lambda} = \begin{cases} \frac{(y_i+\alpha)^{\lambda}-1}{\lambda}, & \lambda_1 \neq 0, \\ log(y_i + \alpha), & \lambda = 0. \end{cases} \quad (2)$$

An even more general formula:

$$\tau(y_i; \lambda, \alpha) = \begin{cases} \frac{(y_i+\alpha)^{\lambda}-1}{\lambda(GM(y))^{\lambda-1}}, & \lambda \neq 0, \\ GM(y)\ ln(y_i + \alpha), & \lambda = 0, \end{cases} \quad (3)$$

where $GM(y) = (y_1 \cdots y_n)^{1/n}$.

Since all variables are positive, formula 1 was used to determine the λ parameter. Table 4.2 shows the estimated Lambda and the sample Skewness (coefficient of skewness of the distribution of the variable).

Table 4.2: Lambda and Skewness metrics

Variable	Estimated Lambda	Sample Skewness
Height.cm.	2	-0,415
Diameter.cm.	2	-0,638
Prunning.firewood.mg.	0,9	-0,0651
N11	0,6	0,064
P11	0,4	0,19
K11	0,9	0,009
Ca11	0,6	0,19
Mg11	0,9	0,0506
B12	1,1	-0,465

Variable	Estimated Lambda	Sample Skewness
Cu12	-0,5	0,927
Fe12	-0,1	0,763
Zn12	0,2	0,54
Mn12	0,3	0,379
N21	2	-0,768
P21	2	-0,464
K21	0	0,315
Ca21	-1	0,607
Mg21	1,2	-0,0716
B22	2	0,959
Cu22	0,2	0,951
Fe22	-0,6	0,636
Zn22	0,5	0,361
Mn22	-1,6	1,46
N31	2	-0,695
P31	0,2	0,22
K31	-1,9	0,906
Ca31	-1,1	0,658
Mg31	2	-0,441
B32	0,9	-0,365
Cu32	2	-0,888
Fe32	-1,5	0,701
Zn32	-1,1	0,854
Mn32	0,6	0,116
N41	-2	0,629
P41	0,5	0,181

Variable	Estimated Lambda	Sample Skewness
K41	2	0,937
Ca41	-0,4	0,267
Mg41	1,1	-0,0844
B42	1,3	0,818
Cu42	0,6	0,132
Fe42	-1,6	0,817
Zn42	-0,4	0,533
Mn42	-0,5	1,31
SPAD.30.05.2014	1,7	-0,87
SPAD.28.06.2014	2	-1,15
SPAD.07.08.2014	-0,9	0,0768
SPAD.24.07.2015	2	-0,596
SPAD.15.07.2016	2	-0,316
SPAD.22.08.2017	-0,2	0,0692
NDVI.30.05.2014	-2	0,281
NDVI.28.06.2014	2	-1,05
NDVI.07.08.2014	2	-0,529
NDVI.24.07.2015	1,5	-0,03

After applying formula 1 to all the variables, new values of the variables were obtained and the distribution slightly changed its form. An example of a change is shown in Figure 4.3.

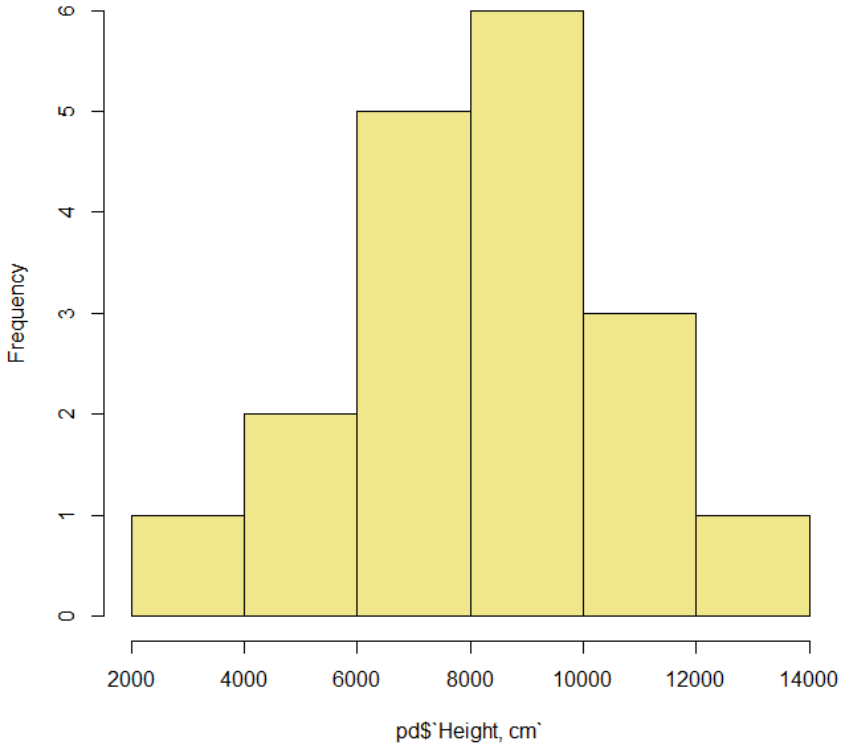

Figure 4.3: Variable “Height, cm” histogram (adjusted), dataset “Plant data”

Then, repeated tests of Shapiro-Wilk and Anderson-Darling were carried out to study the variables transformed by the BoxCox method. The results are presented in the Table 4.3.

Table 4.3: Test results for adjusted variables (normalized by BoxCox method)

Adjusted data frame (pd)				
Variable	**Value**		**P-value**	
	W (Shapiro–Wilk)	**A (Anderson–Darling)**	**W (Shapiro–Wilk)**	**A (Anderson–Darling)**
Height.cm.	0,98531	0,15875	0,9882	0,9392
Diameter.cm.	0,92772	0,61131	0,1771	0,09458
Prunning.firewood.mg.	0,96909	0,22043	0,7805	0,8037
N11	0,9311	0,43644	0,203	0,2646
P11	0,95286	0,32657	0,4716	0,4919
K11	0,96333	0,27059	0,6669	0,6323
Ca11	0,97832	0,17306	0,931	0,9137
Mg11	0,9883	0,1365	0,9967	0,9711

Adjusted data frame (pd)				
Variable	**Value**		**P-value**	
	W (Shapiro–Wilk)	**A (Anderson–Darling)**	**W (Shapiro–Wilk)**	**A (Anderson–Darling)**
B12	0,92621	0,56364	0,1665	0,1233
Cu12	0,93802	0,40691	0,268	0,3137
Fe12	0,96357	0,3397	0,6718	0,4653
Zn12	0,96848	0,25276	0,7686	0,6945
Mn12	0,94438	0,43307	0,3437	0,2698
N21	0,91223	0,79204	0,0942	0,03222
P21	0,98403	0,15101	0,982	0,9517
K21	0,96223	0,23335	0,6451	0,7619
Ca21	0,9692	0,19805	0,7826	0,8667
Mg21	0,97606	0,20519	0,9006	0,8481
B22	0,94366	0,39368	0,3305	0,3383
Cu22	0,82283	1,6745	0,003252	0,0001711
Fe22	0,96742	0,27527	0,7481	0,6167
Zn22	0,98464	0,18848	0,9851	0,8876
Mn22	0,96329	0,32161	0,6661	0,5045
N31	0,96849	0,20325	0,7688	0,8533
P31	0,96081	0,39734	0,6173	0,3313
K31	0,96905	0,20639	0,7797	0,8449
Ca31	0,97834	0,20144	0,9313	0,8581
Mg31	0,94492	0,36765	0,351	0,3918
B32	0,9204	0,58114	0,1314	0,1132
Cu32	0,91254	0,66384	0,09541	0,06916
Fe32	0,95311	0,23815	0,4759	0,7455
Zn32	0,96966	0,25222	0,7913	0,6964

Adjusted data frame (pd)				
Variable	**Value**		**P-value**	
	W (Shapiro–Wilk)	**A (Anderson–Darling)**	**W (Shapiro–Wilk)**	**A (Anderson–Darling)**
Mn32	0,9679	0,27299	0,7574	0,6242
N41	0,96624	0,27685	0,7247	0,6115
P41	0,97851	0,16545	0,9333	0,9274
K41	0,9069	0,63543	0,07589	0,08192
Ca41	0,95039	0,27713	0,4312	0,6106
Mg41	0,96656	0,25591	0,7311	0,6833
B42	0,91881	0,63369	0,1231	0,08277
Cu42	0,81013	1,6704	0,002119	0,0001753
Fe42	0,9768	0,15233	0,9112	0,9497
Zn42	0,97036	0,25182	0,8044	0,6978
Mn42	0,94209	0,4234	0,3145	0,2853
SPAD.30.05.2014	0,95014	0,39807	0,4271	0,3299
SPAD.28.06.2014	0,82259	1,4419	0,003225	0,0006782
SPAD.07.08.2014	0,95525	0,2647	0,5132	0,6525
SPAD.24.07.2015	0,96686	0,353379	0,737	0,4233
SPAD.15.07.2016	0,96576	0,37358	0,7152	0,379
SPAD.22.08.2017	0,95133	0,32982	0,4463	0,4832
NDVI.30.05.2014	0,87247	0,84441	0,01955	0,2359
NDVI.28.06.2014	0,79674	1,6631	0,001367	0,0001831
NDVI.07.08.2014	0,81593	1,6782	0,002573	0,0001674
NDVI.24.07.2015	0,93862	0,48618	0,2744	0,1974

After transformation according to the Shapiro-Wilk test, the main hypothesis (the distribution of the values of the variable is normal) was rejected already for 6 variables (“Cu22”, “Cu42”, “SPAD, 28.06.2014”, “NDVI, 30.05.2014”, “NDVI, 28.06.2014”,

"NDVI, 07.08.2014"), according to the Anderson-Darling criterion also 6 ("N21", "Cu22", "Cu42", "SPAD, 28.06.2014", "NDVI, 28.06.2014", "NDVI, 07.08.2014").

4.2 Applying of methods of Mathematical Statistics

MLR models were constructed. As the intercept (independent variable) were chosen: "Pruning, mg", "Height, cm" and "Diameter, cm". The test results for the intercept "Pruning, mg" are presented in the Table 4.4.

Table 4.4: Results of testing MLR models for the "Pruning, mg" intercept

#	The Model	R2	R2 (adj)
1	Pruning, mg ~ Height, cm + Diameter, cm + Reiteration + Height, cm * Diameter, cm + Height, cm * Reiteration + Diameter, cm * Reiteration	0,6436	0,2426
2	Pruning, mg ~ Height, cm + Diameter, cm + Reiteration + Height, cm * Diameter, cm + Height, cm * Reiteration	0,6157	0,3466
3	Pruning, mg ~ Height, cm + Diameter, cm + Reiteration + Height, cm * Diameter, cm + Diameter, cm * Reiteration	0,5671	0,2641
4	Pruning, mg ~ Height, cm + Diameter, cm + Reiteration + Height, cm * Diameter, cm	0,5658	0,3849
5	Pruning, mg ~ Height, cm + Diameter, cm + Reiteration	0,5583	0,4223
6	Pruning, mg ~ Height, cm + Diameter, cm + Height, cm * Diameter, cm	0,5512	0,4551
7	Pruning, mg ~ Height, cm + Diameter, cm	0,5485	0,4883
8	Pruning, mg ~ Reiteration + Height, cm * Reiteration + Diameter, cm * Reiteration	0,6432	0,3261
9	Pruning, mg ~ Height, cm + Height, cm * Diameter, cm	0,5512	0,4551
10	Pruning, mg ~ Diameter, cm + Height, cm * Diameter, cm	0,5512	0,4551
11	Pruning, mg ~ Reiteration + Height, cm	0,4981	0,3906
12	Pruning, mg ~ Reiteration + Diameter, cm	0,501	0,3941
13	Pruning, mg ~ Reiteration + Height, cm * Reiteration	0,5539	0,368
14	Pruning, mg ~ Reiteration + Diameter, cm * Reiteration	0,5069	0,3014
15	Pruning, mg ~ N11 + P11 + K11 + Ca11 + Mg11 + B12 +Cu12 + Fe12 + Zn12 + Mn12	0,9243	0,8163
16	Pruning, mg ~ N11 + P11 + K11 + Ca11 + Mg11 + B12 +Cu12 + Fe12 + Zn12	0,9224	0,835
17	Pruning, mg ~ N11 + P11 + K11 + Ca11 + Mg11 + B12 +Cu12 + Fe12	0,906	0,8225
18	Pruning, mg ~ N11 + P11 + K11 + Ca11 + Mg11 + B12 +Cu12	0,9036	0,8361
19	Pruning, mg ~ N11 + P11 + K11 + Ca11 + Mg11 + B12	0,8953	0,8382
20	Pruning, mg ~ N11 + P11 + K11 + Ca11 + Mg11	0,892	0,847
21	Pruning, mg ~ N11 + P11 + K11 + Ca11	0,8908	0,8572
22	Pruning, mg ~ N11 + P11 + K11	0,8479	0,8153
23	Pruning, mg ~ N11 + P11	0,8479	0,8276
116	Pruning, mg ~ NDVI, 30.05.2014 + NDVI, 28.06.2014 + NDVI, 07.08.2014 + NDVI, 24.07.2015	0,1139	-0,1587
117	Pruning, mg ~ NDVI, 30.05.2014 + NDVI, 28.06.2014 + NDVI, 07.08.2014	0,09585	-0,0979
118	Pruning, mg ~ NDVI, 30.05.2014 + NDVI, 28.06.2014	0,01384	-0,1176
119	Pruning, mg ~ SPAD, 30.05.2014 + SPAD, 28.06.2014 + SPAD, 07.08.2014 + SPAD, 24.07.2015 + SPAD, 15.07.2016 + SPAD, 22.08.2017	0,1758	-0,2737
120	Pruning, mg ~ SPAD, 30.05.2014 + SPAD, 28.06.2014 + SPAD, 07.08.2014 + SPAD, 24.07.2015 + SPAD, 15.07.2016	0,1642	-0,1841
121	Pruning, mg ~ SPAD, 30.05.2014 + SPAD, 28.06.2014 + SPAD, 07.08.2014 + SPAD, 24.07.2015	0,1617	-0,09627
122	Pruning, mg ~ SPAD, 30.05.2014 + SPAD, 28.06.2014 + SPAD, 07.08.2014	0,1502	-0,0319
123	Pruning, mg ~ SPAD, 30.05.2014 + SPAD, 28.06.2014	5,94E-05	-0,1333

The test results for the "Height, cm" intercept are presented in Table 4.5.

Table 4.5: Results of testing MLR models for the “Height, cm” intercept

#	The Model	R2	R2 (adj)
24	Height, cm ~ Reiteration + Diameter, cm	0,6237	0,5431
25	Height, cm ~ Reiteration + Diameter, cm + N21 + P21 + K21 + Ca21 + Mg21 + B22 +Cu22 + Fe22 + Zn22 + Mn22	0,9112	0,6227
26	Height, cm ~ Diameter, cm + N21 + P21 + K21 + Ca21 + Mg21 + B22 +Cu22 + Fe22 + Zn22 + Mn22	0,8944	0,7007
27	Height, cm ~ N21 + P21 + K21 + Ca21 + Mg21 + B22 +Cu22 + Fe22 + Zn22 + Mn22	0,456	-0,3211
28	Height, cm ~ N21 + P21 + K21 + Ca21 + Mg21 + B22 +Cu22 + Fe22 + Zn22	0,4558	-0,1564
29	Height, cm ~ N21 + P21 + K21 + Ca21 + Mg21 + B22 +Cu22 + Fe22	0,389	-0,1542
30	Height, cm ~ N21 + P21 + K21 + Ca21 + Mg21 + B22 +Cu22	0,387	-0,04206
31	Height, cm ~ N21 + P21 + K21 + Ca21 + Mg21 + B22	0,3784	0,03934
32	Height, cm ~ N21 + P21 + K21 + Ca21 + Mg21	0,2306	-0,09002
33	Height, cm ~ N21 + P21 + K21 + Ca21	0,2256	-0,1266
34	Height, cm ~ N21 + P21 + K21	0,06949	-0,1299
35	Height, cm ~ N21 + P21	0,06931	-0,05478
36	Height, cm ~ Reiteration + Diameter, cm + N31 + P31 + K31 + Ca31 + Mg31 + B32 +Cu32 + Fe32 + Zn32 + Mn32	0,8718	0,4553
37	Height, cm ~ Diameter, cm + N31 + P31 + K31 + Ca31 + Mg31 + B32 +Cu32 + Fe32 + Zn32 + Mn32	0,869	0,6288
38	Height, cm ~ N31 + P31 + K31 + Ca31 + Mg31 + B32 +Cu32 + Fe32 + Zn32 + Mn32	0,596	0,01876
39	Height, cm ~ N31 + P31 + K31 + Ca31 + Mg31 + B32 +Cu32 + Fe32 + Zn32	0,5902	0,1291
40	Height, cm ~ N31 + P31 + K31 + Ca31 + Mg31 + B32 +Cu32 + Fe32	0,5897	0,2249
41	Height, cm ~ N31 + P31 + K31 + Ca31 + Mg31 + B32 +Cu32	0,3801	-0,053388
42	Height, cm ~ N31 + P31 + K31 + Ca31 + Mg31 + B32	0,3731	0,03112
43	Height, cm ~ N31 + P31 + K31 + Ca31 + Mg31	0,1432	-0,2138
44	Height, cm ~ N31 + P31 + K31 + Ca31	0,1254	-0,1438
45	Height, cm ~ N31 + P31 + K31	0,1169	-0,07229
46	Height, cm ~ N31 + P31	0,07037	0,05358
47	Height, cm ~ Reiteration + Diameter, cm + N41 + P41 + K41 + Ca41 + Mg41 + B42 +Cu42 + Fe42 + Zn42 + Mn42	0,9924	0,9675
48	Height, cm ~ Diameter, cm + N41 + P41 + K41 + Ca41 + Mg41 + B42 +Cu42 + Fe42 + Zn42 + Mn42	0,884	0,6712
49	Height, cm ~ N41 + P41 + K41 + Ca41 + Mg41 + B42 +Cu42 + Fe42 + Zn42 + Mn42	0,659	0,1718
50	Height, cm ~ N41 + P41 + K41 + Ca41 + Mg41 + B42 +Cu42 + Fe42 + Zn42	0,5268	-0,005475
51	Height, cm ~ N41 + P41 + K41 + Ca41 + Mg41 + B42 +Cu42 + Fe42	0,5071	0,06901
52	Height, cm ~ N41 + P41 + K41 + Ca41 + Mg41 + B42 +Cu42	0,5071	0,1621
53	Height, cm ~ N41 + P41 + K41 + Ca41 + Mg41 + B42	0,3175	-0,05483
54	Height, cm ~ N41 + P41 + K41 + Ca41 + Mg41	0,254	-0,05678
55	Height, cm ~ N41 + P41 + K41 + Ca41	0,1468	-0,1158
56	Height, cm ~ N41 + P41 + K41	0,08882	-0,1064
57	Height, cm ~ N41 + P41	0,02441	-0,1057
59	Height, cm ~ Reiteration + Diameter, cm + SPAD, 30.05.2014 + SPAD, 28.06.2014 + SPAD, 07.08.2014 + SPAD, 24.07.2015 + SPAD, 15.07.2016 + SPAD, 22.08.2017	0,787	0,5474
60	Height, cm ~ Diameter, cm + SPAD, 30.05.2014 + SPAD, 28.06.2014 + SPAD, 07.08.2014 + SPAD, 24.07.2015 + SPAD, 15.07.2016 + SPAD, 22.08.2017	0,7856	0,6355
61	Height, cm ~ SPAD, 30.05.2014 + SPAD, 28.06.2014 + SPAD, 07.08.2014 + SPAD, 24.07.2015 + SPAD, 15.07.2016 + SPAD, 22.08.2017	0,2584	-0,1461
62	Height, cm ~ SPAD, 30.05.2014 + SPAD, 28.06.2014 + SPAD, 07.08.2014 + SPAD, 24.07.2015 + SPAD, 15.07.2016	0,2229	-0,1008
63	Height, cm ~ SPAD, 30.05.2014 + SPAD, 28.06.2014 + SPAD, 07.08.2014 + SPAD, 24.07.2015	0,2056	-0,03879
64	Height, cm ~ SPAD, 30.05.2014 + SPAD, 28.06.2014 + SPAD, 07.08.2014	0,2033	0,03255
65	Height, cm ~ SPAD, 30.05.2014 + SPAD, 28.06.2014	0,05823	-0,06734
66	Height, cm ~ Reiteration + Diameter, cm + NDVI, 30.05.2014 + NDVI, 28.06.2014 + NDVI, 07.08.2014 + NDVI, 24.07.2015	0,6746	0,4469
67	Height, cm ~ Diameter, cm + NDVI, 30.05.2014 + NDVI, 28.06.2014 + NDVI, 07.08.2014 + NDVI, 24.07.2015	0,671	0,5339
68	Height, cm ~ NDVI, 30.05.2014 + NDVI, 28.06.2014 + NDVI, 07.08.2014 + NDVI, 24.07.2015	0,1624	-0,09533
69	Height, cm ~ NDVI, 30.05.2014 + NDVI, 28.06.2014 + NDVI, 07.08.2014	0,1585	-0,02181
70	Height, cm ~ NDVI, 30.05.2014 + NDVI, 28.06.2014	0,006337	-0,1262

The results of testing MLR models for the “Diameter, cm” intercept are presented in the Table 4.6.

Table 4.6: Results of testing MLR models for the “Diameter, cm” intercept

#	The Model	R2	R2 (adj)
71	Diameter, cm ~ Reiteration + Height, cm + N21 + P21 + K21 + Ca21 + Mg21 + B22 + Cu22 + Fe22 + Zn22 + Mn22	0,9138	0,6338
72	Diameter, cm ~ Height, cm + N21 + P21 + K21 + Ca21 + Mg21 + B22 + Cu22 + Fe22 + Zn22 + Mn22	0,8948	0,702
73	Diameter, cm ~ N21 + P21 + K21 + Ca21 + Mg21 + B22 + Cu22 + Fe22 + Zn22 + Mn22	0,4584	-0,3153
74	Diameter, cm ~ N21 + P21 + K21 + Ca21 + Mg21 + B22 + Cu22 + Fe22 + Zn22	0,4559	-0,1562
75	Diameter, cm ~ N21 + P21 + K21 + Ca21 + Mg21 + B22 + Cu22 + Fe22	0,3107	-0,302
76	Diameter, cm ~ N21 + P21 + K21 + Ca21 + Mg21 + B22 + Cu22	0,9879	-0,1722
77	Diameter, cm ~ N21 + P21 + K21 + Ca21 + Mg21 + B22	0,2192	-0,2067
78	Diameter, cm ~ N21 + P21 + K21 + Ca21 + Mg21	0,107	-0,2651
79	Diameter, cm ~ N21 + P21 + K21 + Ca21	0,1065	-0,1684
80	Diameter, cm ~ N21 + P21 + K21	0,1021	-0,09034
81	Diameter, cm ~ N21 + P21	0,07034	-0,053362
82	Diameter, cm ~ Reiteration + Height, cm + N31 + P31 + K31 + Ca31 + Mg31 + B32 + Cu32 + Fe32 + Zn32 + Mn32	0,8599	0,4046
83	Diameter, cm ~ Height, cm + N31 + P31 + K31 + Ca31 + Mg31 + B32 + Cu32 + Fe32 + Zn32 + Mn32	0,8432	0,5558
84	Diameter, cm ~ N31 + P31 + K31 + Ca31 + Mg31 + B32 + Cu32 + Fe32 + Zn32 + Mn32	0,5166	-0,1741
85	Diameter, cm ~ N31 + P31 + K31 + Ca31 + Mg31 + B32 + Cu32 + Fe32 + Zn32	0,5035	-0,05516
86	Diameter, cm ~ N31 + P31 + K31 + Ca31 + Mg31 + B32 + Cu32 + Fe32	0,4932	0,04268
87	Diameter, cm ~ N31 + P31 + K31 + Ca31 + Mg31 + B32 + Cu32	0,3303	-0,1385
88	Diameter, cm ~ N31 + P31 + K31 + Ca31 + Mg31 + B32	0,3097	-0,06675
89	Diameter, cm ~ N31 + P31 + K31 + Ca31 + Mg31	0,2027	-0,1295
90	Diameter, cm ~ N31 + P31 + K31 + Ca31	0,177	-0,07623
91	Diameter, cm ~ N31 + P31 + K31	0,1585	-0,02183
92	Diameter, cm ~ N31 + P31	0,01801	-0,1129
93	Diameter, cm ~ Reiteration + Height, cm + N41 + P41 + K41 + Ca41 + Mg41 + B42 + Cu42 + Fe42 + Zn42 + Mn42	0,9773	0,9037
94	Diameter, cm ~ Height, cm + N41 + P41 + K41 + Ca41 + Mg41 + B42 + Cu42 + Fe42 + Zn42 + Mn42	0,7816	0,3812
95	Diameter, cm ~ N41 + P41 + K41 + Ca41 + Mg41 + B42 + Cu42 + Fe42 + Zn42 + Mn42	0,3582	-0,5587
96	Diameter, cm ~ N41 + P41 + K41 + Ca41 + Mg41 + B42 + Cu42 + Fe42 + Zn42	0,2789	-0,5324
97	Diameter, cm ~ N41 + P41 + K41 + Ca41 + Mg41 + B42 + Cu42 + Fe42	0,271	-0,377
98	Diameter, cm ~ N41 + P41 + K41 + Ca41 + Mg41 + B42 + Cu42	0,2705	-0,2402
99	Diameter, cm ~ N41 + P41 + K41 + Ca41 + Mg41 + B42	0,2345	-0,1831
100	Diameter, cm ~ N41 + P41 + K41 + Ca41 + Mg41	0,2343	-0,08468
101	Diameter, cm ~ N41 + P41 + K41 + Ca41	0,1945	-0,05337
102	Diameter, cm ~ N41 + P41 + K41	0,1915	0,01827
103	Diameter, cm ~ N41 + P41	0,01612	-0,1151
104	Diameter, cm ~ Reiteration + Height, cm + SPAD, 30.05.2014 + SPAD, 28.06.2014 + SPAD, 07.08.2014 + SPAD, 24.07.2015 + SPAD, 15.07.2016 + SPAD, 22.08.2017	0,8082	0,5924
105	Diameter, cm ~ Height, cm + SPAD, 30.05.2014 + SPAD, 28.06.2014 + SPAD, 07.08.2014 + SPAD, 24.07.2015 + SPAD, 15.07.2016 + SPAD, 22.08.2017	0,8071	0,672
106	Diameter, cm ~ SPAD, 30.05.2014 + SPAD, 28.06.2014 + SPAD, 07.08.2014 + SPAD, 24.07.2015 + SPAD, 15.07.2016 + SPAD, 22.08.2017	0,3327	-0,0313
107	Diameter, cm ~ SPAD, 30.05.2014 + SPAD, 28.06.2014 + SPAD, 07.08.2014 + SPAD, 24.07.2015 + SPAD, 15.07.2016	0,3245	0,043
108	Diameter, cm ~ SPAD, 30.05.2014 + SPAD, 28.06.2014 + SPAD, 07.08.2014 + SPAD, 24.07.2015	0,3025	0,08799
109	Diameter, cm ~ SPAD, 30.05.2014 + SPAD, 28.06.2014 + SPAD, 07.08.2014	0,1178	-0,07122
110	Diameter, cm ~ SPAD, 30.05.2014 + SPAD, 28.06.2014	0,03863	-0,08955
111	Diameter, cm ~ Reiteration + Height, cm + NDVI, 30.05.2014 + NDVI, 28.06.2014 + NDVI, 07.08.2014 + NDVI, 24.07.2015	0,7344	0,5486
112	Diameter, cm ~ Height, cm + NDVI, 30.05.2014 + NDVI, 28.06.2014 + NDVI, 07.08.2014 + NDVI, 24.07.2015	0,7343	0,6235
113	Diameter, cm ~ NDVI, 30.05.2014 + NDVI, 28.06.2014 + NDVI, 07.08.2014 + NDVI, 24.07.2015	0,3234	0,1153
114	Diameter, cm ~ NDVI, 30.05.2014 + NDVI, 28.06.2014 + NDVI, 07.08.2014	0,2945	0,1433
115	Diameter, cm ~ NDVI, 30.05.2014 + NDVI, 28.06.2014	0,006278	-0,1262

In order for the analysis of the obtained results of testing MLR models to be adequate, the variables (intercept and predictors) were tested for multicollenarity. The results are presented in Figure 1 – 7 in the Appendix 1 (model numbering is presented in tables 4.4, 4.5 and 4.6 in this chapter).

Brief conclusions on the resutates of analysis are presented below:

1) The intercept is the “Pruning, mg” variable.

For this independent variable, 31 models were constructed (see Table 4). The variance is best determined in 9 models: 15 – 23 (with the results R2: [0,8479 ; 0,9243], R2-adjusted: [0,8153 ; 0,8572]), but the multicollinearity test showed a high value correlations between all variables involved in modeling.

The following 14 models coped worse with the explanation of variance: 1 – 14 (with the results R2: [0,4981 ; 0,6436], R2-adjusted: [0,2426 ; 0,4883]).

The worst results were in 8 models: 116 – 123 (with the results R2: [5,94E-05 ; 0.1758], R2- adjusted: [-0,2737 ; -0,0319]).

2) The intercept is the “Height, cm” variable.

For this independent variable, 46 models were constructed (see Table 5). The dispersion is best determined in 18 models: 24 – 26, 36 – 40, 47 – 52, 59 – 60, 66 – 67 according to the results of R2: [0,5071 ; 0,9924] and in 9 models: 24 – 26, 37, 47 – 48, 59 – 60, 67 according to the results of R2-adjusted: [0,5339 ; 0,9675]. Checking for multicollinearity showed a generally low and acceptable correlation between all variables involved in the simulation.

The following 15 models coped worse with the explanation of variance: 27 – 33, 41 – 42, 53 – 54, 61 – 64 based on the results of R2: [0,2033 ; 0,456] and 5 models: 36, 40, 49, 52, 66 according to the results of R2-adjusted: [0,1621 ; 0,4553].

The worst results were in 13 models: 34 – 35, 43 – 46, 55 – 57, 65, 68 – 70 according to the results of R2: [0,006337; 0.1624] and in 32 models: 27 – 35, 38 – 39, 41 – 46, 50 – 51, 53 – 57, 61 – 65, 68 – 70 according to the results of R2- adjusted: [-0,2737 ; -0,0319].

3) The intercept is the “Diameter, cm” variable.

For this independent variable, 45 models were constructed (see Table 6). The dispersion is best determined in 11 models: 71 – 72, 76, 82 – 83, 93 – 94, 104 – 105, 111 – 112 according to the results of R2: [0,7343 ; 0,9879] and in 8 models: 71 – 72, 83, 93, 104 – 105, 111 – 112 based on the results of R2-adjusted: [0,5486 ; 0,9037]. Checking for

multicollinearity showed a generally low and acceptable correlation between all variables involved in the simulation.

The following 17 models performed worse in explaining the variance: 73 – 75, 84 – 88, 95 – 98, 106 – 108, 113 – 114 according to the results of R2: [0,2705 ; 0,5166] and 2 models: 82 and 94 based on the results of R2-adjusted: [0,3812 ; 0,4046].

The worst results were in 17 models: 77 – 81, 89 – 92, 99 – 103, 109 – 110 and 115 according to the results of R2: [0,006278; 0,2345] and 35 models: 73 – 81, 84 – 92, 95 – 103, 106 – 110, 113 – 115 according to the results of R2- adjusted: [-0,5587 ; 0,1433].

4.3 Applying Machine Learning methods

As a representative of ML methods, the agglomerative HC method was studied and applied to the dataset “Plant data”. Appendix 2 contains most of the results of this method. The subjects were selected models that showed the best results (high coefficient of determination) in the analysis of MLR (see conclusions of the privios paragraph of this chapter). Model numbering is presented in tables 4.4, 4.5 and 4.6 of the privios paragraph of this chapter.

During the study of dendrograms, it was determined that some models have completely identical dendrograms.

The first example is 26 models and a common dataset model. In this dendrogram I would like to highlight 4 main clusters: 15-18, 13-14, 2-10, 11-7 (from left to right).

The second example is dendrograms of models 37 and 39. In this example, 2 main clusters are clearly distinguished: 15-16 and 3-12 (from left to right). The dendrogram of this group of models is shown in Figure 4.4.

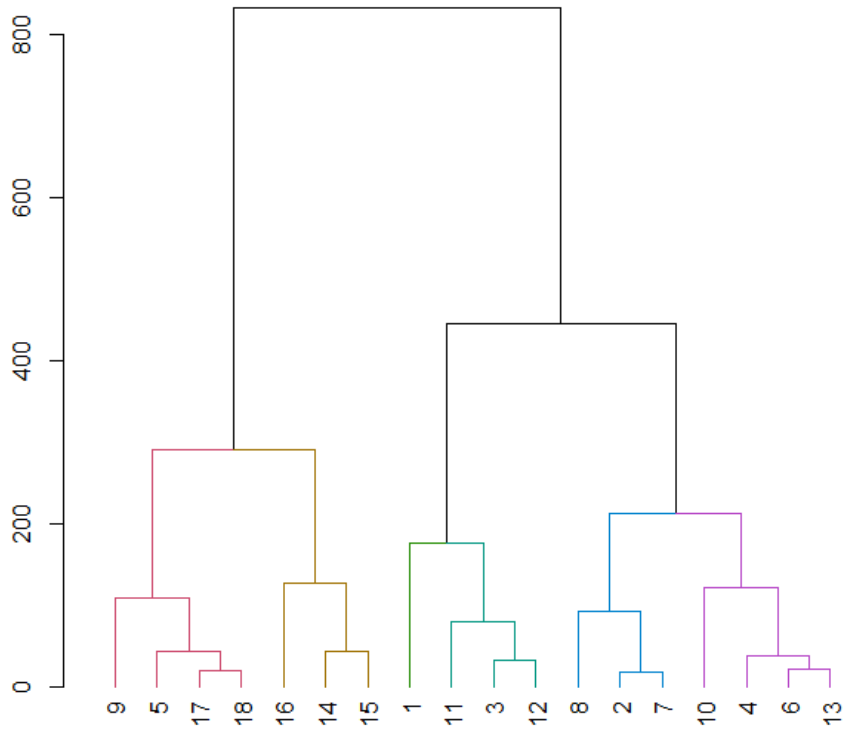

Figure 4.4: Model 16 HC of the Plant Data dataset

In the third example, 3 models are identical at once: 48, 60 and 67. The dendrograms of the models in example three also clearly distinguish 2 main clusters: 15-14 and 3-12 (from left to right).

Models 16 and 76 turned out to be unique. In model 16, 2 main clusters (9-15 and 1-13) and 5 subclusters (9-18, 16-15, 1-12, 8-7, 10-13) can be clearly distinguished with a deeper section of the dendrogram (also on the left right). Model 76 can be primarily divided into 3 main clusters (17-18, 13-9, 8-10) and 6 sub-clusters (17-18, 13-9, 8-12, 14-11, 2-6, 15-10) with a deeper cut of the dendrogram (also from left to right), the first 2 of which are also the main clusters.

Also in Appendix 2, for each model, graphs of the distribution of the distances between the values of the variables. For example, distribution of distances between observations of model 16 is shown in Figure 4.5.

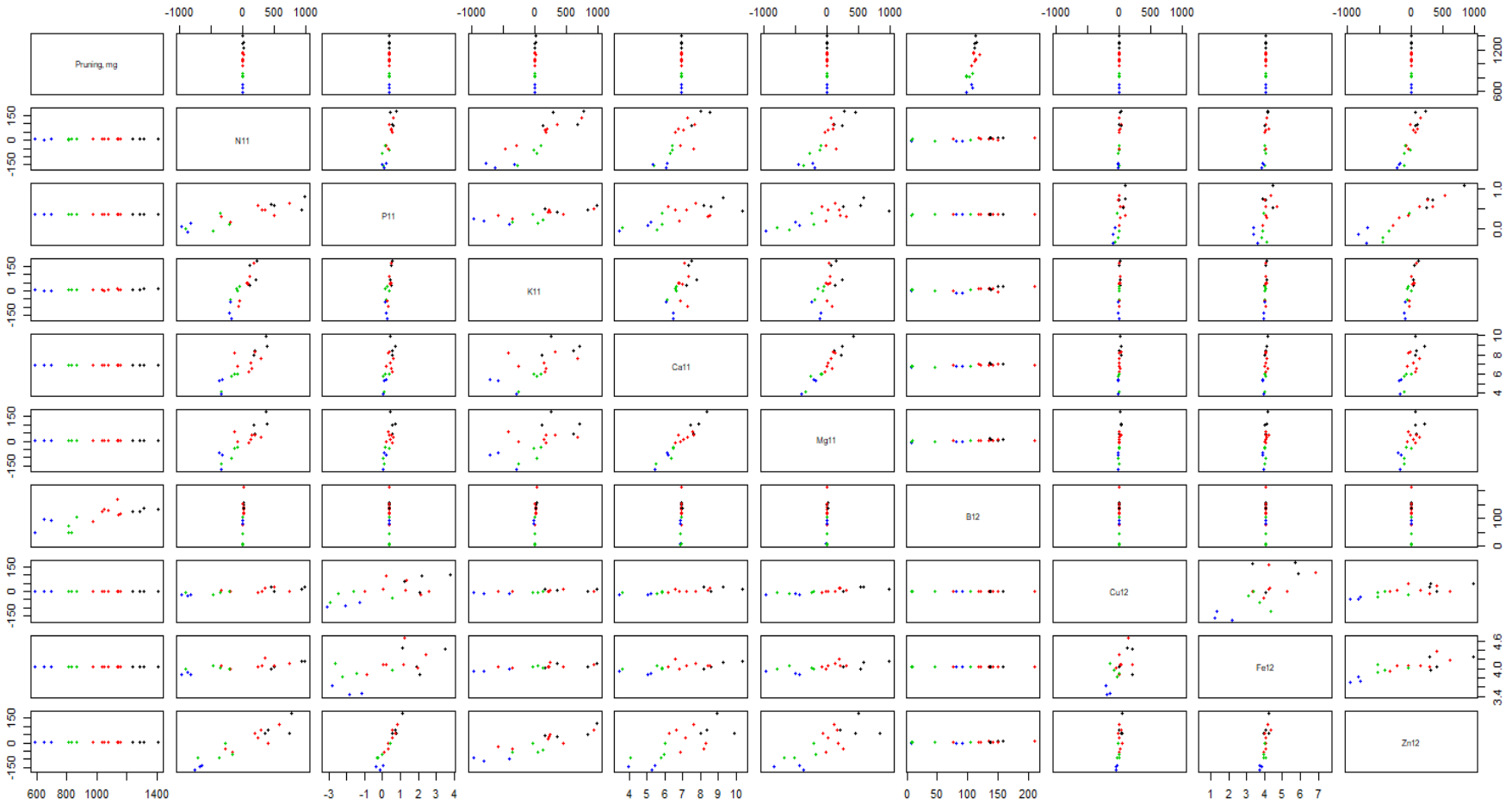

Figure 4.5: Distribution of distances between observations of model 16 of dataset “Plant Data”

4.4 Chapter Summary

So, in this chapter, the methods of classical MS (MLR) and ML (Agglomerative HC) were applied on the same dataset.

The testing of MLR models allowed us to draw conclusions regarding the effect of a set of nutrients on some characteristics of seedlings: the height of the stem, its width, as well as the amount of trimming of seedlings after four years of cultivation, provided that they are fed with the above set of nutrients. Although, as already written above, one cannot ignore the high correlation between the variables of the studied models, which was detected during the verification of the models for multicollinearity.

The study of the same dataset by the method of HC made it possible to find out how many clusters can be distinguished in a particular model, as well as to estimate the magnitude of the distances between observations of various variables.

Summing up the results of this chapter, we can confidently say that the ability to study data using methods of both classical statistics and ML allows you to expand the range of useful information extracted during the analysis. I also want to note that only the tool that combines both of these approaches in the logic of its work can be considered a sufficiently adaptive and modern DST.

Chapter 5

Design of a Decision Support Tool

In the previous chapters, devoted to the theory of methods and directly to practical implementation, the dataset “Plant Data” was used for analysis using both the MLR method and the HC method. In order to remain within the same research framework for this work, the following decision was made: the DST will be designed taking into account the model variables contained in this date set, namely:

1) “Pruning, mg”;
2) “Height, cm”;
3) “Diameter, cm”;
4) “N11”, “N21”, “N31”, “N41”;
5) “P11”, “P21”, “P31”, “P41”;
6) “K11”, “K21”, “K31”, “K41”;
7) “Ca11”, “Ca21”, “Ca31”, “Ca41”;
8) “Mg11”, “Mg21”, “Mg31”, “Mg41”;
9) “B12”, “B22”, “B32”, “B42”;

10) “Cu12”, “Cu22”, “Cu32”, “Cu42”;

11) “Fe12”, “Fe22”, “Fe32”, “Fe42”;

12) “Zn12”, “Zn22”, “Zn32”, “Zn42”;

13) “Mn12”, “Mn22”, “Mn32”, “Mn42”;

14) “NDVI, 30.05.2014”, “NDVI, 28.06.2014”, “NDVI, 07.08.2014”, “NDVI, 24.07.2015”;

15) “SPAD, 30.05.2014”, “SPAD, 28.06.2014”, “SPAD, 07.08.2014”, “SPAD, 24.07.2015”, “SPAD, 15.07.2016”, “SPAD, 22.08.2017”.

This means that the projected DST will involve processes related to monitoring both the concentration of the above nutrients (N, P, K, etc.) and soil moisture, as well as processes that affect the timely adjustment of these factors [22].

5.1 Context modeling of the general scheme of the tool

This should start, of course, with the general scheme of operation of this tool. It is shown in Figure 5.1.

Figure 5.1: Diagram of the DST main inputs and outputs

The chart has a clear separation between inputs and outputs.

In the input data:

- information about the soil - here we mean the concentration of nutrients at a certain depth of the soil (5, 10, 15 cm);
- information about the plant - the type of crop, and the main parameters (plant height, trunk diameter, etc.);
- other information - using the weather forecast as an example, as this can directly influence the watering recommendation for a segment or the entire field.

The following are the results of the tool - outputs such as:

- information on request - reports and recommendations that are generated after a direct request from the operator;
- periodic information - weekly or monthly reports, which are required without fail to compile adequate statistics on the processes under study;
- emergency information - critical situations (washing out a large amount of nutrient necessary for growth and development, prolonged drought, which causes a critical level of moisture in the soil, etc.).

The main functions of this tool include analyzing input information and providing reports and recommendations to decision-makers [23, 24].

5.2 Modeling the flowchart for the output

The sequence of user actions to obtain the output of the tool (report or recommendation) is shown in Figure 5.2.

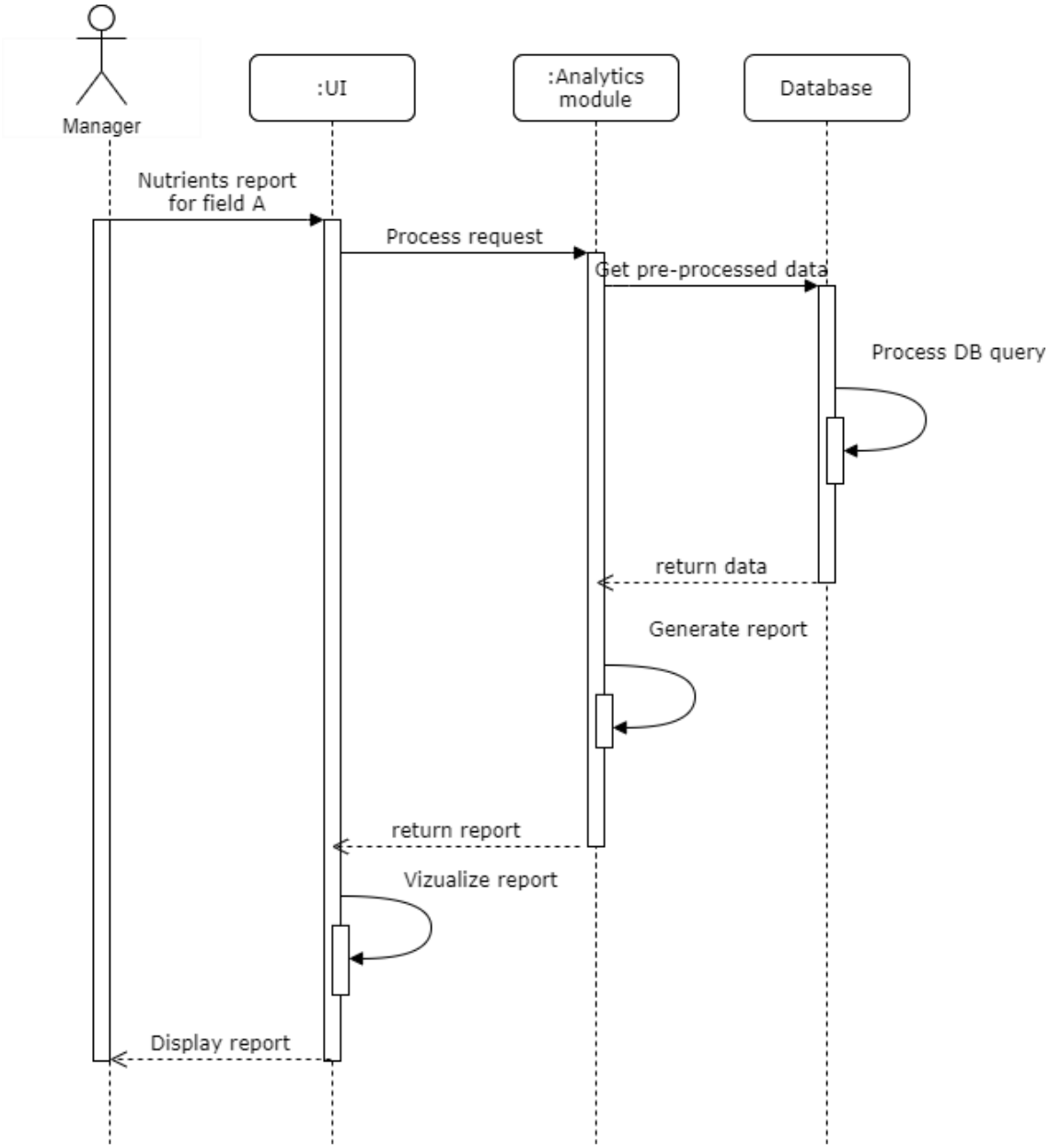

Figure 5.2: Tool sequence diagram

In order to initiate the process of receiving the output data, the operator must select the type of report or recommendation by interacting with the user interface. Next, there will be a request to generate a report to the Analytical Module [25, 26]. A request to obtain data for analysis will be initiated from the analytical module to the database. After receiving the analysis results, the report will be generated for the analytics module. This will then be rendered into the user interface and displayed on the user's screen.

5.3 Modeling the general structure of the tool

In order to represent the main scheme of work of the entire developed DST, a general structure was created, which is shown in Figure 5.3.

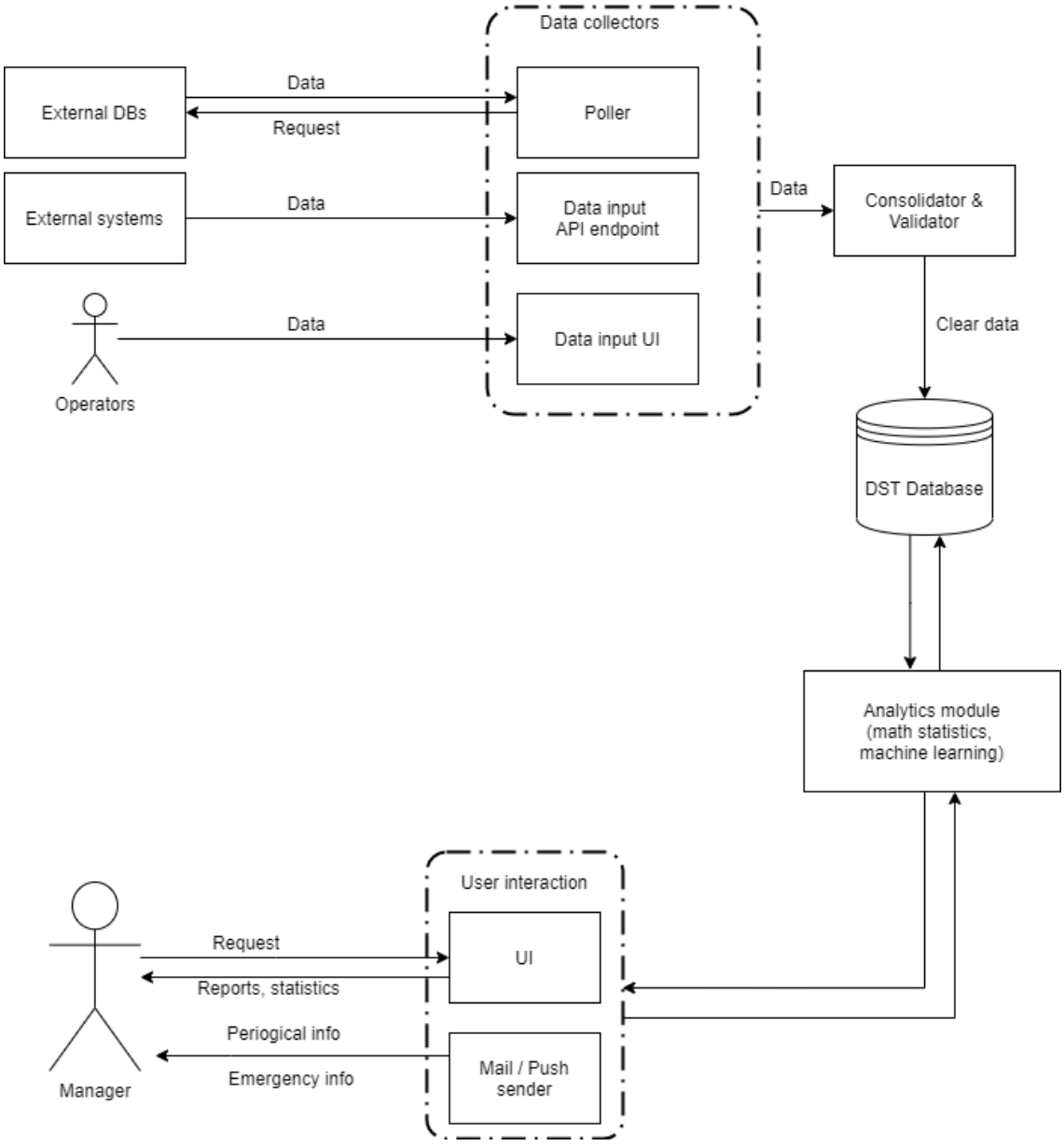

Figure 5.3: General structure scheme of the DST

The diagram shows 2 types of user interaction with the tool:

- manual input of input information into the data collection system;
- receiving reports and recommendations on request, at a specified frequency or in emergency situations.

The schema contains a group of Data Collectors – units Poller, Data input/API endpoint and Data input UI. Poller interacts with external databases to obtain input information (for example, about the concentration of nutrients or the level of moisture in the soil) Data input/API endpoint interacts with external systems (for example, irrigation and fertilization systems). And Data input performs the function of the receiver of user data by analogy with Poller.

After the input data has been received, they are sent to the work of the Consolidator & Validator unit to perform consonant functions with the data - consolidation and validation. Further, the clean data that has passed the pre-preparation stage is sent to the database of the DST [27].

On the other hand, the chain of processes associated with the DB DST (Data Base DST) is two-sided and looks as follows. The user interacts with a group of units User Interaction, which contains units of the UI (User Interface) and Mail/Push Sender. The user sends a request to receive a report or recommendation through the User Interface, then the request goes through the Analytical module, where it gets into the database in a transformed form. Then the response to the request goes through the same chain and is returned in the form corresponding to the request. The unit transmits to the user periodic or emergency reports via Mail/Push Sender using e-mail or in the form of push notifications, respectively, the request for which is initiated by the internal system of the tool.

5.4 Chapter Summary

The design of the DST in this chapter was carried out with a view to further development and possible integration into a pilot training farm in order to assist farmers in integrating PF elements into their small business.

Chapter 6

Conclusions and future work

To solve the problem posed at the beginning of writing the dissertation, the following goals were realized:

1) investigated the current state of affairs in the field of intellectual farming, including:
    a) studied the main components affecting the condition of the soil;
    b) technologies and technical PF means;
    c) studied the impact of public events on the development of scientific progress in the field of intellectual farming;
    d) researched the works of researchers involved in the study of related issues;
2) the theoretical foundations of some methods of MS (ANOVA analysis of variance and MLR) and ML (K-means clustering, HC, DT method, and the ANN approach) were studied;
3) the advantages and disadvantages of the above methods were analyzed;
4) these methods were implemented in the study of a specific set of data;

a) preparation and cleaning of data from the set was made;

b) a detailed description of the analysis results was made both by the method of MS and by the method of ML;

5) conclusions were drawn about the diversity of the analysis results and their possible practical application;

6) a DST was also designed based on the findings on the target suitability of these methods, which will be able to assist farmers in integrating some elements of PF into their pilot training farms.

A possible future of work lies in the implementation of already designed DST models. This will be followed by testing on a small experimental plantation. If the testing is successful, it is possible to integrate this tool into real fields and plantations.

Bibliography

[1] Margarida Arrobas, Sandra Afonso, Isabel Q. Ferreira, José Moutinho–Pereira, Carlos M. Correia, M. Ângelo Rodrigues, 2017. Liming and application of nitrogen, phosphorus, potassium, and boron on a young plantation of chestnut, Turk J Agric For 41: 441–451.

[2] Margarida Arrobas, Sandra Afonso, M. Ângelo Rodrigues, 2018. Diagnosing the nutritional condition of chestnut groves by soil and leaf analyses, Scientia Horticulturae 228: 113–121.

[3] Manuel Ângelo Rodrigues, João Ilídio Lopes, Isabel Queirós Ferreira, Margarida Arrobas, 2018. Olive tree response to the severity of pruning, Turk J Agric For 42: 103–113.

[4] Margarida Arrobas, Isabel Q. Ferreira, Sandra Afonso & M. Ângelo Rodrigues, 2018. Sufficiency ranges and crop nutrient removals for peppermint (Mentha X piperita L.) established from field and pot fertilizer experiments, Communications in Soil Science and Plant Analysis vol. 49, no. 14: 1719–1730.

[5] M. Ângelo Rodrigues, Sandra Afonso, Nelson Tipewa, Arlindo Almeida, Margarida Arrobas, 2019. Quantification of loss in oilseed rape yield caused by delayed sowing date in a Mediterranean environment, Archives of Agronomy and Soil Science, vol. 65, no. 12: 1630–1645.

[6] Isabel Q. Ferreira, M. Ângelo Rodrigues and Margarida Arrobas, 2019. Soil and foliar applied boron in olive: tree crop growth and yield, and boron remobilization within plant tissues, Spanish Journal of Agricultural Research, vol. 17, no. 1, e0901

[7] Margarida Arrobas, António Ribeiro, David Barreales, Ermelinda L. Pereira, M. Ângelo Rodrigues, 2019. Soil and foliar nitrogen and boron fertilization of almond trees grown under rainfed conditions, European Journal of Agronomy 106: 39–48.

[8] M. Ângelo Rodrigues, Valentim Coelho, Margarida Arrobas, Eugénia Gouveia, Soraia Raimundo, Carlos M. Correia, Albino Bento, 2019. The effect of nitrogen fertilization on the incidence of olive fruit fly, olive leaf spot and olive anthracnose in two olive cultivars grown in rainfed conditions, Scientia Horticulturae 256: 108658.

[9] Manuel Ângelo Rodrigues, Vagner Grade, Valdemar Barroso, Abel Pereira, Luís César Cassol, Margarida Arrobas, 2019. Chestnut Response to Organo-mineral and Controlled-Release Fertilizers in Rainfed Growing Conditions, Journal of Soil Science and Plant Nutrition 20: 380-391.

[10] N. Ohana-Levi, A. Ben-Gal, A. Peeters, D. Termin, R. Linker, S. Baram, E. Raveh, T. Paz-Kagan, 2020. A comparison between spatial clustering models for determining N-fertilization management zones in orchards, Precision Agriculture: An International Journal on Advances in Precision Agriculture.

[11] Takashi Kosaki & Anthony S.R. Juo, 2012. Multivariate approach to grouping soils in small fields. II. soil grouping technique by cluster analysis, Soil Science and Plant Nutrition, 35:4, 517-525, DOI: 10.1080/00380768.1989.10434787.

[12] Konstantin Potashev, Natalia Sharonova, Irina Breus, 2014. The use of cluster analysis for plant grouping by their tolerance to soil contamination with hydrocarbons at the germination stage, Science of the Total Environment: 485–486.

[13] Fionn Murtagh, Pedro Contreras, 2012. Algorithms for hierarchical clustering: an overview, WIREs Data Mining Knowl Discov 2: 86–97 doi: 10.1002/widm.53

[14] Fionn Murtagh, Pedro Contreras, 2017. Algorithms for hierarchical clustering: an overview, II, WIREs Data Mining Knowl Discov 7:e1219. doi: 10.1002/widm.1219

[15] R. Bongiovanni, J. Lowenberg-Deboer, 2004. Precision Agriculture and Sustainability, Kluwer Academic Publishers. Manufactured in The Netherlands, Precision Agriculture, 5, 359–387, 2004

[16] S. E. Cook and R. G. V. Bramley, 1998. Precision agriculture — opportunities, benefits and pitfalls of site-specific crop management in Australia, Australian Journal of Experimental Agriculture, 1998, 38, 753–63

[17] Kutter, T., Tiemann, S., Siebert, R., & Fountas, S. (2011). The role of communication and co-operation in the adoption of precision farming. Precision Agriculture, 12(1), 2-17.

[18] Adeyemi, O., Grove, I., Peets, S., & Norton, T. (2017). Advanced monitoring and management systems for improving sustainability in precision irrigation. Sustainability, 9(3), 353.

[19] Zhao, B., Ata-Ul-Karim, S. T., Liu, Z., Ning, D., Xiao, J., Liu, Z., ... & Duan, A. (2017). Development of a critical nitrogen dilution curve based on leaf dry matter for summer maize. Field Crops Research, 208, 60-68

[20] Zhang, N., Wang, M., & Wang, N. (2002). Precision agriculture—a worldwide overview. Computers and electronics in agriculture, 36(2-3), 113-132.

[21] Wolfert, S., Ge, L., Verdouw, C., & Bogaardt, M. J. (2017). Big data in smart farming–a review. Agricultural Systems, 153, 69-80.

[22] Hevner, A., March, S., Park, J., & Ram, S. (2004). Design science in information systems research. Management Information Systems Quarterly, 28(1), 75–106.

[23] Nute, D., Potter, W., Cheng, Z., Dass, M., Glende, A., Maierv, F., Routh, C., Uchiyama, H., Wang, J., Witzig, S., Twery, M., Knopp, P., Thomasma, S., & Rauscher, H. (2005). A method for integrating multiple components in a decision support system. Computers and Electronics in Agriculture, 29(1) 44–59.

[24] Dave Daas, Toine Hurkmans, Sietse Overbeek, Harry Bouwman, 2013. Developing a decision support system for business model design, Institute of Information Management, University of St. Gallen, Electron Markets (2013) 23:251–265

[25] A. Abbasi, S. Sarker, and R. Chiang. Big data research in information systems: Toward an inclusive research agenda. Journal of the Association for Information Systems, 17(2):3, 2016.

[26] D. Arnott and G. Pervan. A critical analysis of decision support systems research. Journal of information technology, 20(2):67–87, 2005.

[27] R. F. Babiceanu and R. Seker. Big data and virtualization for manufacturing cyber-physical systems: A survey of the current status and future outlook. Computers in Industry, 81:128–137, 2016.

[28] R. H. Bonczek, C. W. Holsapple, and A. B. Whinston. Foundations of decision support systems. Academic Press, 2014[29] Hierarchical clustering

[29] Hierarchical clustering, https://en.wikipedia.org/wiki/Hierarchical_clustering

[30] Greedy algorithm, https://en.wikipedia.org/wiki/Greedy_algorithm

[31] https://blog.onesoil.ai/ru/what-data-farmer-needs

[32] https://www.fendt.com/int/xaver

[33] https://www.dtn.com/agriculture/agribusiness/clearag/

[34] http://www.adapt-n.com/

[35] https://www.fendt.com/au/smart-farming/isobus

[36] https://en.wikipedia.org/wiki/Linear_regression

[37] http://rasyidh44.blogspot.com/2019/11/pengenalan-machine-learning.html

[38] https://meta.wikimedia.org/wiki/User:Sumit.iitp/Research

Appendix 1

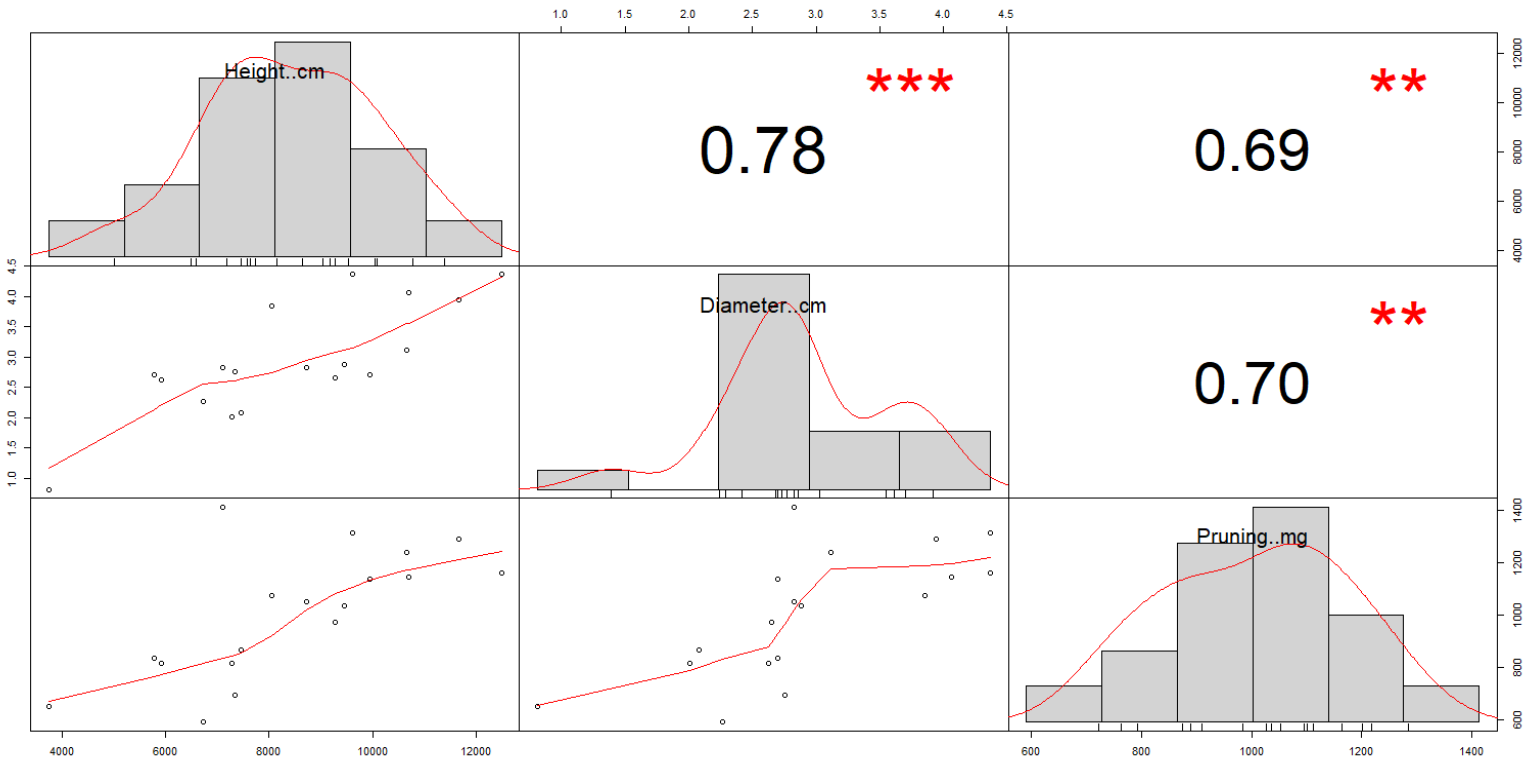

Figure 1: Checking for multicollinearity of variables in model 7

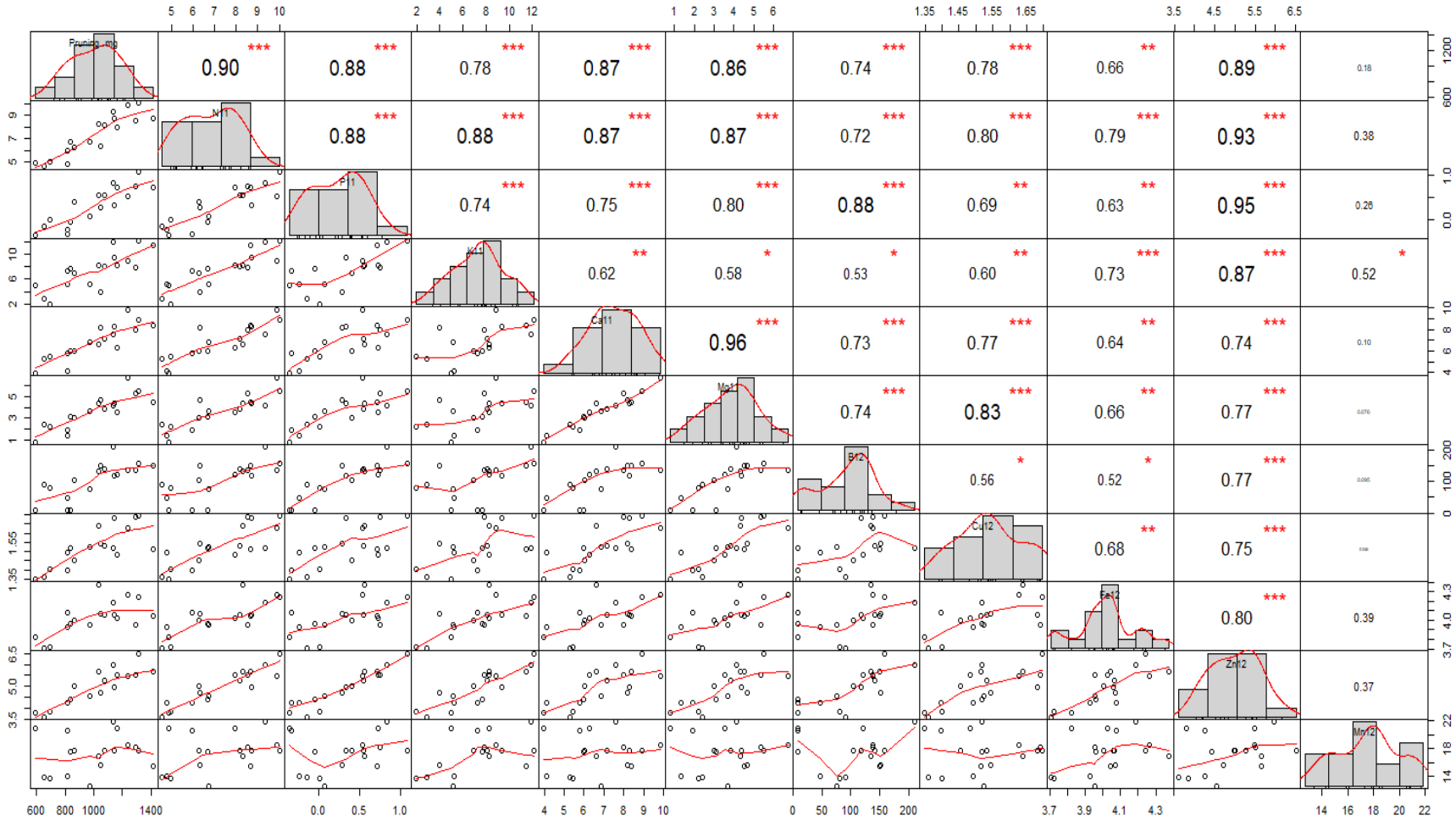

Figure 2: Checking for multicollinearity of variables in model 15

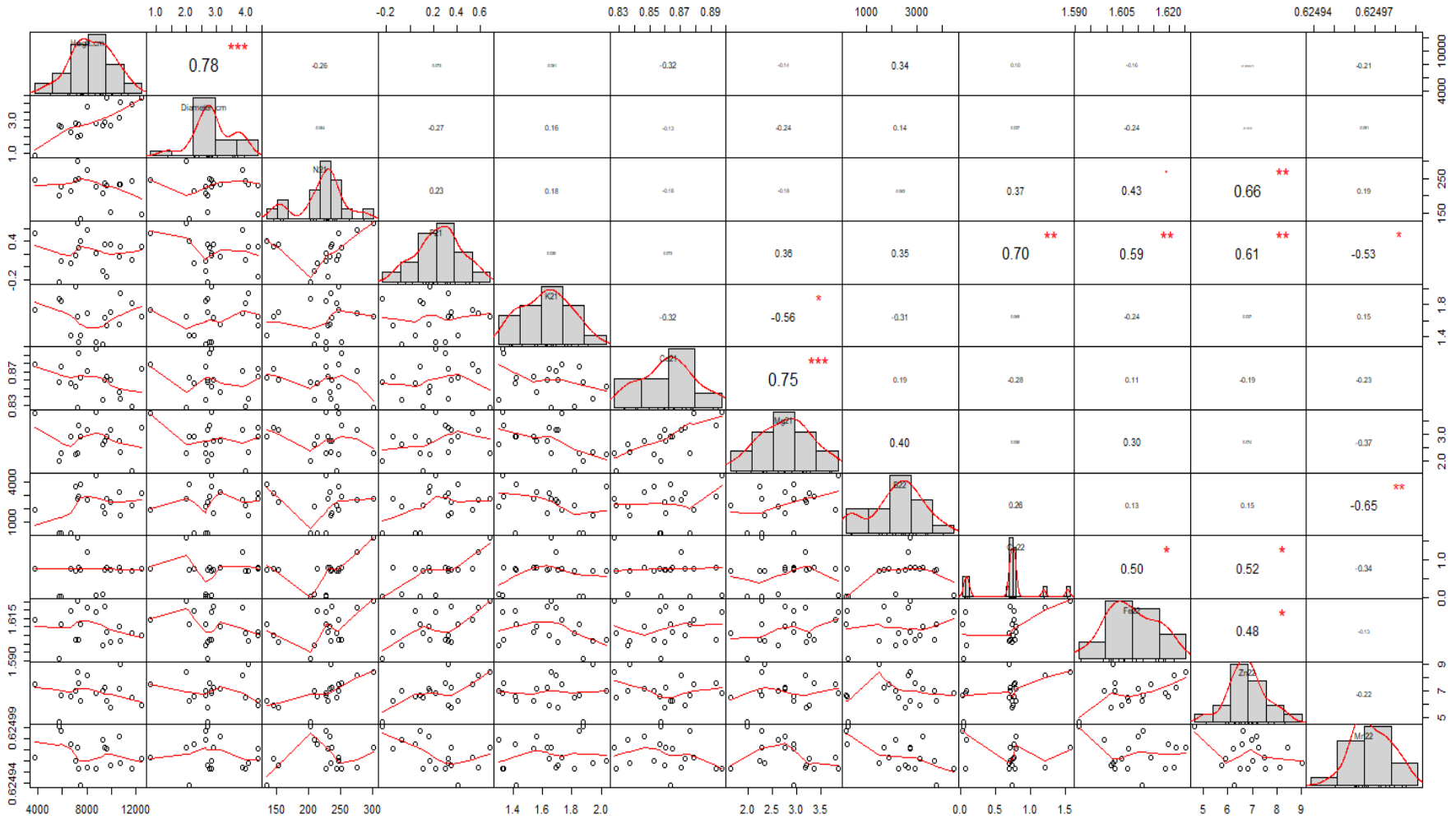

Figure 3: Checking for multicollinearity of variables in model 26

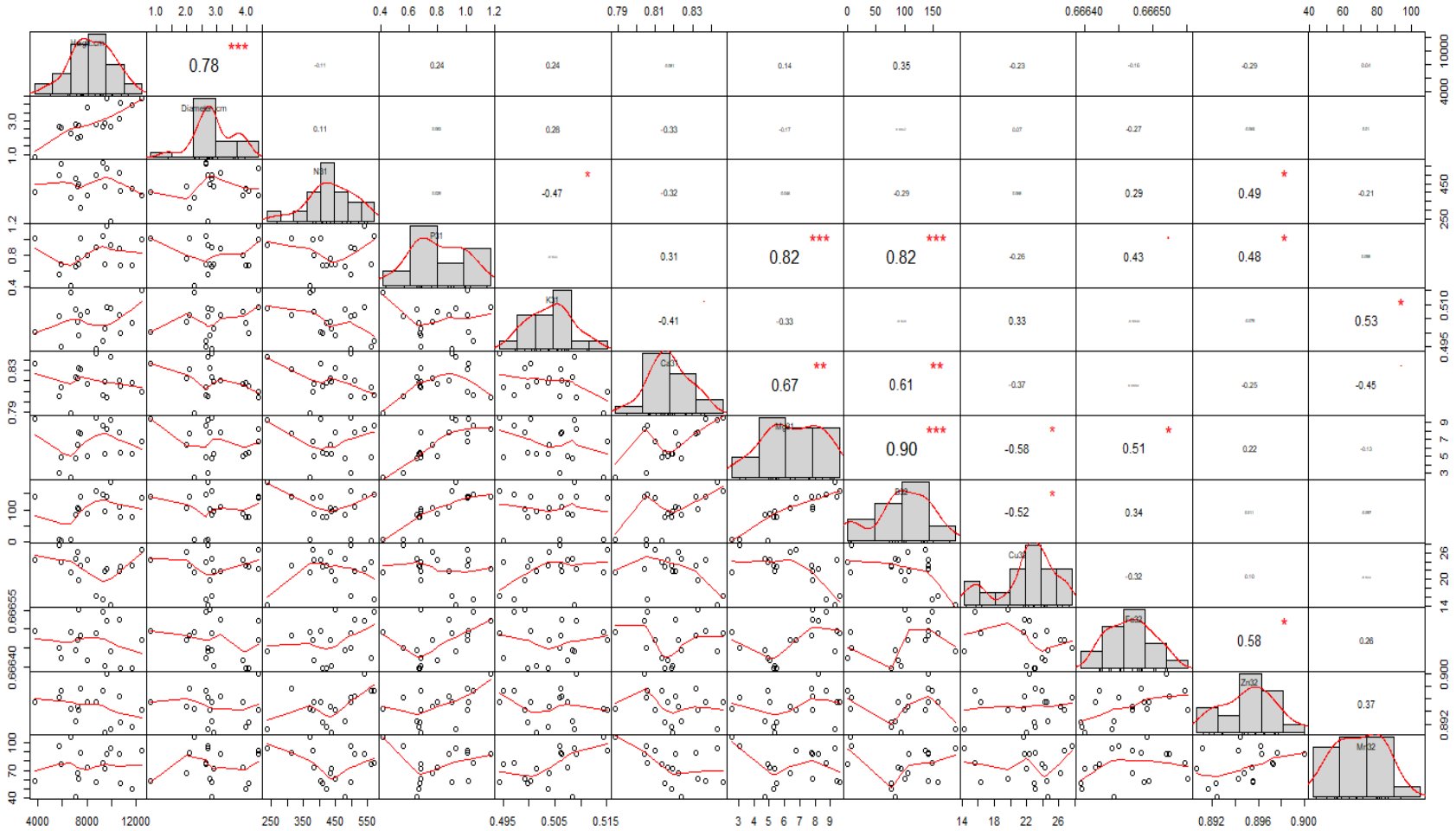

Figure 4: Checking for multicollinearity of variables in model 31

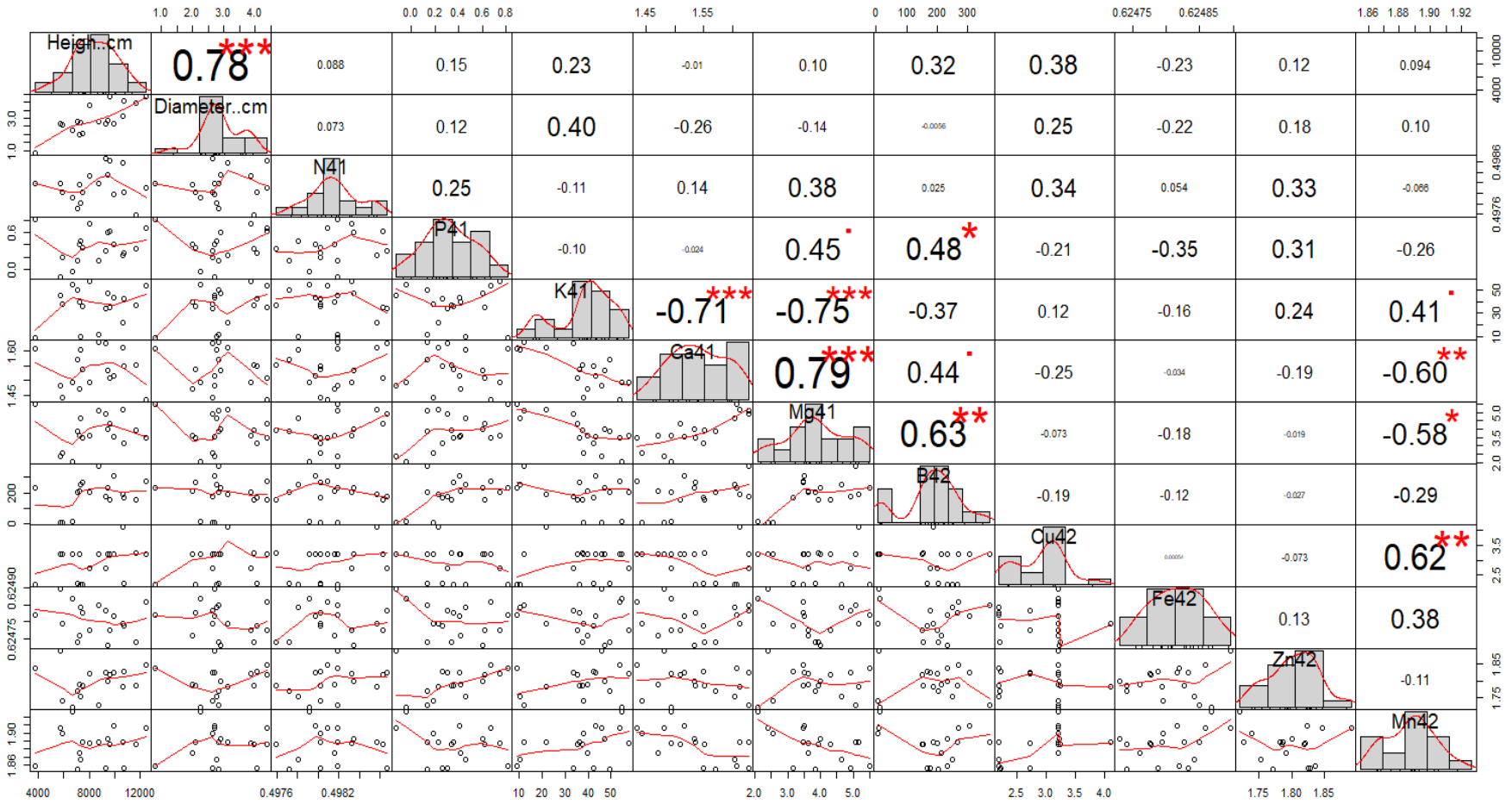

Figure 5: Checking for multicollinearity of variables in model 48

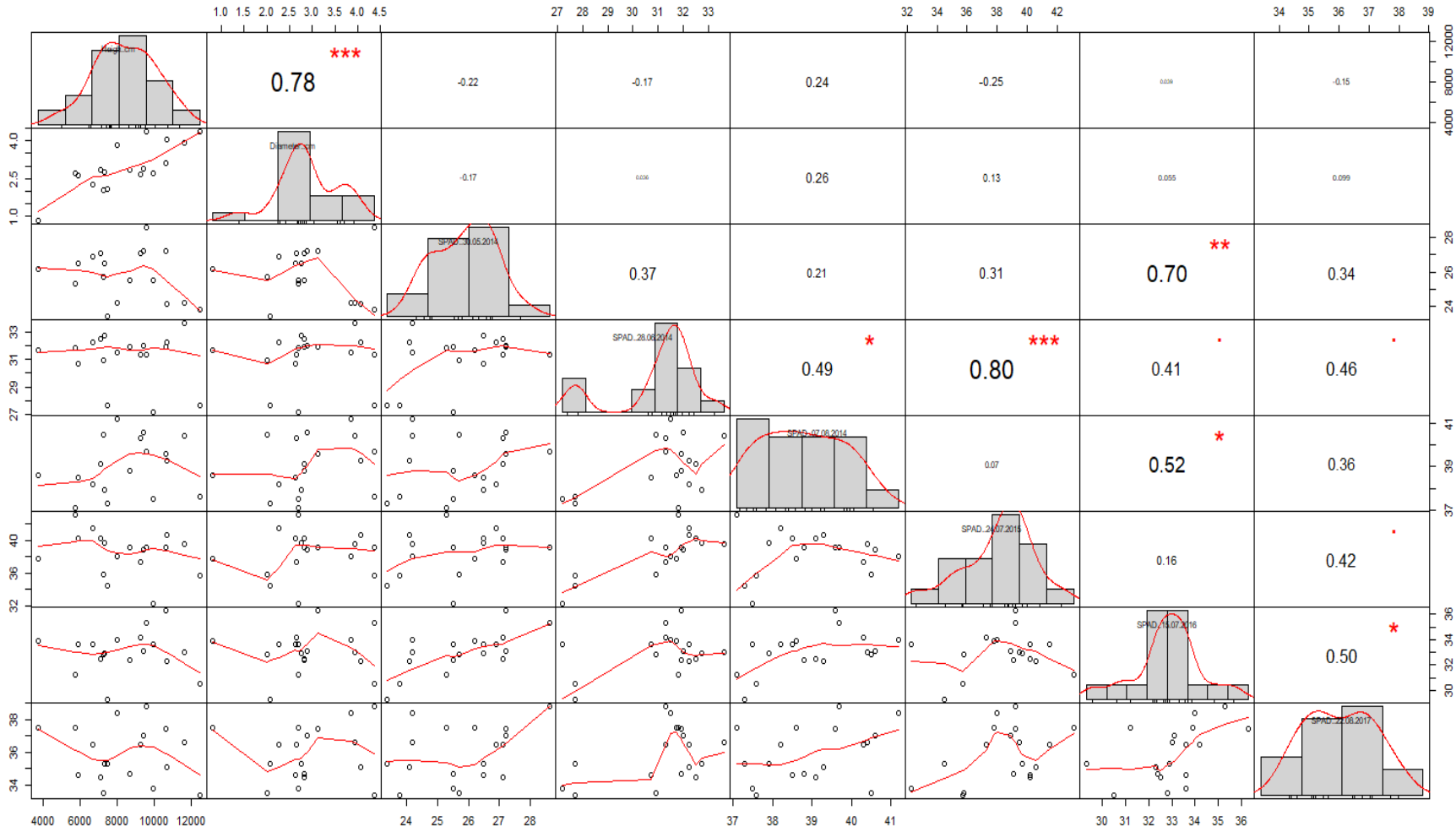

Figure 6: Checking for multicollinearity of variables in model 60

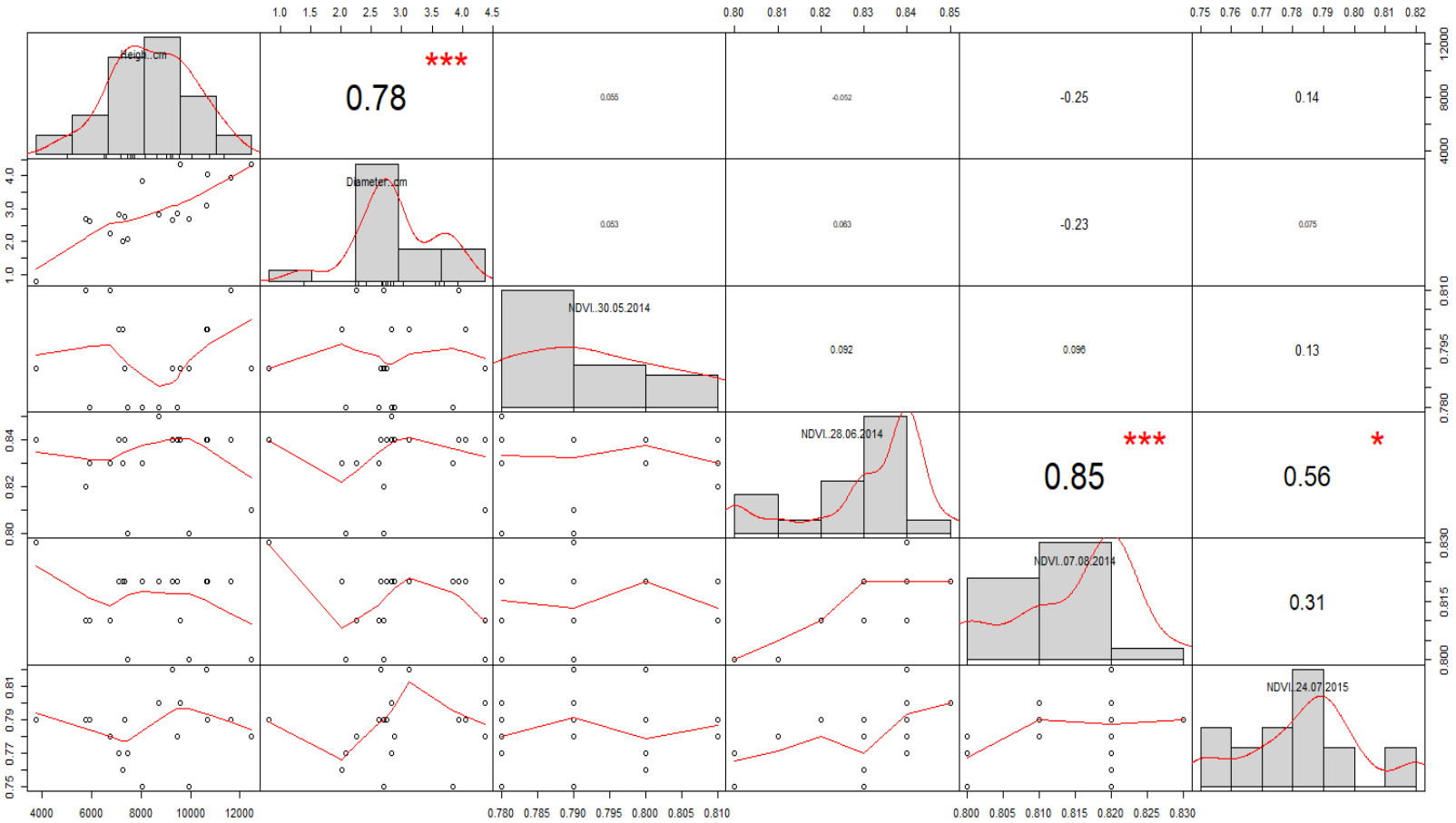

Figure 7: Checking for multicollinearity of variables in model 67

Appendix 2

Model numbering is presented in the Chapter 4: tables 4.4, 4.5 and 4.6.

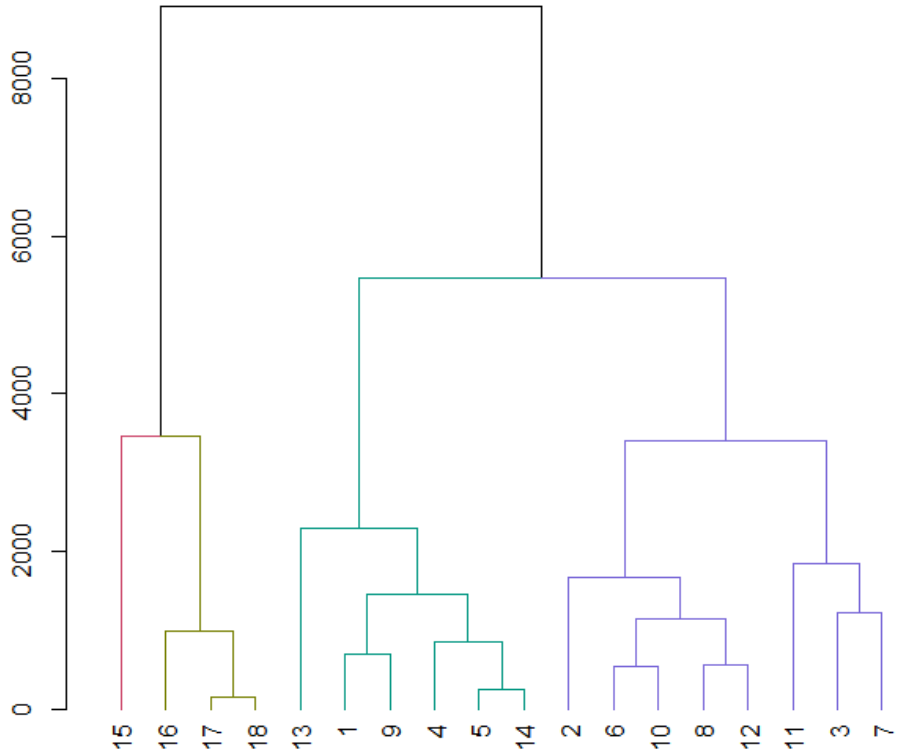

Figure 1: HC of the Plant Data dataset

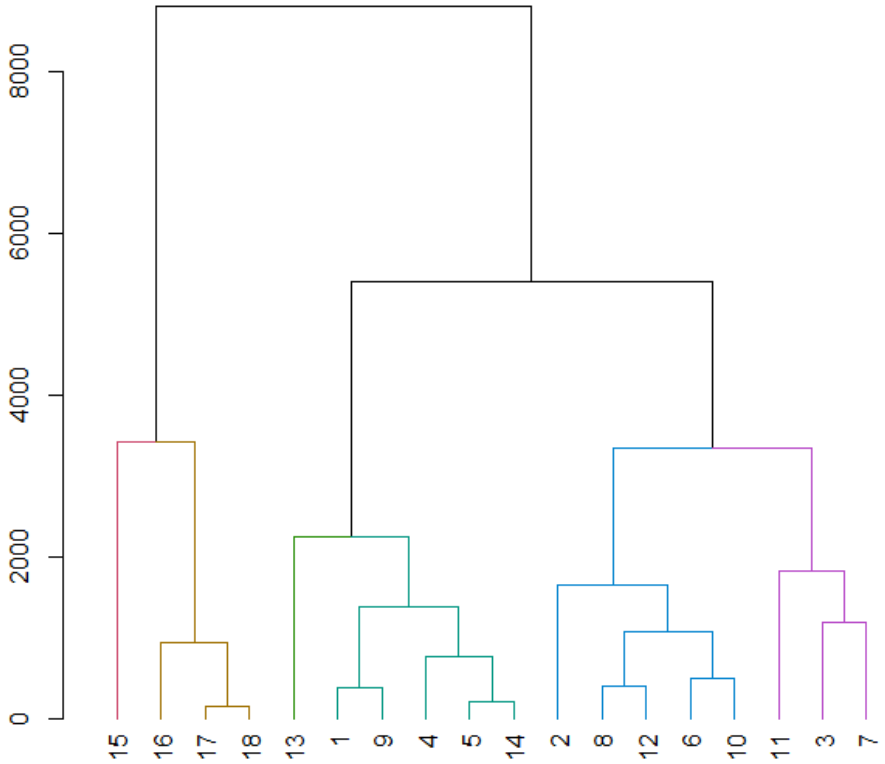

Figure 2: Model 26 (72) HC of the Plant Data dataset

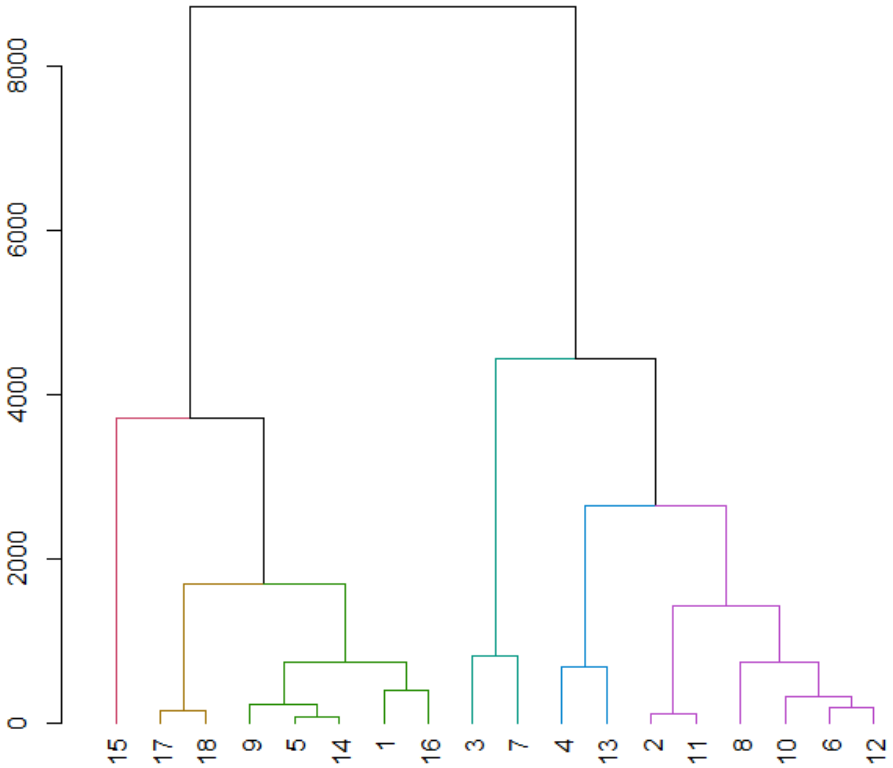

Figure 3: Model 37 (83) HC of the Plant Data dataset

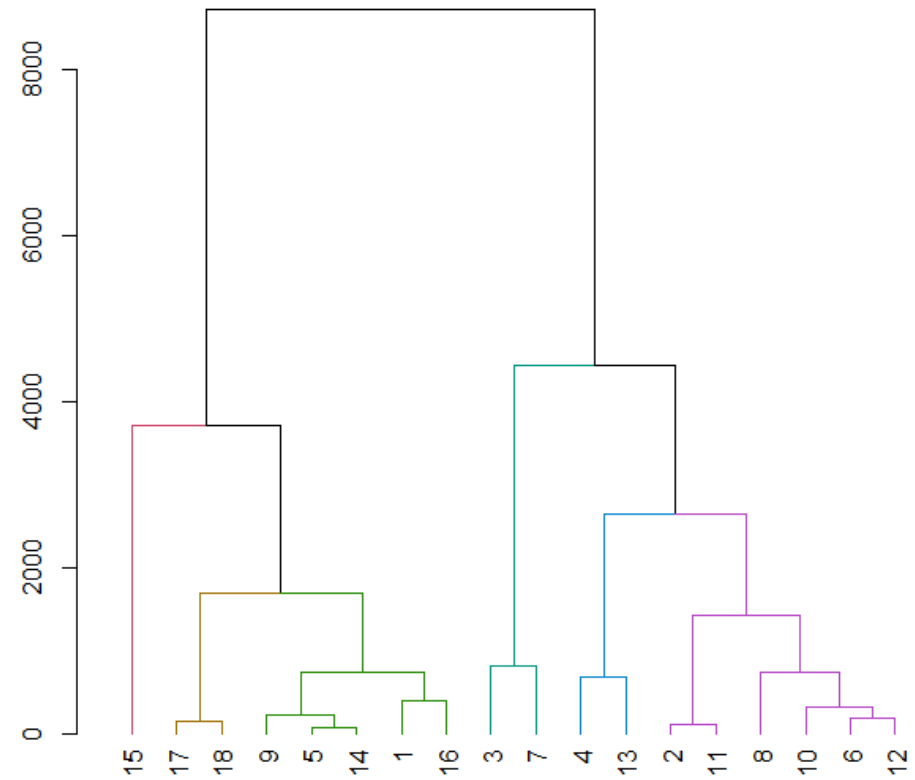

Figure 4: Model 39 HC of the Plant Data dataset

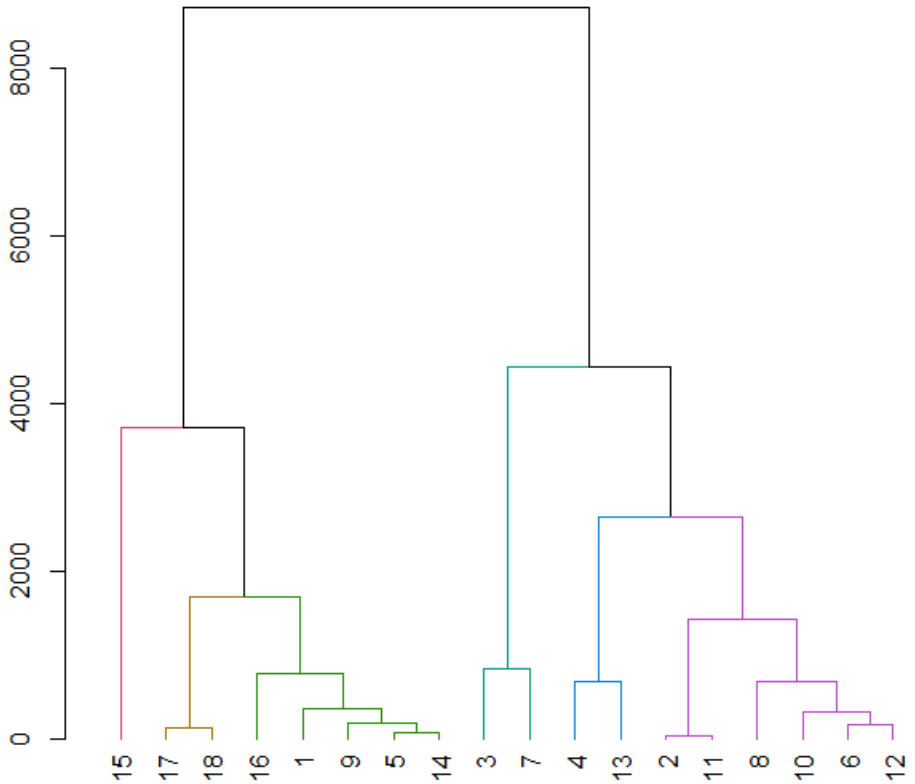

Figure 5: Model 48 (94) HC of the Plant Data dataset

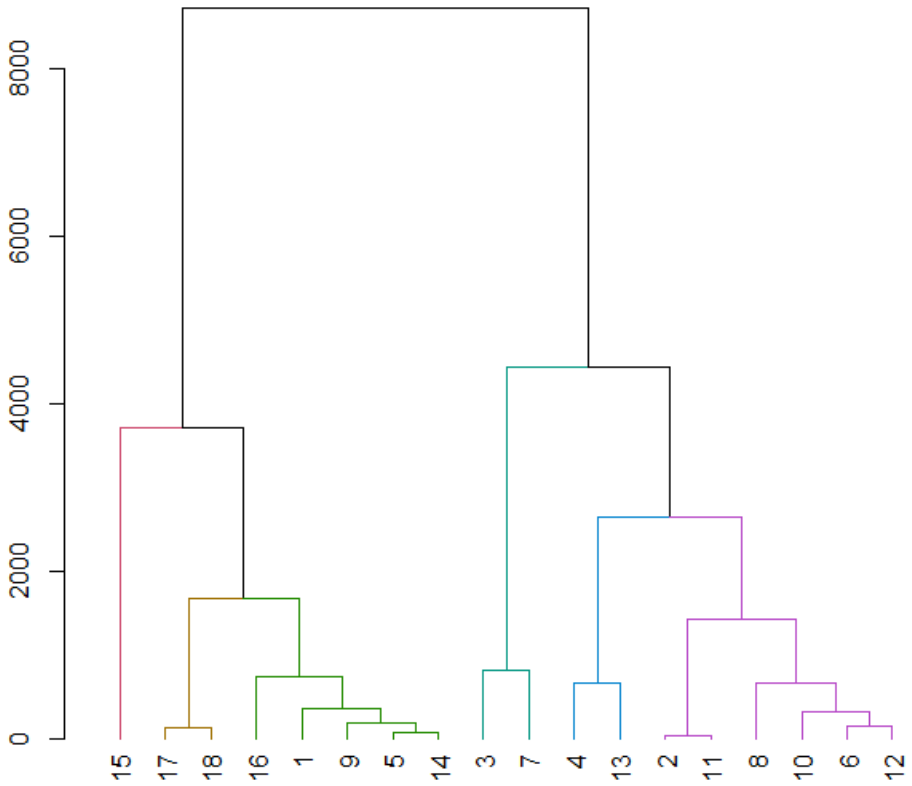

Figure 6: Model 60 (105) HC of the Plant Data dataset

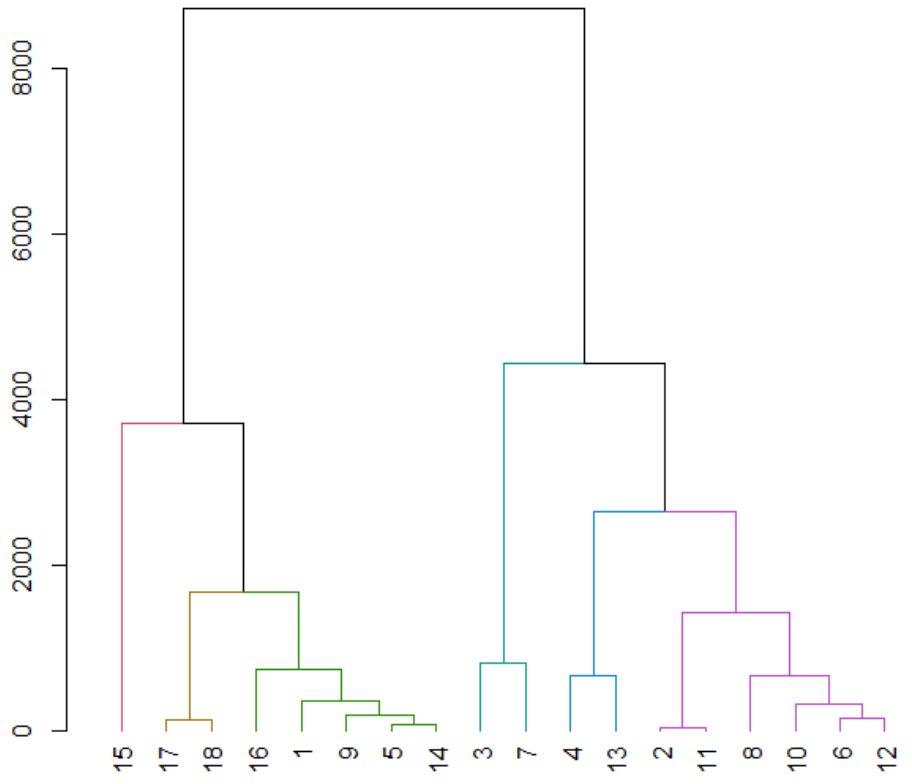

Figure 7: Model 67 (112) HC of the Plant Data dataset

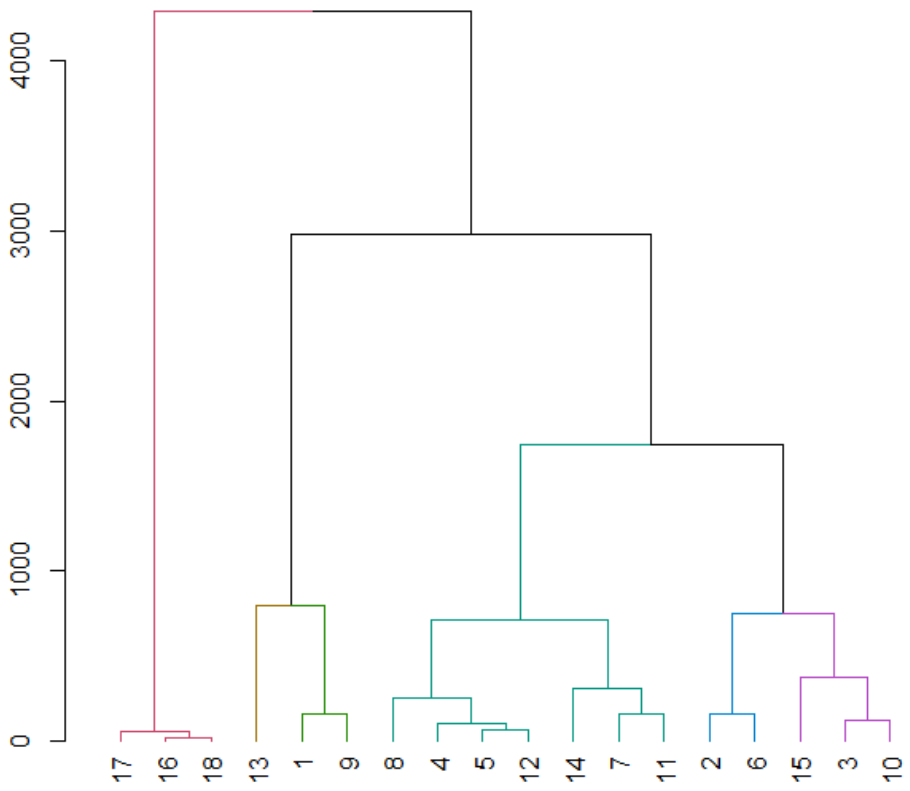

Figure 8: Model 76 HC of the Plant Data dataset

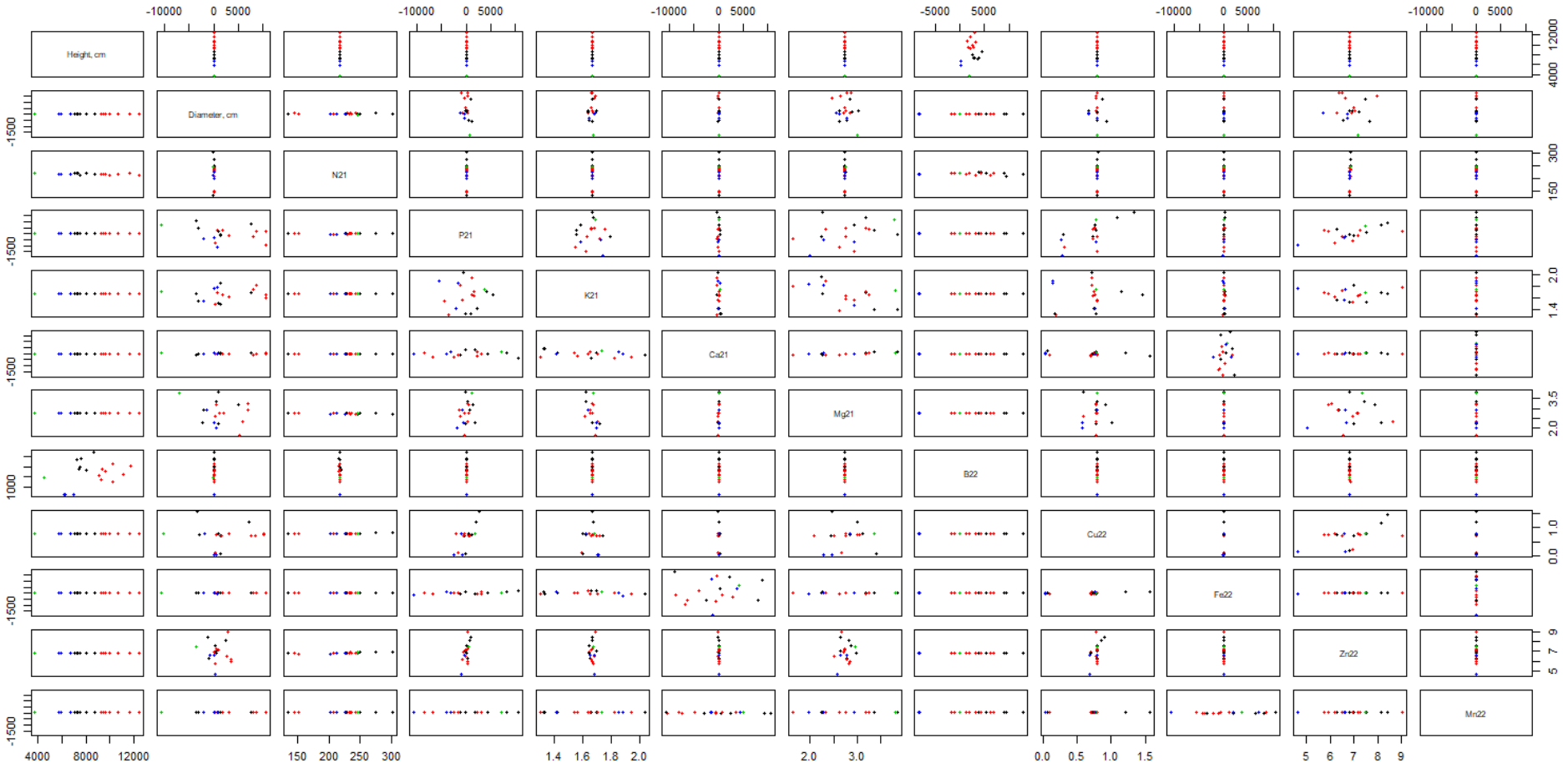

Figure 9: Distribution of distances between observations of model 26 (72) of dataset “Plant Data”

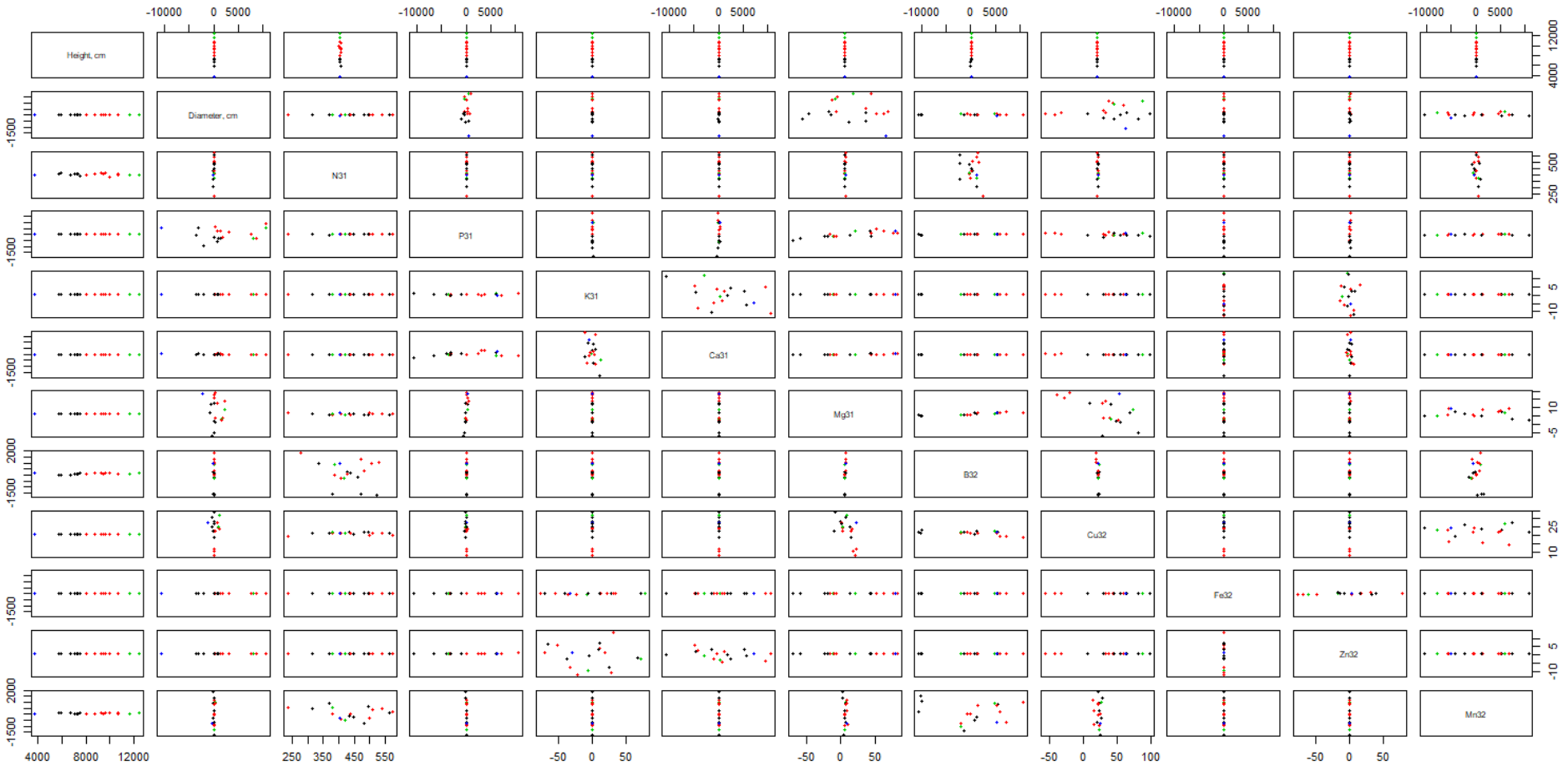

Figure 10: Distribution of distances between observations of model 37 (83) of dataset “Plant Data”

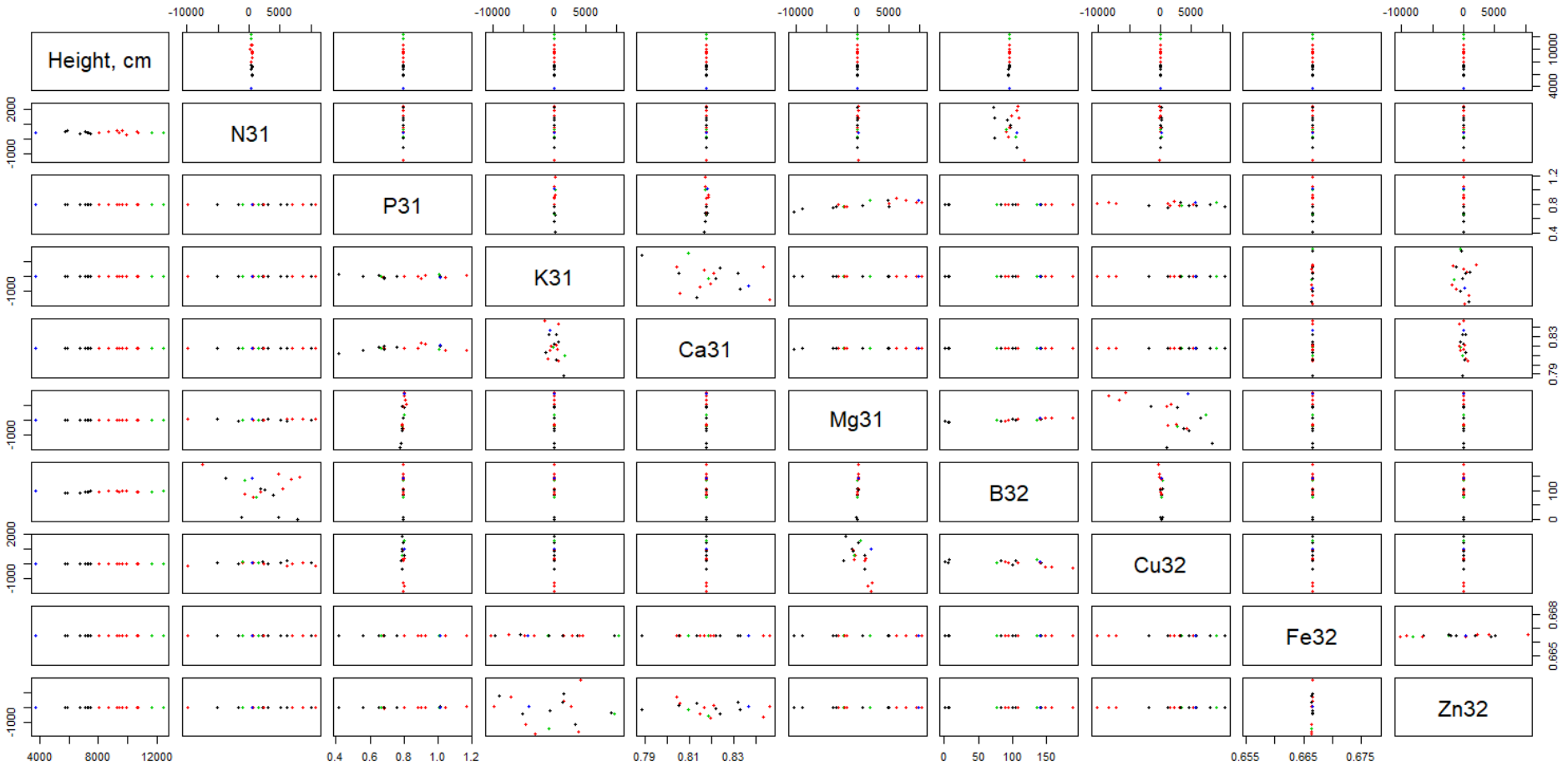

Figure 11: Distribution of distances between observations of model 39 of dataset “Plant Data”

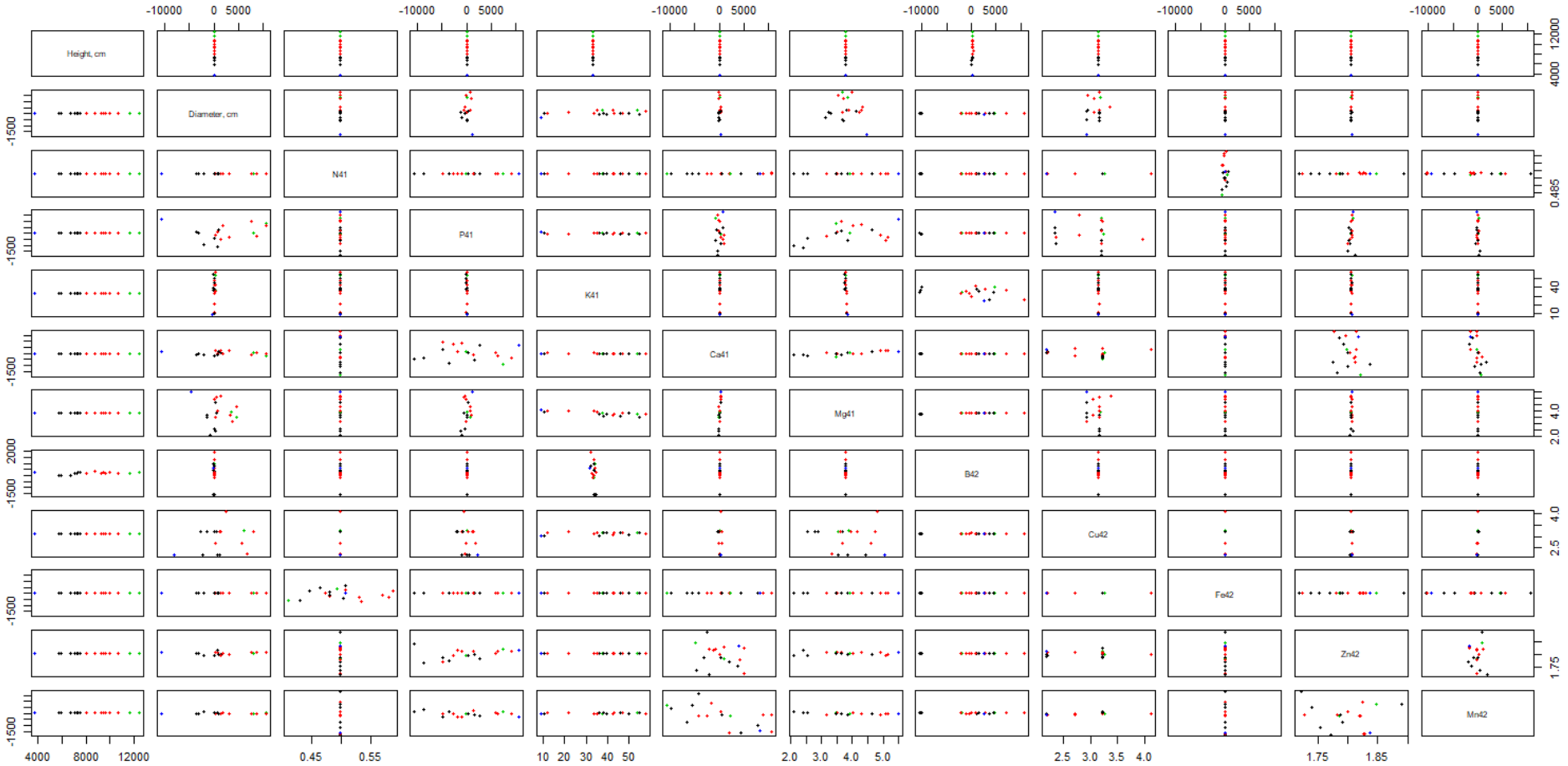

Figure 12: Distribution of distances between observations of model 48 (94) of dataset “Plant Data”

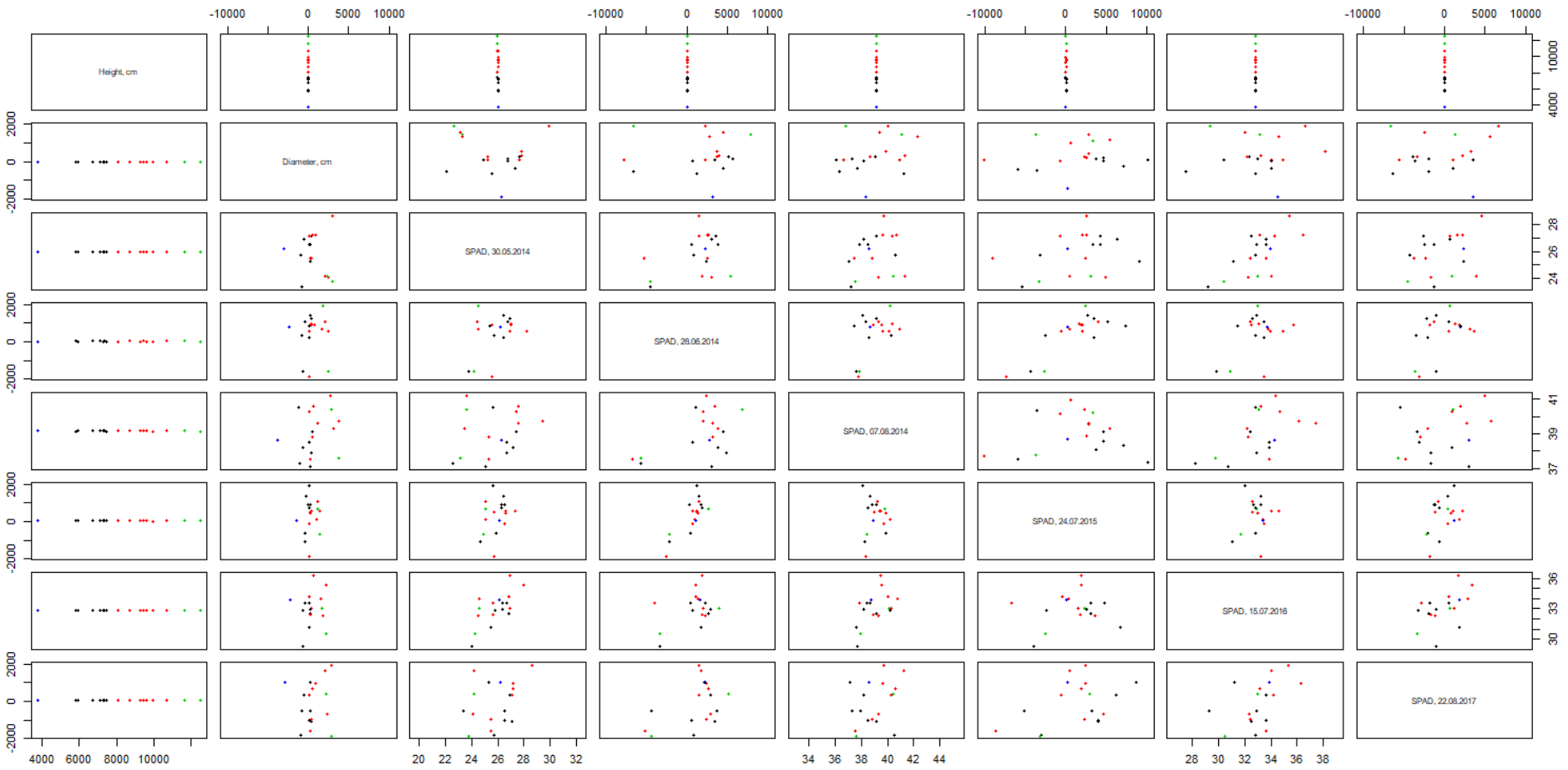

Figure 13: Distribution of distances between observations of model 60 (105) of dataset “Plant Data”

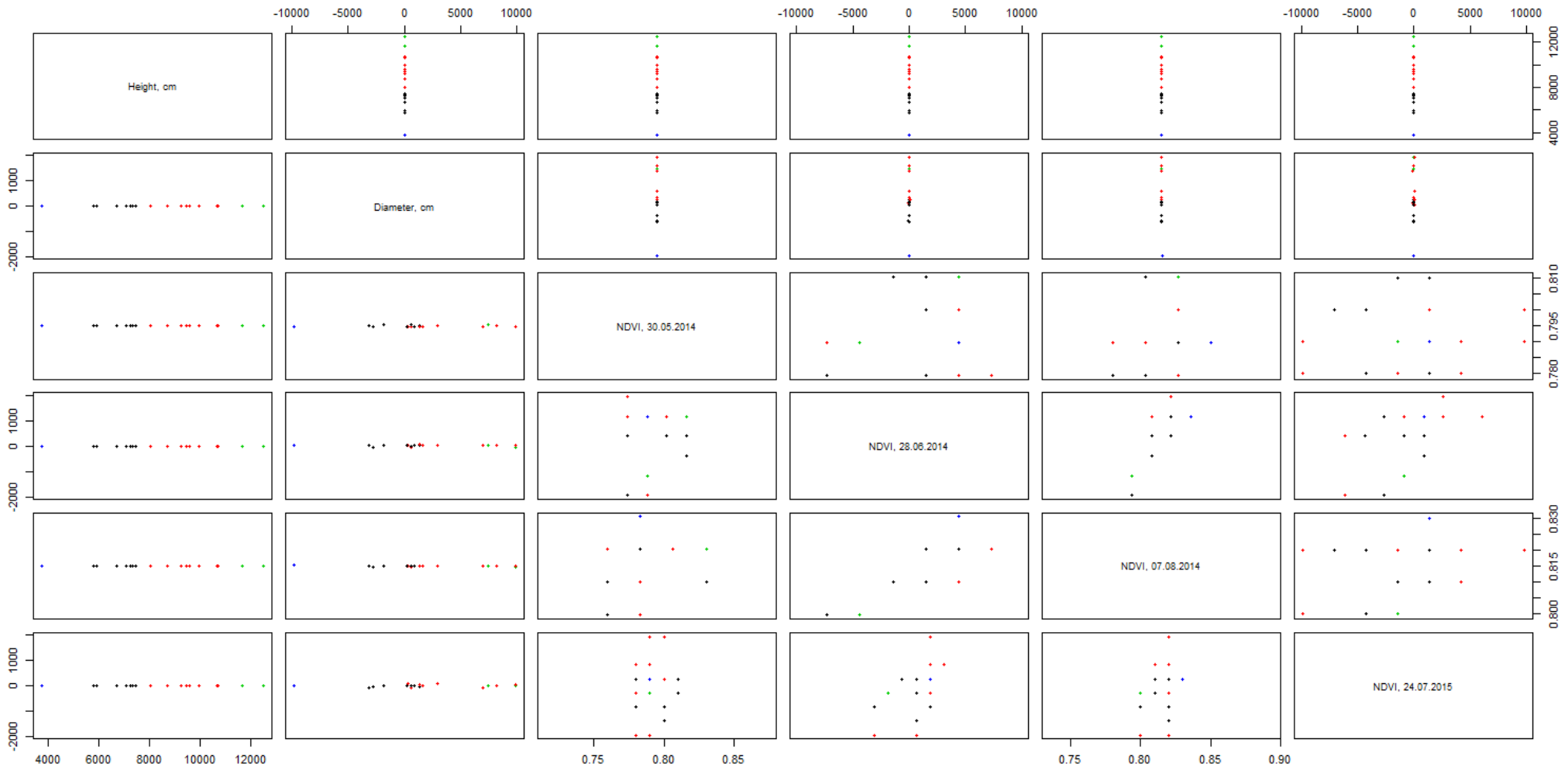

Figure 14: Distribution of distances between observations of model 67 (112) of dataset “Plant Data”

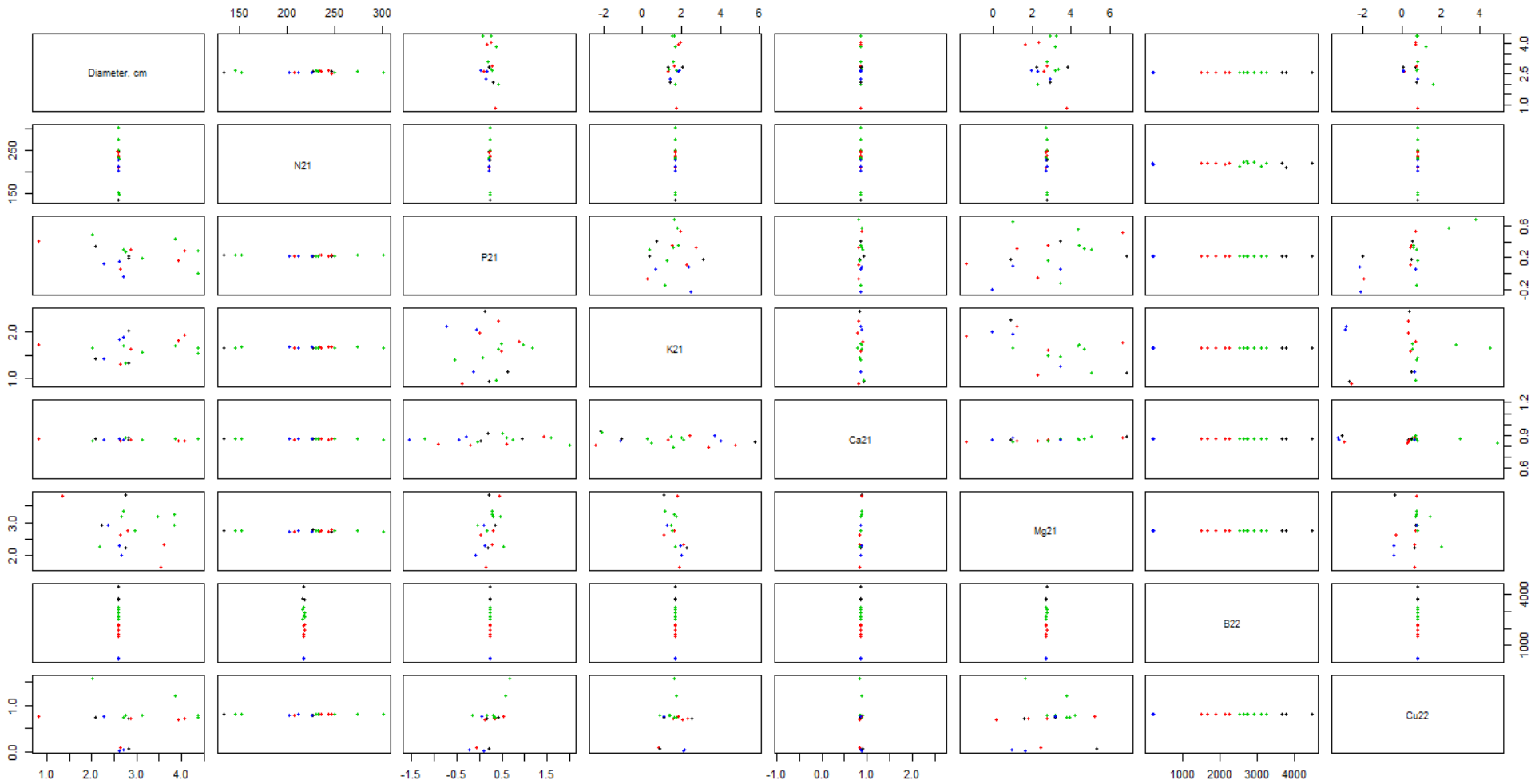

Figure 15: Distribution of distances between observations of model 76 of dataset “Plant Data