UNIVERSIDAD DE OVIEDO

PROGRAMA DE DOCTORADO EN INGENIERÍA INFORMÁTICA

# Planning and Decision Support Systems in the Frame of Precision Agriculture

Ph.D. Thesis

Author: Rodolfo de Benito Arango

Supervisors
PhD Irene Díaz Rodríguez
PhD Antonio Manuel Campos López

June, 2016

# RESUMEN DEL CONTENIDO DE TESIS DOCTORAL

| 1.- Título de la Tesis | |
|---|---|
| Español/Otro Idioma:<br>**Sistemas de Planificación y Soporte a la Decisión en el Ámbito de la Agricultura de Precisión** | Inglés:<br>**Planning and Decision Support Systems in the Frame of Precision Farming** |

| 2.- Autor | |
|---|---|
| Nombre:<br>**Rodolfo de Benito Arango** | |
| Programa de Doctorado:<br>**Ingeniería Informática** | |
| Órgano responsable:<br>**COMISIÓN ACADÉMICA DEL PROGRAMA DE DOCTORADO DE INGENIERÍA** | |

**RESUMEN (en español)**

La conservación del medioambiente es una prioridad para las autoridades Europeas. De hecho, varios de los programas principales de investigación de la Unión Europea (UE) están orientados al estudio y monitorización de la Tierra desde diferentes ámbitos. Estas iniciativas proporcionan servicios relacionados con políticas de protección ambiental, ingeniería forestal, cambio climático, desarrollo sostenible o agricultura. En este marco, el sector agrícola tiene una importancia estratégica para la economía global.

Uno de los servicios más relevantes en el dominio agrícola y medioambiental es la monitorización terrestre. Principalmente, estos servicios proporcionan información geográfica sobre la cobertura terrestre y el estado de la vegetación, con cobertura global y local. Esta información es crítica para los procesos de planificación y la toma de decisiones. De hecho, hay varios proyectos Europeos intentando proporcionar soluciones en este contexto. Sin embargo, los datos deben ser tratados y analizados adecuadamente. El procesamiento de estos datos puede resultar complejo debido a que se actualizan rápidamente, creciendo de una forma que su monitorización, análisis y almacenamiento se convierte en un reto. Además, la utilización de estos datos con procesos de toma de decisiones y aprendizaje automático puede entenderse como un problema de análisis de *big data*.

La Inteligencia Artificial (AI) puede verse como una herramienta para analizar este tipo de problemas y también para proporcionar aplicaciones y servicios de gran valor para los procesos de planificación y toma de decisiones relacionados con el dominio agrícola y medioambiental. Tales decisiones pueden llevar a una agricultura sostenible en un contexto *eco-friendly*. De hecho, los servicios de Agricultura de Precisión pueden ser mejorados y potenciados por medio del uso de algoritmos de aprendizaje automático y otras técnicas de IA. Por tanto, el estudio y desarrollo de modelos robustos para proporcionar servicios *smart-agro* alimentados con fuentes heterogéneas de datos representa un reto hoy en día.

Con esta finalidad, la presente tesis está orientada al estudio y desarrollo de modelos para servicios smart-agro relacionados con la delimitación automática del terreno y la identificación de zonas homogéneas de terreno cultivable, por medio del análisis de imágenes satélite combinado con algoritmos de aprendizaje automático. Esta tesis también propone un modelo para la predicción de la producción de uva, combinando fuentes heterogéneas de datos tales como imágenes satélite, registros históricos de producción y análisis de suelo, utilizando algoritmos de aprendizaje automático. Además, estos servicios *smart-agro* podrían ser integrados en plataformas agrícolas especializadas tales como *Farm-Oriented Open Data in Europe* (FOODIE) (http://www.foodie-project.eu/).

**RESUMEN (en Inglés)**

Environmental conservation is a priority for European authorities. In fact, several main research European Union (EU) programs are focused on the study and monitoring of the Earth from different scopes. These initiatives provide services related with environmental protection policies, forest engineering, climate change, sustainable development or agriculture. In this framework, the agriculture sector has a strategic importance for the global economy.

One of the most relevant services in the agriculture and environmental domains is land monitoring. These services mainly provide geographical information about the land cover and the state of the vegetation, with both global and local coverage. This information is critical for planning and for decision-making processes. In fact, there are some European projects trying to provide solutions in this context. However, data must be properly treated and analyzed. Processing these data may become complex because the available datasets are quickly updated, growing in a way that their monitoring, analysis and storage become challenging. In addition, the learning and decision making processes using these data can be understood a big data analysis problem.

Artificial intelligence (AI) can be seen as a tool for both analyzing this kind of problems and also for providing specific and high-value applications and services for planning and decision-making processes related to the agricultural and environmental domains. These decisions may lead to a sustainable agriculture in an eco-friendly context. In fact, Precision Agriculture services could be improved by means of machine learning algorithms and other AI techniques. Hence, the study and development of robust models for smart agro-services relaying on heterogeneous data sources using machine learning algorithms represent a challenge nowadays.

To that end, this thesis aims to study and develop models for smart agro-services related with the automatic cultivable land delimitation and the automatic identification of homogeneous land zones by means of the analysis of satellite imagery combined with machine learning algorithms. The thesis also proposes a model for forecasting grape production, combining heterogeneous data sources such as agro-meteorological stations, satellite imagery, historical records of yield production and soil analysis, using machine learning algorithms. In addition, these smart agro-services could be integrated on agricultural specialized platforms as the Farm-Oriented Open Data in Europe (FOODIE) (http://www.foodie-project.eu/).

**SR. DIRECTOR DE DEPARTAMENTO DE INFORMÁTICA/**
**SR. PRESIDENTE DE LA COMISIÓN ACADÉMICA DEL PROGRAMA DE DOCTORADO EN INFORMÁTICA**

UNIVERSIDAD DE OVIEDO

PROGRAMA DE DOCTORADO EN INGENIERÍA INFORMÁTICA

# Planning and Decision Support Systems in the Frame of Precision Agriculture

Ph.D. Thesis

Author: Rodolfo de Benito Arango

Supervisors
PhD Irene Díaz Rodríguez
PhD Antonio Manuel Campos López

June, 2016

# Acknowledgement

I would like to express my sincere gratitude to my supervisors PhD Irene Díaz Rodríguez and PhD Antonio M. Campos for their knowledge, work, support and expert guidance on this research. I would also like to thank PhD Elías Fernández-Combarro Álvarez for his support, patience, enthusiasms, knowledge and expertise.

My sincere thanks also goes to Emilio Rodríguez Canas, technical director of Bodegas Terras Gauda, for his patience and detailed explanations regarding all the aspects of the vineyard.

I thank my partners on the EU project Farm-Oriented Open Data in Europe (FOODIE) because I've learnt so much from them.

Finally, I would like to thank my family. My parents, for encouraging me always to pursue my dreams and, especially to Ana, my companion in the adventure of life, and to Carlos, my son, for they support and love.

# Abstract

Environmental conservation is a priority for European authorities. In fact, several main research European Union (EU) programs are focused on the study and monitoring of the Earth from different scopes. These initiatives provide services related with environmental protection policies, forest engineering, climate change, sustainable development or agriculture. In this framework, the agriculture sector has an strategic importance for the global economy.

One of the most relevant services in the agriculture and environmental domains is land monitoring. These services mainly provide geographical information about the land cover and the state of the vegetation, with both global and local coverage. This information is critical for planning and for decision-making processes. In fact, there are some European projects trying to provide solutions in this context. However, data must be properly treated and analyzed. Processing these data may become complex because the available datasets are quickly updated, growing in a way that their monitoring, analysis and storage become challenging. In addition, the learning and decision making processes using these data can be understood a big data analysis problem.

Artificial intelligence (AI) can be seen as a tool for both analyzing this kind of problems and also for providing specific and high-value applications and services for planning and decision-making processes related to the agricultural and environmental domains. These decisions may lead to a sustainable agriculture in an eco-friendly context. In fact, Precision Agriculture services could be improved by means of machine learning algorithms and other AI techniques. Hence, the study and development of robust models for smart agro-services relaying on heterogeneous data sources using machine learning algorithms represent a challenge nowadays.

To that end, this thesis aims to study and develop models for smart agro-services related with the automatic cultivable land delimitation and the automatic identification of homogeneous land zones by means of the analysis of satellite imagery combined with machine learning algorithms. The thesis also proposes a model for forecasting grape production, combining heterogeneous data sources such as agro-meteorological stations, satellite imagery, historical records of yield production and soil analysis, using machine learning algorithms. In addition, these smart agro-services could be integrated on agricultural specialized platforms as the Farm-Oriented Open Data in Europe (FOODIE) (http://www.foodie-project.eu/).

# Table of Contents

# List of Figures

# List of Tables

# Chapter 1

# INTRODUCTION

Agricultural production was intensified by the end of the 20th century to satisfy food demand. An increased application of fertilizers, massive inversions on irrigation systems and the reduction of maturity cycles are some of the factors responsible for the increase in production. However, the demand for food continues increasing and yield potential seems to be reaching a ceiling (Cassman, 1999). Therefore, in order to overcome that barrier, a Precision Agriculture (PA) approach is required (McBratney et al., 2005). Such approach may also benefit economical and environmental issues(Bongiovanni and Lowenberg-DeBoer, 2004).

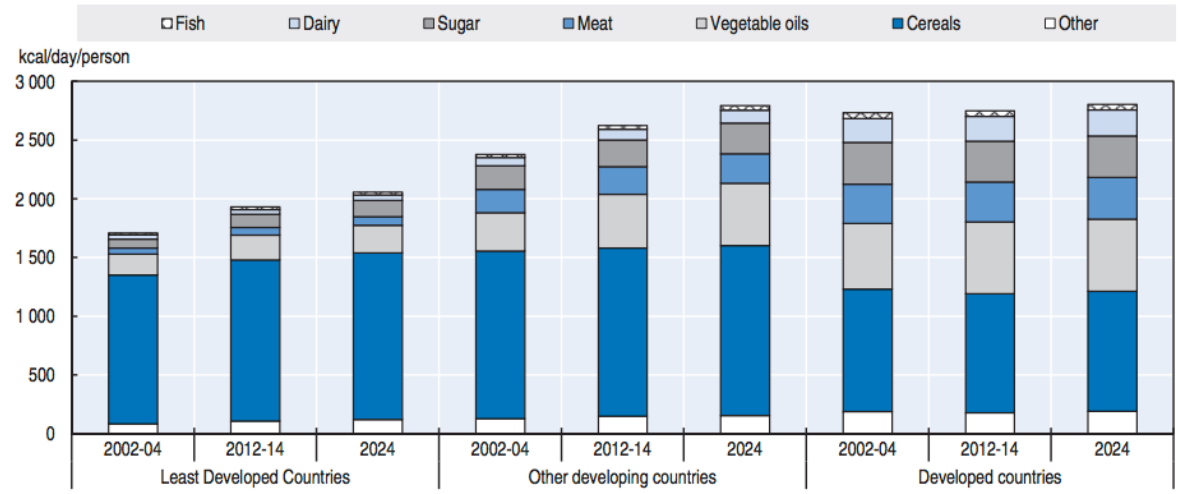
In this regard, Information and Communications Technology (ICT) and Artificial Intelligence (AI) techniques may contribute to develop and improve agricultural Decision Support Systems (DSS) which are necessary to the full adoption of PA (McBratney et al., 2005).

This chapter is an introduction to PA, describing the challenges of the integration in this domain of ICT and AI in order to model Smart Agro-Services for planning and DSS. The chapter is organized as follows. Section 1.1 introduces the concept of PA. Section 1.2 introduces AI and Smart Agro-services. Section 1.2 introduces the location and characteristics of the vineyards used in the case study. Finally, Section 1.4 explains the objectives of this thesis.

## 1.1 Precision agriculture

Environmental conservation is a priority for European authorities. In fact, several main research programs from the European Union (EU) are focused on the study and monitoring of the Earth from different scopes(http://www.2020horizon.es). The EU also has a hub for Earth monitoring called Copernicus (http://www.Copernicus.eu). It collects data from different sources such as satellite imagery and meteorological stations. These initiatives provide services related with environmental protection policies, forest engineering, climate change, sustainable development or agriculture.

In fact, the agriculture sector has an strategic importance for the global economy and it has also a crucial role on the human food supply (see Figure 1.1). For instance, cereals are




Note: The category "other" includes eggs, roots and tubers. Vegetables, fruits, pulses and other food items are not included in this figure.
Source: OECD/FAO (2015), "OECD-FAO Agricultural Outlook", *OECD Agriculture Statistics* (database), http://dx.doi.org/10.1787/agr-outl-data-en.
StatLink http://dx.doi.org/10.1787/888933228737

Figure 1.1: Caloric intake per capita in least developed, other developing and developed countries

responsible for providing about two-thirds of all food energy in human diets (Cassman, 1999).

However, 21st-century agriculture faces multiple changes (FAO et al., 2009) including those related to the growth of food and fibre production in order to satisfy the food demand of a rising population (see Figs. 1.2 and 1.3). It is also important to increase feed-stocks for future new bioenergy markets. Fulfilling this demand with a smaller rural labour force (FAO et al., 2009) without depleting the resources and protecting the environment requires to adopt more efficient production methods (Cassman, 1999). In fact, it will require precise management of all agricultural production elements in time and space dimensions (Cassman, 1999). The foundation of this kind of management is to provide the right inputs at the right place and time (Bongiovanni and Lowenberg-DeBoer, 2004) for maximizing yield production with agricultural practices that are economically viable and eco-friendly. This is the idea behind PA that will be explained on the following section.

### 1.1.1 The concept of Precision Agriculture

PA is related with Site-specific Management (SSM), a concept that it is defined in (Lowenberg-DeBoer and Swinton (1997)) as the "electronic monitoring and control applied to data collection, information processing and decision support for the temporal and spatial allocation of inputs for crop production". In this framework, PA provides a technological frame for automating SSM by means of ICT (Bongiovanni and Lowenberg-DeBoer, 2004). Thus, SSM could be a viable approach in commercial agriculture.

However, the definition of PA is elusive because it is evolving along with technological changes (McBratney et al., 2005). Initially, this concept was only associated to Variable-Rate Application (VRA) of fertilizers (Robert, 1993) which is intended provide the right inputs taking into account site-specific needs of crops.

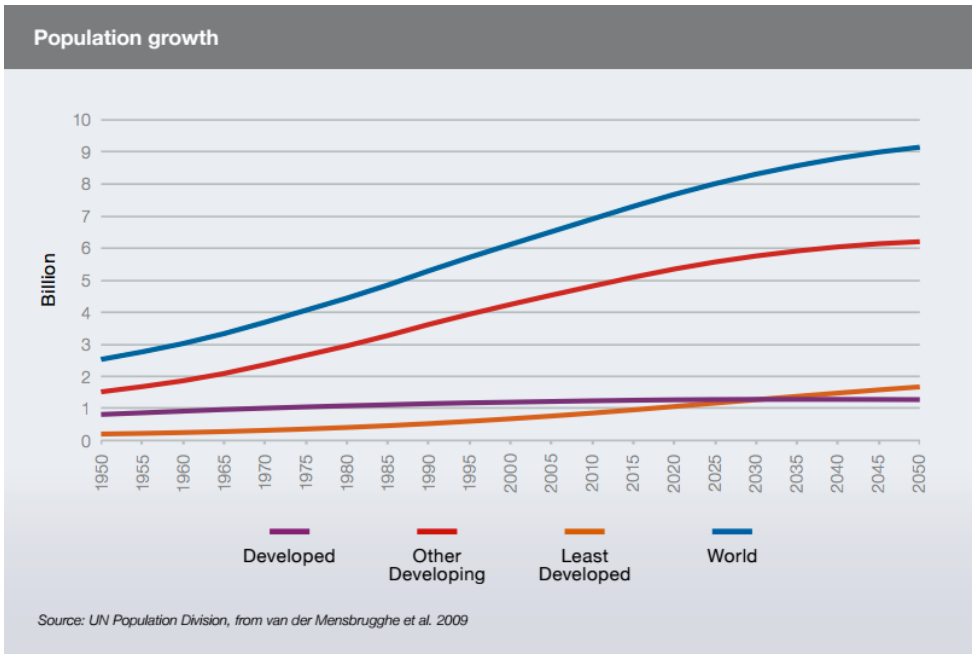



Figure 1.2: Population growth. Tendency from 1950 to 2050

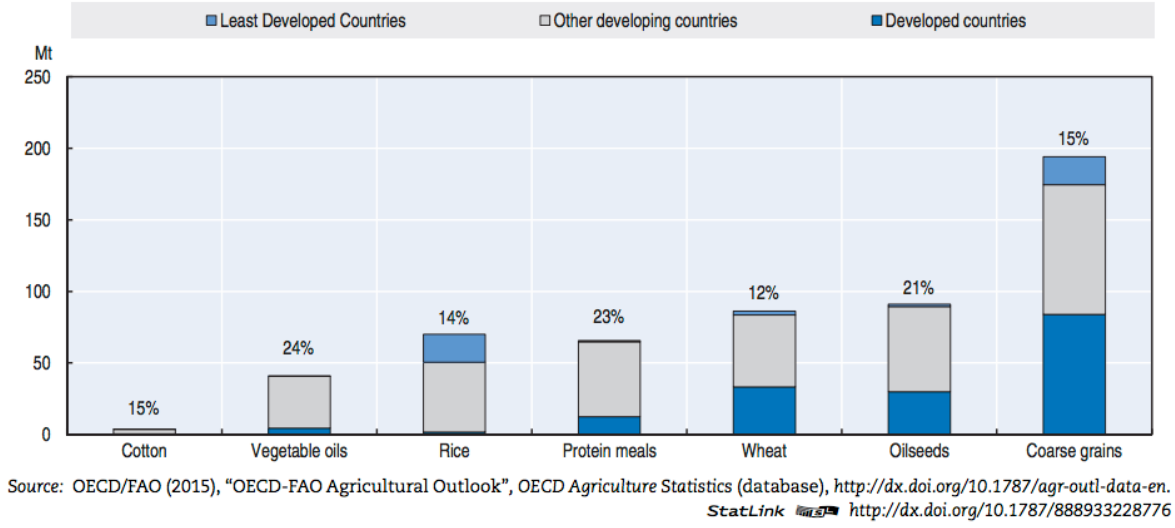



Figure 1.3: Projected growth of crop production in least developed, other developing and developed countries. Increase in volume and percentage, 2024 relative to 2012-14

In fact, the use of VRA techniques for nutrients and pesticides in a field provide an opportunity for effective management of inputs through PA, while increasing production efficiency (Hatfield, 2000). This idea of managing inputs, as it is required by the crops, is as old as agriculture (Bongiovanni and Lowenberg-DeBoer, 2004) and it is strongly related to the variations existing within a field. Such variations were categorized (Hatfield (2000)) as:

- **Natural** variations, such as soil composition.
- **Random** variations, for instance those caused by rainfall or weather events.
- **Managed** variations, such as agricultural interventions.

In order to address site-specific needs instead of applying uniform rates of inputs over large areas it is necessary to be aware of within-field spatial variability (Bongiovanni and Lowenberg-DeBoer, 2004). In fact, managed variations are well known by farmers whilst random variations could be difficult to foresee. On the other hand, the quantification of natural variations includes the knowledge of soil variation, biological variation and soil process variation (Hatfield, 2000). Spatial variation of the physical and chemical properties of the soil can be determined with soil analysis or predicted using pedotransfer functions such as those derived from the reflectivity of soil (Ludwig et al., 2008). Biological variations include, among others, populations of insects, weeds, soil microbial populations and disease outbreaks that could be inferred by the harvestable yield (Bongiovanni and Lowenberg-DeBoer, 2004). Finally, soil process variations as those produced by nitrogen dynamics are the most difficult to quantify because of the complex interactions between the physical environment and the biological response (Bongiovanni and Lowenberg-DeBoer, 2000).

However, PA is an holistic approach that not only manages the spatial dimension (see Fig1.4). The temporal dimension, often referred as Developmental Stage (DS) (Swinton, 1997), also plays a crucial role and comprises the information about the life cycles of agricultural crops or diseases. Hence, a broader definition of PA including both spatial and temporal dimensions could be: "that kind of agriculture that increases the number of (correct) decisions per unit area of land per unit time with associated net benefits" (McBratney et al., 2005). Those decisions and their associated benefits will be reviewed on the following section.

### 1.1.2 Precision agriculture for sustainability and environmental protection

The American Society of Agronomy (ASA) (1989) defines Sustainable Agriculture as "the one that, over the long term, enhances environmental quality and the resource base in which agriculture depends; provides for basic human food and fiber needs; is economically viable; and enhances the quality of life for farmers and the society as a whole".

In this regard, PA could have a positive impact on both the economical and environmental dimensions. These positive outputs can be categorized as follows (Bongiovanni and Lowenberg-DeBoer, 2000):

Figure 1.4: Spatial variability managed with the aid of WebGIS tools vs traditional maps

- **Nutrient management**. Regarding the amount of nitrogen fertilizer applied using Variable Rate Technology (VRT-N), the reduction could reach around the 36% (Griepentrog and Kyhn, 2000) maintaining the production. The environmental benefits of VRT-N are also remarkable (Thrikawala et al., 1999) and it can be concluded that VRT-N treatments produce larger decreases in residual soil nitrate than the uniform application of this fertilizer (Hergert et al., 1996; Delgado et al., 2001; Whitley et al., 2000).

- **Herbicides and Pesticides**. The use of VRT in the application of herbicides (see Figure 1.5) shows a decrease in environmental damage and reduces consumption about the 40–60% (Stafford and Miller, 1996; Timmermann et al., 2001) reaching savings of 66–75% according to some studies (Heisel et al., 1996). Regarding the use of pesticides for treatments related with diseases, it is stated in (Hatfield, 2000) than can be treated similarly to weeds using the same VRA principles.

- **Insecticides**. Site-specific pest management can contribute to the slow development of insecticide resistance and to conserve natural enemies (Midgarden et al., 1997). In addition, it can reduce insecticide inputs by 30–40% compared to a whole-field strategy.

- **Soil quality**. Soil quality can be defined as "the capacity of a soil to function in a productive and sustained manner, while maintaining or improving the resource base, environment, and plant, animal, and human health" (Larson and Pierce, 1991). In this regard, PA practices help to prevent soil erosion while environmental impacts are reduced and profits are increased (Meyer-Aurich et al., 2001).

- **Record keeping**. The registration of information regarding to applied inputs (as for

Figure 1.5: Use of VRT machinery for the application of herbicides and pesticides

instance the use of phytosanitary products) helps to implement environmental regulations based on observed practices (Swinton, 1997). For instance, Spanish farmers are required to keep a register with all kind of agricultural operations in a document called Cuaderno de Explotación Agrícola (Agricultural Holding Notebook). This notebook consists of several sections including one for detailing the phytosanitary treatments applied in the holding. Such information allows local and central administrations to perform spatial analysis tasks in order to generate maps of use of certain products or to study their impact on water quality.

### 1.1.3 Barrier for adopting precision agriculture

Considering the economical benefits and positive impact on the environment derived from PA (see Section 1.1.2) use, it is reasonable to think that farmers tend to embrace this set of technologies. However, there are some barriers that may hinder the use of PA. Wiebold et al. (1998) identifies the following ones:

- **Costs of technology adoption**, uch as the cost of the equipment or the time invested in learning how to use technology.
- **Training programs and consultation resources**. For instance, lack of local experts or training deficiencies.
- **Data quality control**. For example, difficulty in storing and retrieving data with different formats or methods to analyze yield data in order to better understand yield limiting factors.
- **Consumer guide for precision agriculture**. Mainly, reports with comparatives of PA equipment, techniques and software.

- **Environmental aspects of precision agriculture**. For instance, documented benefits of PA on the environment.
- **Need for new technology development**. Including the development of new types of sensors or methods for detecting and treating weeds and diseases.

### 1.1.4 Steps for adopting precision agriculture

Section 1.1.1 shows that PA takes advantage of ICT in order to provide valuable information and services to farmers. However, to take advantage of PA potential benefits, a certain degree of knowledge on ICT is required for farmers. In addition, it is necessary to overcome some barriers (see Section 1.1.3). Thus, in (Kitchen et al., 2002) it is proposed a natural learning process for PA based on the following six steps:

1. **Learning and understanding spatial data management**. The aim of this step is to understand within-field spatial variability and to use mapped information for analysis and decision making.
2. **Learning the proper use of sensors**. It includes at least the following:
   - Global Positioning System (GPS) to geo-localize relevant data for a field.
   - Yield monitoring systems for collecting data on-the-go across a field providing yield maps.
   - Remote sensing data to interpret images looking for patterns an to obtain crucial data about the health of crops (see Section 3.3.6) and other variables related.
   - VRT for the use of machinery equipped with VRA for its automatic application.
3. **Learning to use computers and software**. It is required to use of Geographic Information System (GIS) for storing the spatial data collected in previous steps and to manage computerized maps for their visual representation.
4. **Learning to identify relevant and manageable yield influencing factors**. In order to make crop production decisions it is feasible to analyze yield maps looking for spatial patterns identifying the factors that may influence on yield. Such patterns may suggest the partition of fields into Management Zones (MZ), regions with similar properties and, therefore, can be uniformly managed (Kholsa et al., 2001).
5. **Learning to develop site-specific management plan**. Once the information is analyzed and interpreted the following step is to determine achievable goals and to develop a site-specific management plan to achieve them.
6. **Learning to do strategic sampling and on-farm trials**. The final step is the optimization and refinement of the management process by means of monitoring the crop progress.

### 1.1.5 Potential adoption of precision agriculture

Regarding the potential worldwide-adoption of PA, the following topology of regions is identified in (McBratney et al., 2005). It is mainly based on economic criteria and government support for agriculture:

- **Type A. Developed economies with government-supported agriculture**. It includes the European Union, Japan and the USA. This type deals with environmental impact of agriculture due to an increase on inputs to maximize production. PA practices could help to put the stress on environmentally optimal production.
- **Type B. Developed economies with minimally government-supported agriculture**. It comprises Australia, New Zealand, Argentina and Brazil; where PA technology came later than to Europe or the US. This type focus on production quantity and quality and less in protecting the environment because of the economical dependence on agricultural exports.
- **Type C. Developing economies with plantation and/or centrally-planned agriculture**. It refers to most third-world countries such as Brazil, Mauritius, Malaysia and Costa Rica where already PA is being applied.
- **Type D. Developing economies with small-scale or subsistence agriculture**. In this countries PA has been scarcely adopted due to the dependence of technology inherent to PA.

On the other hand, focusing on the use of VRT, it can be considered three levels of adoption of PA practices from a technological perspective (Blackmore et al., 1995). The initial level is taken as a reference and states the traditional within-field practice with no ICT and no PA. Level two includes some investment on ICT that provides the farmer understanding for taking decisions about site-specific treatments. However, in this technological level the agricultural machinery is not equipped with VRT technology and the process must be done manually changing the setting of the equipments. Finally, level three represents fully adoption of VRT.

## 1.2 Artificial intelligence and smart agro-services

AI is a formal discipline which dates from the late 1950s and early 1960s and roots in other disciplines such as Philosophy, Mathematics, Psychology, Linguistics and Computer Science (Goldstein and Papert, 1977).

AI is defined in (Marr, 1977) as "the study of complex information processing problems that have their roots in some aspect of biological information processing". However, the definition of AI moves along different approaches of the AI related with behaviour, thought processes and reasoning (Russell et al., 1995). For instance, following the behavioural approach, AI is defined in (Luger and Stubblefield, 1990) as "The branch of Computer Science that is concerned with the automation of intelligent behaviour". Whilst Bellman in Bellman et al. (1978) defined

it as "[The automation of] activities that we associate with human thinking, activities such as decision-making, problem solving, learning ...". This latter definition follows the aforementioned reasoning approach applied to the branch of IA used in this thesis, Machine Learning (ML).

ML (see Section 2.1) is a branch of artificial intelligence that provides methods with the ability to learn from or to make predictions from data. Other branches of AI study, among others, natural language processing; computer vision, sensing and computer-aided translation; knowledge representation; robotics and expert systems (Bourbakis, 1992).

The following section will introduce the potential of ML in the agricultural domain and specifically in the PA context.

### 1.2.1 Modeling smart agro-services with machine learning algorithms

PA techniques supported by ICT provide knowledge about spatial variability and different characteristics of an specific area, helping to define more efficient and rational crop management plans in relation to more localized use of fertilizers and agrochemicals (Yu et al., 2010; Fernández-Quintanilla et al., 2011). In fact, PA practices require Decision Support Systems (DDS) to provide valuable information for the farmers by means of analysing relevant and manageable factors (Kitchen et al., 2002) and to facilitate the full adoption of PA (McBratney et al., 2005).

These PA services could be improved by means of machine learning algorithms and other AI techniques related with forecasting, classification and clustering that can provide a new kind of services that could be named as smart agro-services. The following services can be categorised as smart agro-services:

- **Automatic delimitation of crop land areas**, that can provide land classification by means of remote-sensed imagery analysis.
- **Automatic delimitation of MZs**, that is based on clustering homogeneous land regions based on spatial and temporal analysis of remote-sensed imagery and other relevant factors.
- **Crop yield forecasting**, based on regression algorithms and heterogeneous data sources such as climatic indices, biophysical features and remote-sensed imagery.
- **Recommendations about harvesting**, intended to predict the dates for harvesting beginning of each MZ and for selective harvesting by means of clustering the crops according to multi-spectral imagery.
- **Recommend site-specific phytosanitary treatments**, using specific diseases forecasting models for predicting outbreaks and progress of common crop disease such as Plasmopara Viticola or Uncinula Necator.
- **Recommend site-specific fertilization and irrigation needs**, using analysis of multi-spectral imagery for advising farmers regarding what MZ or specific regions of crops show water stress or lack of vigour.

Regarding the application of ML techniques in agricultural environments there are different approaches (see Section 2.2). For instance, Support Vector Machines (SVM, see Gualtieri and Cromp (1999)), k-Nearest Network Classifier (k-NN, see Zhu and Basir (2005)) or Random forest (Gislason et al., 2006) are algorithms often applied for land cover classification. Artificial Neural Networks (NN) were also used for this purpose in (Kavzoglu and Mather, 2003). However, when the variable to predict is continuous (as in yield prediction), the methods commonly used are M5 (Quinlan et al., 1992), k-NN (Altman, 1992) or Support Vector Regression (SVR, see Basak et al. (2007)). Bayesian techniques have been also used in PA for pest prediction (Tripathy et al., 2011) or for predicting coffee diseases (Perez-Ariza et al., 2012). ML techniques based on SVM are also extensively used for predicting pests (Wang and Ma, 2011). All these different approaches show the benefits of applying ML techniques in agricultural environment, pointing out that this area is promising.

## 1.3 Scenery for the cases of study

Bodegas Terras Gauda is a well-known Spanish wine producer under the Rías Baixas denomination of origin (DO) in Galicia, Spain. The region of Rías Baixas has an Atlantic climate with moderate temperatures and significant rainfall well distributed along the year, decreasing in the summer stage. However, it is also characterized by daily variations in weather. Such variations influence in the phenological development stages of the crops and their production (Lorenzo et al., 2013).

Figure 1.6: Terras Gauda vineyards

The scenery for the cases of study of this thesis is located in the Terras Gauda vineyards, in Rìas Baixas. Specifically, at O Rosal Valley, 41° 56'N 8° 47'W, (see Fig. 1.7 and 1.8). The extension of the vineyard considered in this thesis is about 60 hectares and includes varieties of Vitis vinifera such as Albariño, Caiño, Loureiro and Treixadura. In fact, Albariño, is one of the most important varieties of Vitis vinifera in Galicia (Loureiro et al., 1998) and produces high-quality wines.

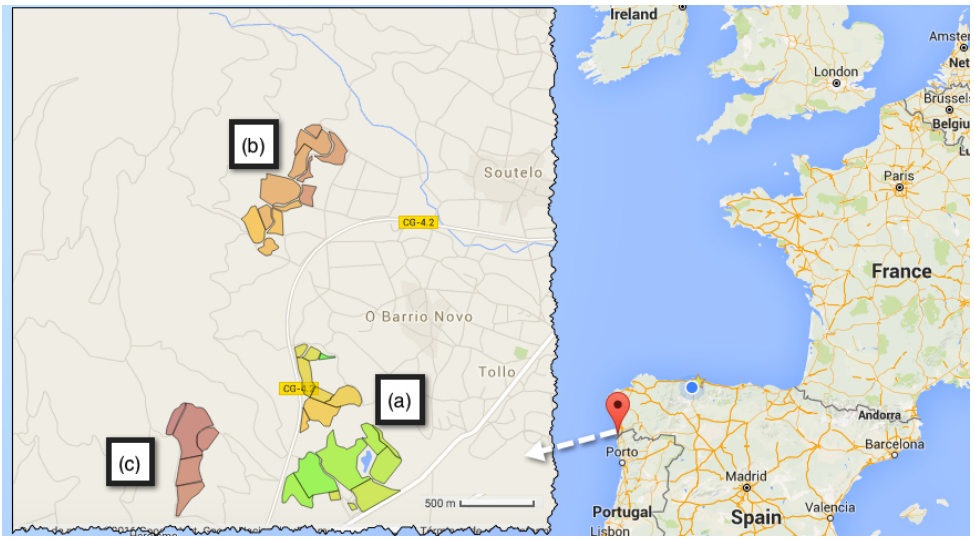



Figure 1.7: Location of the Terras Gauda vineyards. (a) Land parcel 1, (b) Land Parcel 2, (c) Land Parcel 3.

Figure 1.8: Detailed view of the Terras Gauda Parcels. Land parcel 1 (left), Land parcel 2 (middle), Land parcel 3 (right)

## 1.4 Thesis objectives

The aim of this research is to analyse and design models of smart agro-services for planning and decision support systems in the framework of Precision Agriculture. The goal is twofold: on one hand, a massive volume of heterogeneous data sources such as satellite imagery, agro-meteorological stations and data provided by public repositories like Copernicus is analysed. On the other hand, it will develop and adapt models for decision support in this kind of context.

In particular, the following objectives are identified:

- Evaluate potential data sources for smart agro-services, such as open-data satellite imagery or agro-meteorological networks (see Section 3).
- Analyze the effect of satellite temporal resolution in automatic delimitation of land (see

Section 4).

- Develop a method for automatic delimitation of cultivable land (see Section 5).
- Design a procedure for agricultural MZs identification(see Section 6).
- Establish models for predicting crop yield (see Section 7).

Sections 3 to 7 describe the proposals of this thesis. Section 2 describes the foundations of this work providing an overview of concepts, methods and algorithms from machine learning. Finally, Section 8 extracts conclusions and future work.

# Chapter 2

# MACHINE LEARNING IN THE PRECISION AGRICULTURE CONTEXT

Any decision-making system strongly depends on both the intelligent system used to produce decisions and the data to learn from. In particular, modelling smart agro-services requires the selection of the right ML algorithm for each kind of problem depending on the positive impact on the results and also on the computational cost (Domingos, 2012).

On the other hand, data is also a challenge in the smart-agro context because it requires collecting, processing and analyzing massive quantities of data (Lynch, 2008) from heterogeneous data sources. In addition, sometimes it is necessary to apply dimensional reduction methods in order to reduce the amount of data. In this chapter, the basic ML techniques as well as feature reduction techniques are described in order to a better understanding of the proposed smart agro-services.

Finally, Section 2.2 reviews related work in regard with the use of ML algorithms and techniques in the agricultural domain. In particular, those applications related with this thesis and with future works.

## 2.1 Introduction to Machine learning

AI follows two approaches to artificial learning (Hutchinson, 1994). The first one studies the learning process of the human mind and aims to translate these processes into algorithms and computer programs. The second approach is ML, which involves the development of algorithms that learn from data. In fact, ML is a branch of artificial intelligence that provides methods with the ability to learn from or to make predictions on data. These methods build a model from example inputs in order to make predictions (Mitchell et al., 1997).

ML does not make any assumption about the structure of the data model, allowing the

construction of complex non-linear models. There are many different paradigms in ML: lazy methods such as k-Nearest Neighbors (Altman, 1992), methods based on tree construction as, for instance, C4.5 (Quinlan, 1993) or Neural or Bayesian networks (Mitchell et al., 1997). All of them have been successfully used in many different domains.

Learning how to classify grapevine varieties, for example considering certain characteristics of their leaves, could be treated as a ML problem. The idea is to train an algorithm for distinguishing the varieties using examples already classified. The examples would have a set of input features with the values of morphological attributes of the leaf such as width, leaf area and so on; and a label identifying the corresponding variety of each input (see Fig.2.2). Once the algorithm is trained it predicts the grapevine variety from other input data with the same structure as the training set.

Other example regarding ML would be how to predict the grapevine yield based on features such as the area of the parcel, the grapevine variety and the total solar irradiance. In this case the training set would consist on data from past harvests for such features and their corresponding yield. Once the algorithm is trained it predicts the grapevine yield for the next campaign based on new input data (see Fig.2.5).

These examples follow a Supervised Machine Learning (SML) approach. In particular, the first one, predicting grapevine varieties, is a classification problem whilst the last one, forecasting yield, is a regression problem. Both classifiers and regression models are described in Section 2.1.1.

Following with this practical approach by means of examples, let's consider data collected in a vineyard days before the harvesting by a drone equipped with a multi-spectral camera. Assuming that the values of the reflectivity of the grapes are in some way correlated with their quality, it would be interesting to obtain different groups of grapes sharing similar reflectivity values. Once the groups of grapes are formed and represented in a map, it could be relatively easy to do selective harvesting without mixing the grapes from different qualities. This problem could be treated as a ML problem considering the reflectivity values as the training data. It should be noted that, in this case, the learner do not have any information about the expected quality of the grape based on the reflectivity values, as a SML would have. Hence, this problem correspond to the family of Unsupervised Machine Learning (UML) algorithms. These algorithms will be described on Section 2.1.3.

All these examples would arise some questions. For instance:

- **How to evaluate the predictions?** For performance evaluation of ML algorithms, Section 2.1.6 describes several performance measures (Jardine and van Rijsbergen, 1971)
- **Are the groups well formed?** Regarding the goodness of the groups obtained with UML algorithms, Section 2.1.4 describes some metrics such as the silhouette coefficient (Rousseeuw, 1987) and the Calinski-Harabasz index (Caliński and Harabasz, 1974).
- **In the case of a large number of features, which ones should be considered as input data?** This question is addressed by means of feature selection techniques described

in Section 2.1.5 such as Correlation-based Feature Selection (Hall and Smith, 1997).

### 2.1.1 Supervised learning algorithms

This family of algorithms is called supervised because the learning process requires the output (label) for each sample of the training data. In the case of the aforementioned examples, the output corresponds to the grapevine variety and the yield of the parcel, respectively. Formally, Supervised Learning Algorithms (SLA) uses as input a vector of discrete and/or continuous feature values $x_i = (x_{i,1}, ..., x_{i,d})$ and outputs $y_i$ (Domingos, 2012). In order to obtain that output, the algorithms require to be trained with a set of examples $(xt_i, yt_i)$ where $xt_i$ has the same dimensionality as the vector of input features $x_i$ and $yt_i$ is the label.

Considering whether the nature of the output to predict $y_i$ is discrete or continuous, the learning problem is approached by means of classifiers or regression models, respectively. It should be noticed that there are ML models and algorithms specifically designed to classify such as Naive Bayes (Nilsson, 1965) and other ones specialized on regression problems such as regression trees (Morgan and Sonquist, 1963). However, neural networks (Ripley and Hjort, 1995), for example, could be applied to both kind of problems.

#### Classifiers

Classification provides solutions such as deciding whether an email is considered SPAM or genuine (Guzella and Caminhas, 2009) or classify agricultural l (Duro et al., 2012) according to the aforementioned example of the grapevine leaves. SML algorithms classification task is related to decide the membership of an input vector of features $x_i$ to a known class $y_i$. In this regard, there are different approaches considering whether it is assigned a class to each $x_i$ or if it is assigned a probability of class membership $P(y_i|x_i)$. For instance, with the first approach a grapevine leaf would be classified as Treixadura. However, in the second approach the leaf would be classified as Treixadura with a probability $p$.

SVM (Cortes and Vapnik, 1995) are one the most representative algorithms of the first approach (Dreiseitl and Ohno-Machado, 2002). On the contrary, Naive Bayes (Duda et al., 1973) follows the probabilistic approach.

Classification is formally defined by the process of determining to which class an input vector of features belongs. This process requires the construction of a classifier which is a function that assigns a class label to the input vector (Friedman et al., 1997). The discrimination between two classes is known as a binary classification problem. However, if the classification involves more than two classes the classification problem is referred as multiclass. An example that falls into the first category would be determining if a given set of weather conditions leads to an outbreak of a particular crop disease. Whilst the classification of grapevine leaves into several varieties falls into the second type.

In this regard, specific algorithms are designed for providing solutions to binary classification. That is the case, for instance, of SVM and the Perceptron algorithm (Wasserman, 1989). Other algorithms, such as decision trees, are applicable to both problems. In fact, multiclass problems

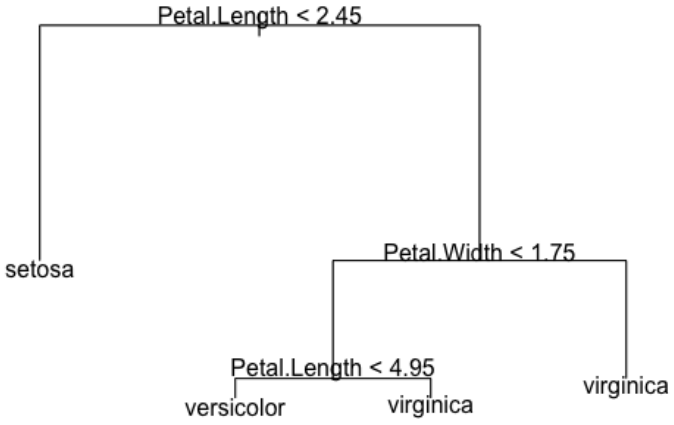



Figure 2.1: Decision tree for the dataset iris

can be addressed using binary classification techniques (Lorena et al., 2008). These approaches use decomposition strategies such as one-against-the-rest (Montanes et al., 2005), for dividing the problem into several binary subproblems and building a binary learner for each one.

On the other hand, the performance of classifiers partially depends on the selection of the training set (Zadrozny, 2004). For instance, unbalanced training sets where there is a predominant class are example of potential low performance of the algorithm.

In the following the SML methods related to the research proposed in this thesis are described (in particular for automatic land classification, see Section 5).

- **Decision trees**.

  C4.5 (Quinlan, 1994) is probably the decision tree classifier that has been most extensively used in the literature (Bae and Kim, 2011; Chou, 2012; Mistikoglu et al., 2015). The algorithm can be applied for classifying unordered discrete values.

  C4.5 constructs the tree from the training data following a greedy approach that splits the training set $T$ successively in each node until the stop condition is reached (see Alg. 2.1). Considering $n$ as the number of classes of $T$ and $a$ as the minimum of classes to stop the splitting process, if required. The tree-construction algorithm is applied recursively to each $T$.

  The first step is to compute the frequency of each class $C_i$ in $T$. If all instances in $T$ have the same class (or the number of classes in $T$ is less than a certain $a$ (Murthy, 1998)) then a leaf node is created with the associated class $C_i$ (or the most frequent class). In other case, the Information Gain (IG) is calculated (see Section 2.1.5) for each attribute $A_j$. Then a decision node is created selecting the attribute with the highest IG for test the node.

  The algorithm continues splitting $T$ into $k$ subsets where. Taking into account that discrete attributes consider as many partitions as possible values takes. Whilst continuous attributes consider two partitions based on whether the elements of $T$ have a value for

this attribute greater than a certain value or not, which has to be determined (Ruggieri, 2002). As a result of this process the algorithm considers each new partition as follows. A new leaf node is generated if the partition is empty, with the most frequent class of the parent mode. The algorithm is applied recursively to all the non-empty partitions.

**Algorithm 2.1** Pseudo-code of a BuildTree function for the construction of C4.5 algorithm (Ruggieri, 2002)

```
for all class C_i in 𝒯 do
  compute the frequency of the classes in T: freq(C_i, T)
end for

if (freq(C_i, T) = n) or (n <a) then
  return a leaf
end if

create a decision node

for all attribute A_j do
  compute InformationGain(A)
end for

nodeTest = attributeWithBestGain

if nodeTest is continuous then
  find threshold for splitting T
end if

for all T´ in the splitting of T do
  if T´ is empty then
    child of nodeTest is a leaf
  else
    child of nodeTest = BuildTree(T´ )
  end if
end for
```

The second stage involving the construction of a tree is pruning. It is intended to remove irrelevant and redundant nodes of the tree resulting in smaller trees (Zhang et al., 2002). In addition, it improves generalization and reduces overfitting (Chawla, 2003). Among the different pruning methods Bradford et al. (1998) identifies cost-complexity pruning (Breiman et al., 1984) and reduced error pruning and pessimistic pruning (Quinlan, 1987) as the most used.

Despite the advantages of C4.5 concerning simplicity, speed and error rate (Lim et al., 2000), the algorithm does not handle classes with continuous values (Lakshminarayan et al., 1996). On the other hand, may be affected by a selection bias regarding continuous variables with numerous distinct values (Quinlan, 1996b). This is related with the calculation of the threshold required for the splitting criteria.

An improved version is the C5.0 algorithm. It is faster than C4.5 and uses less memory than Quinlan's algorithm, in addition to producing trees that are usually smaller than those obtained with C4.5 (see Kuhn and Johnson (2013)), among other improvements.

The algorithm builds the classifier from the training data starting for seeking the feature that best divides the training data. This feature would be the root node of the tree. In this regard, there are several methods available and depending of the problem it has to be decided which method should be used (Murthy, 1998). The same process is recursively repeated for each partition creating sub-trees until no partitions can be done (Kotsiantis, 2007). The result (see Fig.2.1) is a tree where each node represents a feature and each branch its associated value. The leaves correspond to classes. Once the decision tree is build, the process of classify new instances is similar to follow a flowchart and involves beginning from the root following the branches according to the values of the features.

For example, Fig.2.1 shows a decision tree built from the Iris dataset (Fisher, 1936) for classifying iris plants (Setosa, Versicolour and Virginica) based on the length and width of both sepal and petal. For the classification of a new instance with this values: Petal.Length=4.7, Petal.Width=1.4, Sepal.Length=7.0, Sepal.Width=3.2; it is clear that in the root it should be follow the right branch because the value of Petal.Length for the plant to classify is greater than 2.5. In the next level it should be followed the left branch because Petal.Width, 1.4, is lower than 1.75, and so on until reach the bottom of the tree and the node with the class versicolor, which is in fact the right type of plant for this instance.

- **Naive Bayes**. It is a simple probabilistic classifier based on Bayes theorem (see Equ. 2.1) and is a widely used classifier, in many different domains (Lau et al., 2014; Kang and Kim, 2011; Koc et al., 2012).

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)} \tag{2.1}$$

The classifier assumes that the values of the features are statistically independent within each class (Farid et al., 2011) and given a new instance calculates the probability of membership to a class computing posterior probabilities using Equ.2.3.

$$P(y_i|(x_1, x_2, ..., x_n)) = \frac{p(y_i)\prod_{i=1}^{n} P(x_j|y_i)}{P(x_1, x_2, ..., x_n)} \tag{2.2}$$

where $P(x|y)$, $P(y)$ and $P(x)$ are calculated from the training data.

The result of the classifier is a probabilistic summary for each of the possible classes $y_i$. For instance, let's take the aforementioned grapevine leaf classifier assuming a Naive Bayes trained for the types of leaves (classes) $y = \{Loureiro,\ Treixadura\}$ and the discrete attributes: width category (w) and leaf area category (la). For classifying a new leaf (instance) with the attributes $x = \{w = 10,\ la = 3\}$, Naive Bayes will compute the following posterior probabilities:

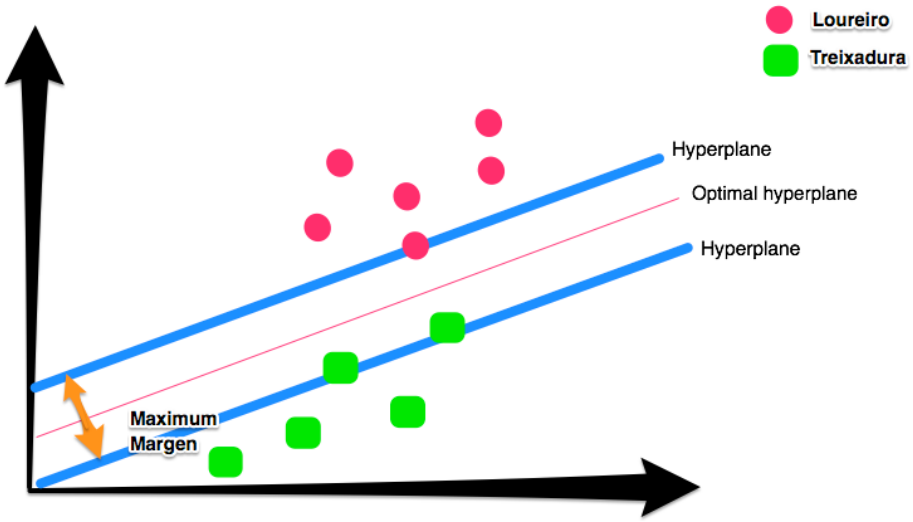



Figure 2.2: SVM and maximum margin

$$P(L|(w=10, la=3)) = \frac{p(L)P(la=3|L)P(w=10|L)}{P(w=10, lac=3)} \tag{2.3}$$

$$P(T|(w=10, la=3)) = \frac{p(T)P(la=3|T)P(w=10|T)}{P(w=10, la=3)} \tag{2.4}$$

where $L$ and $T$ are Loureiro and Treixadura, respectively.

- **Support Vector Machines**: SVM (Cortes and Vapnik, 1995) is one the most popular ML algorithm due to their efficiency and effectiveness in many problems (Díaz et al., 2004). SVM deals with the notion of margin or either side of a hyperplane that separates two classes (Kotsiantis, 2007). The idea is to maximize the margin between the classes in order to reduce an upper bound on the expected generalisation error (Kotsiantis, 2007) (see Fig. 2.2).

  The basic form of SVM learns the parameters $a$ and $b$ of a linear decision rule (Zadrozny, 2004) (see Equ. 2.5) and the sign, positive or negative, will determine the class membership of a new instance to classify.

$$h(x) = sign(a \cdot x + b) \tag{2.5}$$

  The aforementioned margin maximization between the classes is accomplished solving the following optimization problem:

$$\begin{aligned} &\text{minimize: } V(a,b) = 1/2a \cdot a \\ &\text{subject to: } \forall i : y_i \left[a \cdot x_i + b\right] \geq 1 \end{aligned}$$

  However, in practice, it could be difficult to obtain a decision rule that classifies correctly

all the inputs. Thus, the aforementioned optimization is modified with the inclusion of slack variables $\varepsilon_i$ (Cortes and Vapnik, 1995):

$$\begin{aligned} &\text{minimize: } V(a,b,\varepsilon) = 1/2a \cdot a + C\sum_{i=1}^{n}\varepsilon_i \\ &\text{subject to: } \forall i : y_i\left[a \cdot x_i + b\right] \geq 1 - \varepsilon_i,\ \varepsilon_i > 0 \end{aligned}$$

On the other hand, there are classification problems that require non-linear decision boundaries. To that end, it was introduced the use of kernel functions (Aizerman et al., 1964) defined by (Kotsiantis, 2007) as "special class of function that allow inner products to be calculated directly in feature space". Literature proposes a large number of kernels, some popular ones are the following (Kotsiantis, 2007):

- **Polynomial kernels** (Min and Lee, 2005):

$$K(x,y) = (x \cdot y + 1)^P, \tag{2.6}$$

  where $P$ is the degree of the polynomial.

- **Radial Basis Function (RBF)** (Lin and Liu, 2007):

$$K(x,y) = e^{\frac{-\|x-y\|^2}{2\varsigma^2}}, \tag{2.7}$$

  where $\varsigma$ is the width of the radial basis function.

- **k-NN**. It is an example of instance based learning method (Altman, 1992) as it does not produce a model in training time. It is characterized for being one of the simplest ML methods and for its reduced training time. This machine learning approach is widely used in many areas, such as for example text classification (Jiang et al., 2012).

  The idea is to classify new instances by observing the class of their neighbours. In fact, k-NN finds the k most nearest instances of the training set to the new one. Then assigns the most frequent class of those $k$ neighbours to the new instance. This closeness between instances is calculated by means of a distance metric. Some of the most significant (Kotsiantis, 2007) are shown on Table 2.1.

| | |
|---|---|
| Camberra | $D(x,y) = \sum_{i=1}^{m} \frac{\lvert x_i - y_i \rvert}{\lvert x_i + y_i \rvert}$ |
| Chebychev | $D(x,y) = \max_{i=1}^{m} \lvert x_i - y_i \rvert$ |
| Euclidean | $D(x,y) = \left( \sum_{i=1}^{m} \lvert x_i - y_i \rvert^2 \right)$ |
| Manhattan | $D(x,y) = \sum_{i=1}^{m} \lvert x_i - y_i \rvert$ |
| Minkowsky | $D(x,y) = \left( \sum_{i=1}^{m} \lvert x_i - y_i \rvert^r \right)^{1/r}$ |

Table 2.1: Example of distance metrics

There are several approaches for selecting the class from the $k$ neighbours. For instance, using the voting strategy where the class is chosen by selecting the most voted (Chiang et al., 2012), that is, the majority class among the $k$ neighbours. Other approach is the distance-weighting strategy (Wilson and Martinez, 2000) which means that the class of a closer neighbour affects more than the class of a more distant one. Hence, the problem of this algorithm is the selection strategy of the optimum $k$ (Kotsiantis, 2007). In fact, such strategy is often experimentally selected.

- **Artificial neural networks**. Artificial NN (Ripley and Hjort, 1995) are inspired in biological neural networks and have been applied in a wide variety of classification and regression problems (Paliwal and Kumar, 2009; Dreiseitl and Ohno-Machado, 2002). NN are composed by nodes called perceptrons or neurons. Each node takes an input vector and compute a nonlinear summing function (activation function, Dayhoff and DeLeo (2001)) in order to obtain the output $y$ (see Fig. 2.3). Such mapping from the input vector to the desired output uses weight vectors which are recomputed in the training stage (see Equ. 2.8).

$$S_j = \sum_{i=0}^{n} w_j a_i, \tag{2.8}$$

where $w_j$ is the weight for unit $j$ and $a_i$ the activation value for unit $i$.

There are different activate functions. Karlik and Olgac (2011) includes Uni-polar sigmoid, Bi-polar sigmoid, Hyperbolic tangent, Conic Section, and RBF. The Uni-polar sigmoid function (Cybenko, 1989) is the most popular one (Dayhoff and DeLeo, 2001) and is defined as:

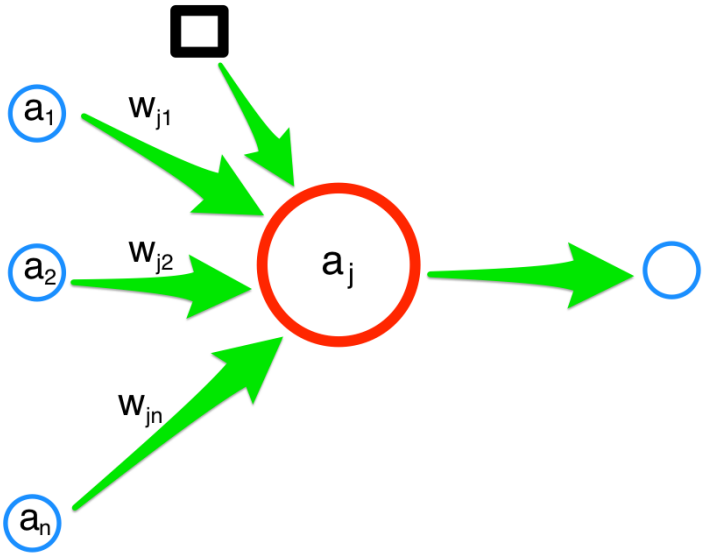



Figure 2.3: Artificial NN processing unit. $a_j$ is the activation value for unit $j$, and $W_{ji}$ is the weight from unit 1 to unit $j$. The square at top-left represent the bias

$$g(x) = \frac{1}{1 + e^{-x}} \tag{2.9}$$

Hence, the activation function $a_j$ would be:

$$a_j = \frac{1}{1 + e^{(-S_j)}}, \tag{2.10}$$

where $S_j$ is the incoming sum for unit $a_j$.

Although the use of the uni-polar sigmoid function is extended, the study of performance analysis of activate functions in (Karlik and Olgac, 2011) concludes that the Hyperbolic tangent performs better recognition accuracy than those of the other functions. The Hyperbolic tangent function is defined as:

$$tanh(x) = \frac{sinh(x)}{cosh(x)} = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{2.11}$$

where $S_j$ is the incoming sum for unit $a_j$.

Regarding the architecture of NN, the most popular is the multilayered perception (MPL) (Dayhoff and DeLeo, 2001) which groups neurons in different interconnected layers (see Fig. 2.4):

- **Input layer**. The neurons of this layer are fed with the input vector.
- **Hidden layers**. Each neuron is fed with the output of all the neurons in the input layer or in the previous hidden layer. MPL have at least one hidden layer and this number is parametrizable for each algorithm.

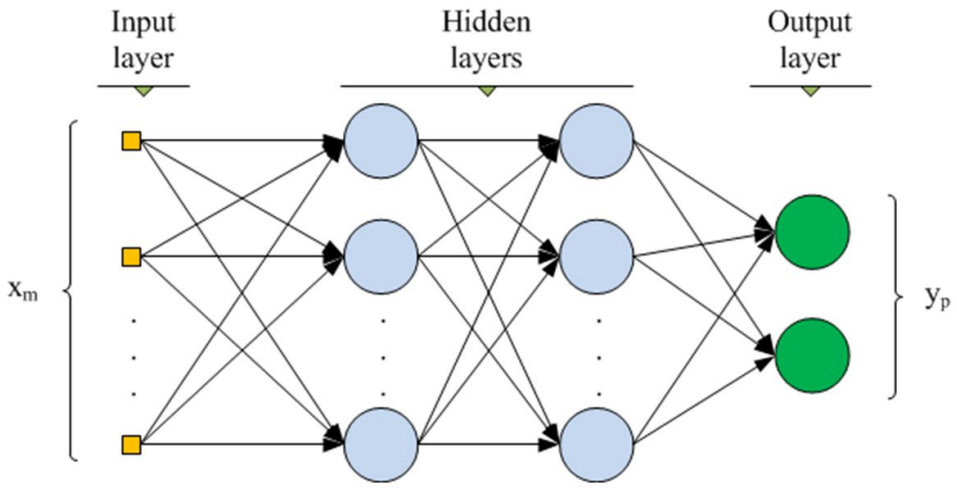



Figure 2.4: Example of NN with MPL architecture. Source: Thiago M. Geronimo, Carlos E. D. Cruz, Fernando de Souza Campos, Paulo R. Aguiar and Eduardo C. Bianchi (2013).

– **Output layer**. Each neuron is fed with the output of all the neurons in the last hidden layer. The output is the class $y$.

Other architectural approaches such as deep convolutional neural networks (Krizhevsky et al., 2012) consider thousands of layers. Recently, this kind of artificial NN seems to be gaining popularity due to notables advances in image recognition (Taigman et al., 2014) and speech recognition (Hinton et al., 2012), among other fields (Hadsell et al., 2009; Collobert et al., 2011).

Regarding improving generalization for avoiding over-fitting, there is a strong dependency on the balance between the training set and the complexity of the NN (Schwartz et al., 1990). In this regard, there are mechanisms for controlling such complexity. Simple Weight Decay is a parameter which controls the complexity of the model by means of limiting the growth of the weights of the NN (Moody et al., 1992). It penalizes large weights and choosing the right value of the parameter may improve generalization.

### Regression models

Regarding the learning problem related to predict the value of a class $y_i$ for a vector of inputs $x_i$ when $y_i$ is a continue value, the idea is to use a function representing the expected relationship between $x_i$ and $y_i$ (see, for example, Fig. 2.5). For instance, for one feature the relationship is expressed in Equ.2.12)

$$y_i = h(x_i) + e_i, \tag{2.12}$$

where $e_i$ is the error and $h(x_i)$ is the hypothesis:

$$h_\vartheta(x_i) = \vartheta_0 + \vartheta_1 x \tag{2.13}$$

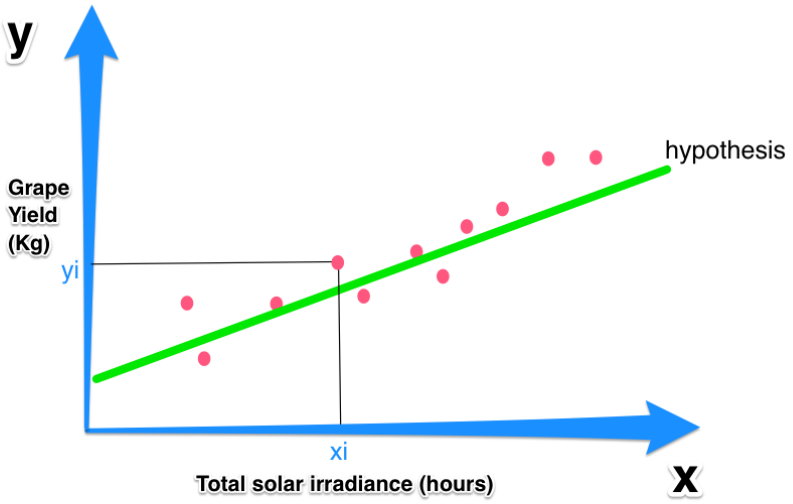



Figure 2.5: Example of linear regression

Considering the aforementioned example, at the beginning of this section, predicting the grapevine yield based on the total solar irradiance; the question would be how to find the "right" values of the parameters $\vartheta_0$ and $\vartheta_1$ in the following equation:

$$h_\vartheta(x_i) = \vartheta_0 + \vartheta_1 TotalSolarIrradiance \tag{2.14}$$

The intuition behind this is to use a cost function:

$$J(\vartheta_0, \vartheta_1) = \frac{1}{2m} \sum_{i=1}^{m} (h_\vartheta(x_i) - y_i)^2 \tag{2.15}$$

with the goal of to minimize that function with algorithms such as gradient descent.

On the other hand, regression trees learners produce a classification based on piecewise linear functions as they partition the space into a set of regions and fit the predicted value within each region using a linear model. A well-known regression tree algorithm is M5-Prime (Wang and Witten, 1996). It is a learner which constructs regression trees producing a classification based on piecewise linear functions as they partition the space into a set of regions and fit the predicted value within each region using a linear model.

The way this method works is the following: Assuming a training set with examples each one defined by its value on a set of attributes (discrete or continuous) and a continuous target, the method constructs a model that relates the target values of the training examples to the values of the variables defining the example.

This model can then be easily applied to predict the target variable: in the first phase, the decision tree (see, for example, Fig.2.6) is used to classify the example into one of the groups;

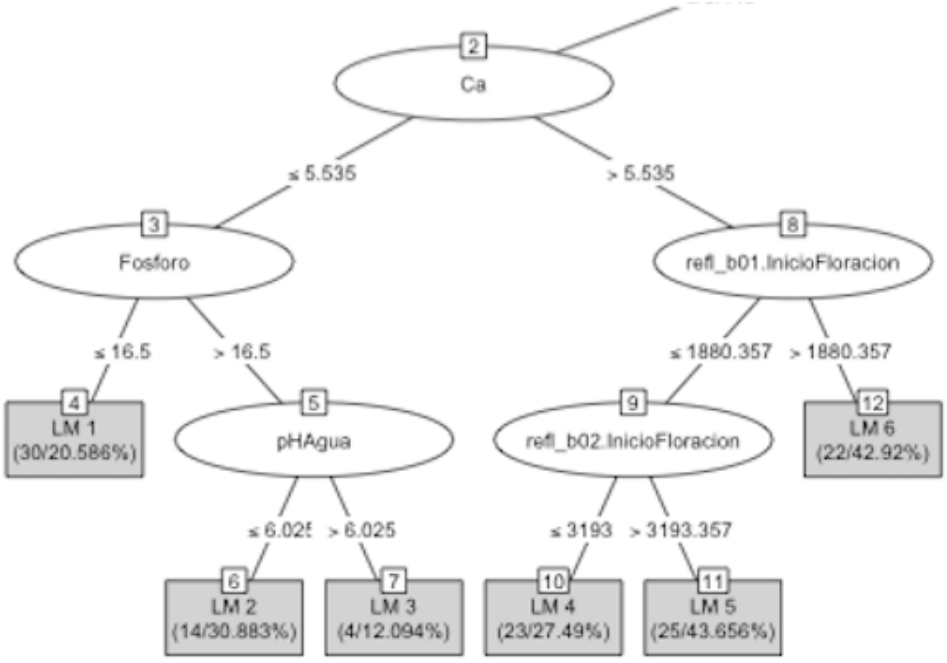



Figure 2.6: Example of regression tree (partial view) where for each leave it shown the corresponding linear model reference, the number of the instances from the training set and the percentage of error. For instance, the linear mode corresponding to leave number 4 is LM 1, it has 30 instances and the error is 20.586%

.

then, the linear equation associated to the particular group the example has been classified into (see Appendix B for examples of these equations).

M5-Prime selects the split that maximizes the expected error reduction. Once the tree is constructed, a multivariate linear model is computed for the examples at each tree node with standard regression techniques and using only attributes that are referenced by tests or linear models somewhere in the sub-tree at this node.

The main characteristics of this method are:

- Regression tree construction:
  - **Splitting criterion**: Maximize $SDR = sd(T) - \sum_i \frac{|T_i|}{|T|} \times sd(T_i)$

    being $T$ the set of examples that reaches the node and $T_1, T_2, \ldots$ the subsets resulting from the node split according to the selected attribute.
  - **Stopping criterion**: Standard deviation below a given threshold (small enough) in all nodes
  - **Pruning**: Heuristic estimation of absolute error of linear regression models according to

    $$\frac{n+v}{n-v}$$

    with $n$ being the number of examples that reach the node and the number of parameters that represents the class value at that node. Pruning greedily removes terms

from linear regression models to minimize the estimated error.

- **Smoothing** is used to compensate discontinuities between adjacent linear models at the leaves of the pruned tree. The smoothing process uses first the leaf model to compute the predicted value and then filters that value along the path back to the root, combining it with the value predicted by the linear model for that node. The modified prediction $p$ is computed by

$$p' = \frac{np + kq}{n + k}$$

with $n$ the number of examples at the smoothed node, $k$ a constant, $p$ and $q$ are respectively the predictions passed to the studied node from below and the value predicted by the model at the studied node. Basically, what this process does is to achieve the effect of incorporating ancestor models into the leaves.

- The value at each leaf is estimated using a linear regression function.
- At each node, it uses only a subset of attributes occurring in the sub-tree.

### 2.1.2 Applying supervised learning algorithms

As a general idea, for the process of applying SLA to a problem Kotsiantis (2007) proposes the following task (see Fig. 2.7):

1. **Identification of required data to the problem**. This task may include dimensionality reduction by means of feature selection techniques (see Section 2.1.5)
2. **Data pre-processing**, cleaning and standardizing the data.
3. **Definition of training set**, selection of random samples from the input data.
4. **Algorithm selection**, studying those SLA more applicable to the problem.
5. **Training**, executing the algorithm with the training set.
6. **Evaluation** with test set, using performance evaluation (see Section 2.1.6).
7. **Verify results**, if the results of the performance evaluation are not the expected then consider repeating previous steps or setting up the parameters of the algorithm.
8. **Apply the classifier** to the input data when the results are the expected.

### 2.1.3 Unsupervised learning algorithms

Unsupervised Machine Learning (UML) algorithms use as input a vector of discrete and/or continuous feature values $x_i = (x_{i,1}, ..., x_{i,d})$. In the case of clustering algorithms, the aim is to group the input data into $k$ partitions which are defined using a measure of similarity between

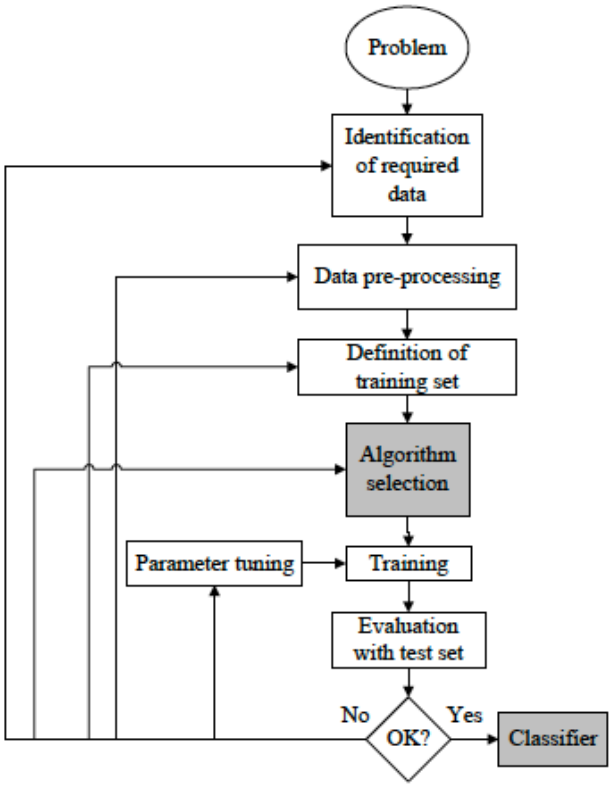



Figure 2.7: The process of supervised ML (Kotsiantis, 2007)

the elements of $x_i$ (see Table 2.1). The clustering is obtained by means of breaking the input data up into groups until some stability condition is reached.

Cluster analysis is often used to bring similar individuals into groups (Jain and Dubes, 1988) and to identify natural structures in a dataset (Jain, 2010; Gurrutxaga et al., 2011). Therefore, its used in many fields such as biology (Sneath et al., 1973), image processing (Chou et al., 2004) and pattern recognition (Mirkin, 2012). For instance, the aforementioned example of selective harvesting grouping grapes is a typical problem that can be addressed with clustering algorithms.

Regarding the clustering methods, Milligan and Cooper (1987) identify four categories: partitioning algorithms, hierarchical methods, overlapping clustering procedures and ordination techniques. This work considers the two main categories (Milligan and Cooper, 1987): partitioning algorithms and hierarchical methods. In particular, the following algorithms were used for MZ identification and for mapping cultivable land: Partition Around Medoids (PAM) and hierarchical clustering.

### Partitioning methods

Although literature proposes many partition methods, PAM and K-means (MacQueen et al., 1967) are among the most popular ones. In fact, they are very similar and the main difference between them is that PAM minimizes a sum of dissimilarities instead of a sum of squared

euclidean distances. The following lines describe PAM which is the partition algorithm used in this thesis.

PAM (Li, 2009) is a partitioning algorithm. Thus, it breaks the input data up into $k$ groups, defined in advance, until some stability condition is reached. It tries to find a set of representative objects called medoids which are centrally located in clusters. The objective function used is the sum of the distances from each object to the closest medoid.

Regarding the distance or dissimilarity among the samples, PAM builds a dissimilarity matrix from the input data based on a particular metric of distance (see Table 2.1). Let's put an example based on the Iris dataset considering these values of characteristic of flowers as input:

| | Sepal.Length | Sepal.Width | Petal.Length | Petal.Width |
|---|---|---|---|---|
| Flower 1 | 5.1 | 3.5 | 1.4 | 0.2 |
| Flower 2 | 5.0 | 3.4 | 1.6 | 0.4 |
| Flower 3 | 6.5 | 2.8 | 4.6 | 1.5 |
| Flower 4 | 6.3 | 2.3 | 4.4 | 1.3 |
| Flower 5 | 5.8 | 2.8 | 5.1 | 2.4 |
| Flower 6 | 4.8 | 3.4 | 1.9 | 0.2 |

Apparently, flowers 1, 2 and 3 are similar each other, the same for flowers 3, 4 and 5. Computing the dissimilarity matrix will show the average distances among them. For instance, these are the results considering the euclidean distance.

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Flower 2 | 0.3162278 | | | | |
| Flower 3 | 3.7920970 | 3.5805028 | | | |
| Flower 4 | 3.6180105 | 3.3985291 | 0.6082763 | | |
| Flower 5 | 4.4170126 | 4.1533119 | 1.2449900 | 1.4832397 | |
| Flower 6 | 0.5916080 | 0.4123106 | 3.4971417 | 3.3045423 | 4.0546270 |

It is clear that flowers 1 and 2 are the most similar (0.3162278), then flowers 2 and 6 (0.4123106), then flowers 1 and 6 (0.5916080) and so on. Computing the dissimilarity matrix using Manhattan distance lead to the same conclusions:

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Flower 2 | 0.6 | | | | |
| Flower 3 | 6.6 | 6.2 | | | |
| Flower 4 | 6.5 | 6.1 | 1.1 | | |
| Flower 5 | 7.3 | 6.9 | 2.1 | 2.8 | |
| Flower 6 | 0.9 | 0.7 | 6.3 | 6.2 | 7.0 |

Regarding how PAM builds the clusters, the algorithm involves three steps:

1. **Initialization**. Select, at random, the $k$ medoids from the data input. For instance, in the previous example, if $k = 2$ the algorithm would choose flower 3 and flower 4 as the representative objects of the clusters.

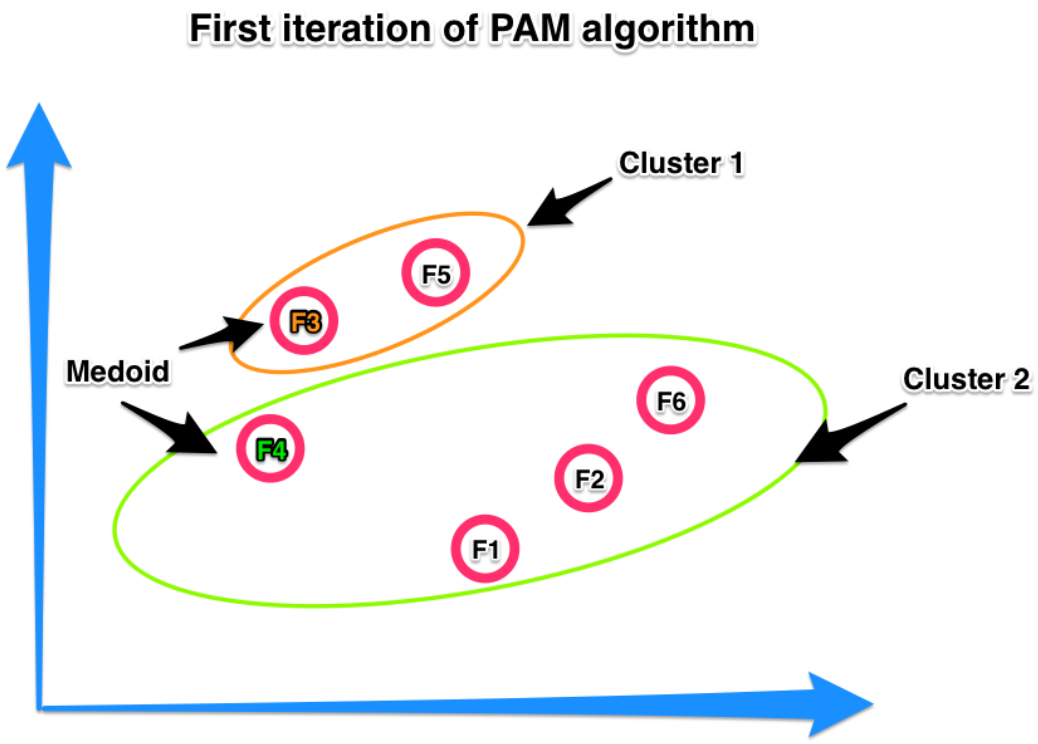



Figure 2.8: First step of the PAM algorithm. $F_i$ denotes the number of the flower.

2. **Assignment**. For each sample of the data input, locate the closest medoid to and assign it to the corresponding cluster. Continuing with the example, this means to compute the distance between the medoids, flowers 3 and 4, and the other flowers. As a result of these assignment, the groups would be (see Fig.2.8):

   - **Group 1**:

   | | Sepal.Length | Sepal.Width | Petal.Length | Petal.Width |
   |---|---|---|---|---|
   | Flower 3 (medoid) | 6.5 | 2.8 | 4.6 | 1.5 |
   | Flower 5 | 5.8 | 2.8 | 5.1 | 2.4 |

   - **Group 2**:

   | | Sepal.Length | Sepal.Width | Petal.Length | Petal.Width |
   |---|---|---|---|---|
   | Flower 1 | 5.1 | 3.5 | 1.4 | 0.2 |
   | Flower 2 | 5.0 | 3.4 | 1.6 | 0.4 |
   | Flower 4 (medoid) | 6.3 | 2.3 | 4.4 | 1.3 |
   | Flower 6 | 4.8 | 3.4 | 1.9 | 0.2 |

3. **Update**. For each cluster, compute the new medoid from the points assigned to the cluster. The new medoid will be the point that minimizes the dissimilarity to the rest of the elements in the cluster. In the case of the example, for the group 1 the new medoid

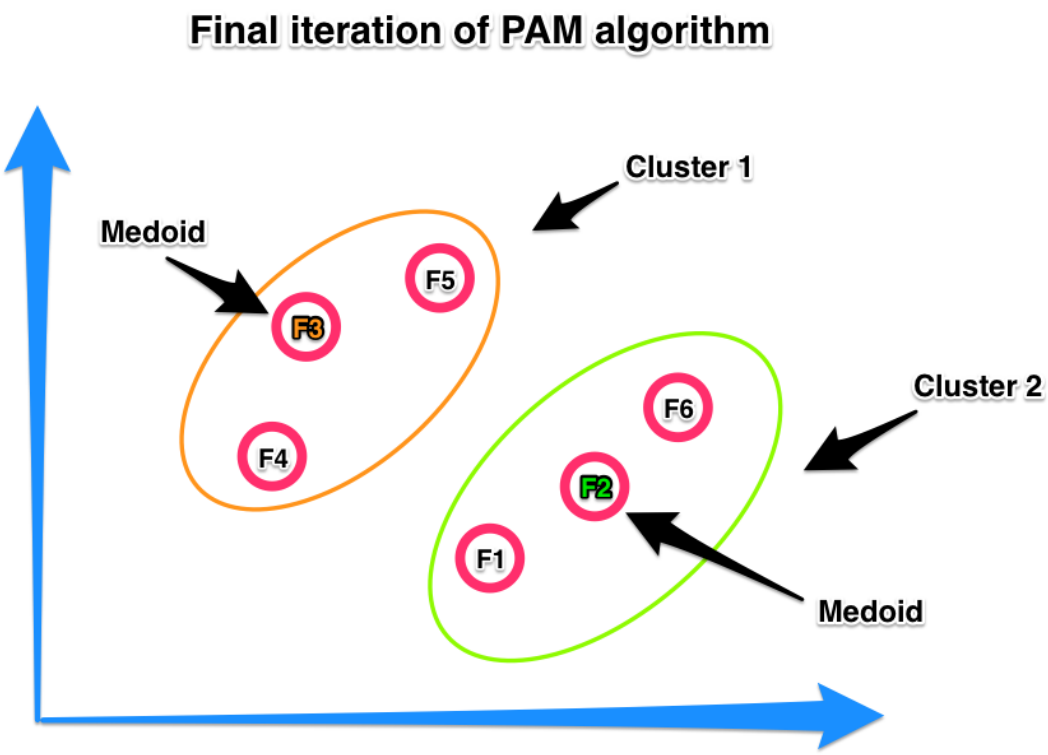



Figure 2.9: Final step of the PAM algorithm. $F_i$ denotes the number of the flower.

does not change. The average distance of flower 5 to the other flowers is greater than this average for flower 3. However, group 2 would have flower 6 as the new medoid.

Steps 2 and 3 are repeated until the clusters are no longer modified. In the case of the example, this condition is reached with the medoids: flowers 2 and 3; obtaining the following clustering (see Fig.2.9):

- **Group 1**:

| | Sepal.Length | Sepal.Width | Petal.Length | Petal.Width |
|---|---|---|---|---|
| Flower 3 (medoid) | 6.5 | 2.8 | 4.6 | 1.5 |
| Flower 4 | 6.3 | 2.3 | 4.4 | 1.3 |
| Flower 5 | 5.8 | 2.8 | 5.1 | 2.4 |

- **Group 2**:

| | Sepal.Length | Sepal.Width | Petal.Length | Petal.Width |
|---|---|---|---|---|
| Flower 1 | 5.1 | 3.5 | 1.4 | 0.2 |
| Flower 2 (medoid) | 5.0 | 3.4 | 1.6 | 0.4 |
| Flower 6 | 4.8 | 3.4 | 1.9 | 0.2 |

On the other hand, PAM requires to know in advance the number of groups $k$ for clustering. However, this number could be unknown. That could be the case, for instance, of clustering

areas of land that share similar properties. Section 2.1.4 explains several techniques to address the challenge of deciding the appropriate number of clusters.

### Hierarchical clustering

Hierarchical clustering provides solutions represented as a tree (see Fig.2.10) denominated dendrogram (Phipps, 1971). The construction of the tree follows two approaches. Agglomerative algorithms were clusters are built from bottom up (Zhao and Karypis, 2002) and partitioning algorithms were hierarchical clusters are built from top down (Milligan and Cooper, 1987). This thesis uses the first approach.

Agglomerative algorithms initially assign each object to its own cluster. Then the algorithm proceeds iteratively, at each stage joining the two most similar clusters and continuing until there is just a single cluster. However, there is other approach that builds the hierarchical clusters from top down using partitioning algorithms (Milligan and Cooper, 1987).

In order to merge the clusters $C_h$ and $C_k$, the algorithm use a *linkage* function based on the dissimilarity between them (Ding and He, 2002). Milligan and Cooper (1987) cites the following methods as linkage functions, $d_{ij}$ is a dissimilarity measure:

- **Single linkage** (Sneath et al., 1973), defined as the minimum dissimilarity between clusters:

$$d_{single}(C_h, C_k) = \min_{i \in C_h \ j \in C_k} d_{ij} \tag{2.16}$$

- **Complete linkage** (King, 1967), defined as the maximum dissimilarity between clusters:

$$d_{complete}(C_h, C_k) = \max_{i \in C_h \ j \in C_k} d_{ij} \tag{2.17}$$

- **Average linkage** (Jain and Dubes, 1988), defined as the average of all distances between two clusters.

$$d_{average}(C_h, C_k) = \frac{1}{|C_h| \ |C_k|} \sum_{i \in C_h, \ j \in C_k} d_{ij} \tag{2.18}$$

- **Minimum variance** (Ward Jr, 1963; Murtagh and Legendre, 2014), Ward's minimum variance method aims at finding compact, spherical clusters minimizing the total within-cluster variance.

Regarding the generic application of these methods in the agglomerative process, in Lance and Williams (1966) it is proposed a generalized recurrence formula for computing the dissimilarity between two clusters (see Equ. 2.19). This approach uses the same formula for each linkage method using the corresponding parameters described in Table 2.2 (Day and Edelsbrunner, 1984).

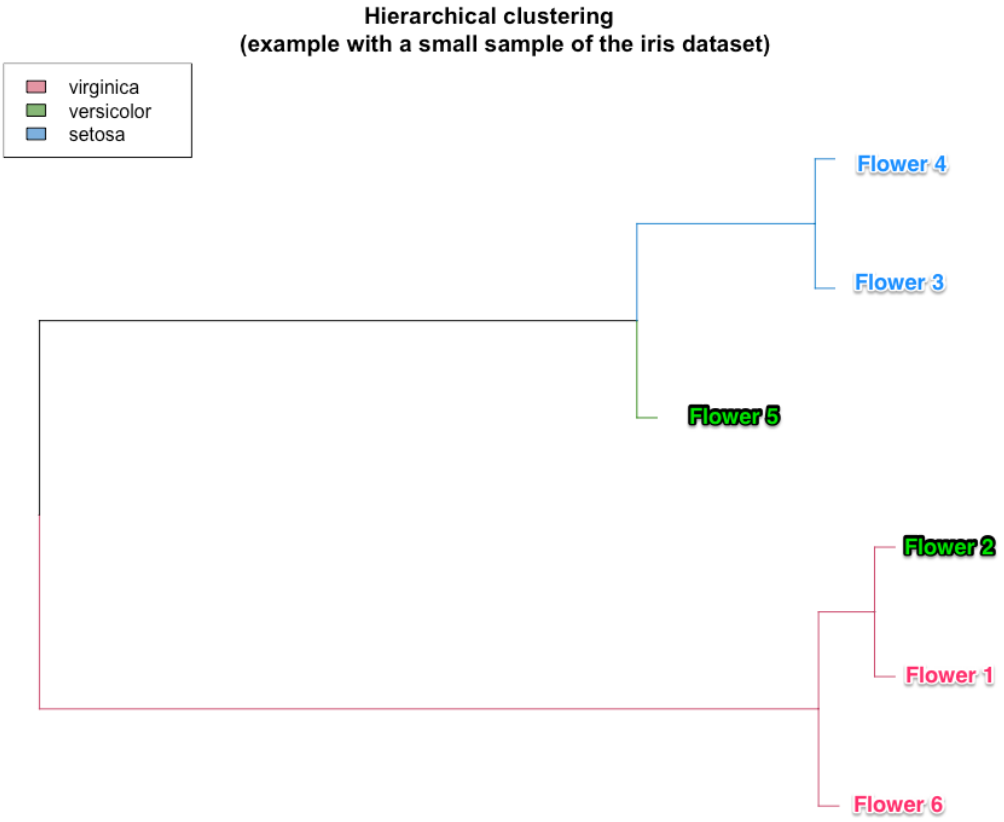



Figure 2.10: Example of dendogram with the same sample used for PAM

$$d(h,k) = \alpha_i d(i,k) + \alpha_j d(j,k) + \beta d(i,j) + \gamma |d(i,k) - d(j,k)|, \tag{2.19}$$

where $\alpha_i$, $\alpha_j$, $\beta$ and $\gamma$ are parameters of the linkage method (see Table 2.2).

| Linkage method | $\alpha_i$ | $\alpha_j$ | $\beta$ | $\gamma$ |
|---|---|---|---|---|
| Single linkage (Nearest neighbor) | $1/2$ | $1/2$ | $0$ | $-1/2$ |
| Complete linkage (Furthest neighbor) | $1/2$ | $1/2$ | $0$ | $1/2$ |
| Average linkage (UPGMA) | $\frac{n_i}{n_i+n_j}$ | $\frac{n_j}{n_i+n_j}$ | $0$ | $0$ |
| Minimum variance (Ward) | $\frac{n_i+n_k}{n_i+n_j+n_k}$ | $\frac{n_j+n_k}{n_i+n_j+n_k}$ | $\frac{-n_k}{n_i+n_j+n_k}$ | $0$ |

Table 2.2: Parameter values for the Lance and Williams formula for computing the dissimilarity between two clusters

### 2.1.4 Clustering validation and selection of the number of partitions

A large number of ways of evaluating the goodness of a clustering algorithm have been proposed in the literature. Such evaluation includes criteria for measuring the compactness of the clusters (within-cluster criteria) and their isolation (between-cluster criteria) (Cormack, 1971). One approach to this evaluation is the use of Cluster Validity Indices (CVI) (Halkidi et al., 2001) such as the Calinski-Harabasz index (Caliński and Harabasz, 1974). However, literature includes numerous CVI and could be challenging to select one of them. In this particular, in Milligan and Cooper (1985) one of the most extensive comparatives regarding the study of indices is studied.

One of the CVI studied by Milligan and Cooper's is the Calinski-Harabasz index. It evaluates the cluster validity considering the sum of the squared errors between different clusters and the squared differences of all objects in a cluster to their respective cluster center. The index is calculated as follows (Maulik and Bandyopadhyay, 2002):

$$CH(K) = \frac{(\sum_{k=1}^{K} n_k |z_k - z|^2)/(K-1)}{(\sum_{k=1}^{K} \sum_{i=1}^{n_k} |x_i^k - z_k|^2)/(n-K)} \tag{2.20}$$

where $n$ is the number of points, $K$ is the number of clusters, $n_k$ is the number of points in cluster $k$, $z_k$ is the centroid of cluster $k$, $x_i^k$ is the $i$-th point in cluster $k$ and $z$ is the centroid of the entire data set.

Other widely-used index (Wang et al., 2009) is the Silhouette coefficient (Rousseeuw, 1987). The index is based on the comparison of cluster tightness and separation. Using this information, a graph called Silhouette is built. This Silhouette graph shows which objects lie well within their cluster and which ones are merely somewhere in between clusters. The average silhouette width (see Equ. 2.21 and Fig. 2.11) provides an evaluation of the clustering validity and can be used to select an "appropriate" number of clusters.

$$s(i) = \frac{b(i) - a(i)}{max(a(i), b(i))}, \tag{2.21}$$

where $a(i)$ is the average dissimilarity of i to all other objects of A; $d(i, C)$ is the average dissimilarity of $i$ to all objects any cluster $C$ different from $A$; and $b = mind(i, C)$ , with $C \neq A$

Regarding the selection of the number $k$ of clusters, a common procedure is to generate the clustering for $k = 2$ to a certain value $n$ and select the $k$ with the best clustering distribution. However, the bottleneck of clustering algorithms is to properly select the best clustering distribution (Hartigan, 1975; Sugar and James, 2003). Let's consider, for example, the Silhouette coefficient as criterion for evaluating the cluster validity. To select the optimum $k$ according to this coefficient, it can be followed a procedure described in (Hennig, 2010). The procedure is briefly described in Alg.2.2 considering that the number of elements to cluster is $n$, $K^*$ is the maximum number of clusters (which is at most $n$) and $d(.,.)$ a distance.

To sum up, clustering processes can be executed following the steps described in (Milligan

**Algorithm 2.2** Selection of optimum $k$ according to Silhouette coefficient

**for** $j = 1, K^*$ **do**

**for** $i = 1, n$ **do**

$a(i) = \frac{\sum\limits_{j \in C_i} d(i,j)}{\#C_i}$, $C_i$ cluster of element $i$.

$b(i) = \min\limits_{k=1...j} d(i, C_l), \forall$ cluster $C_l$

$s(i) = \frac{b(i)-a(i)}{max\{(b(i),a(i)\}}$,

**end for**

$s^j_{avg} = \frac{\sum_{i=1}^{n} s(i)}{n}$

**end for**

$k = argmax\{s^j_{avg}\}$

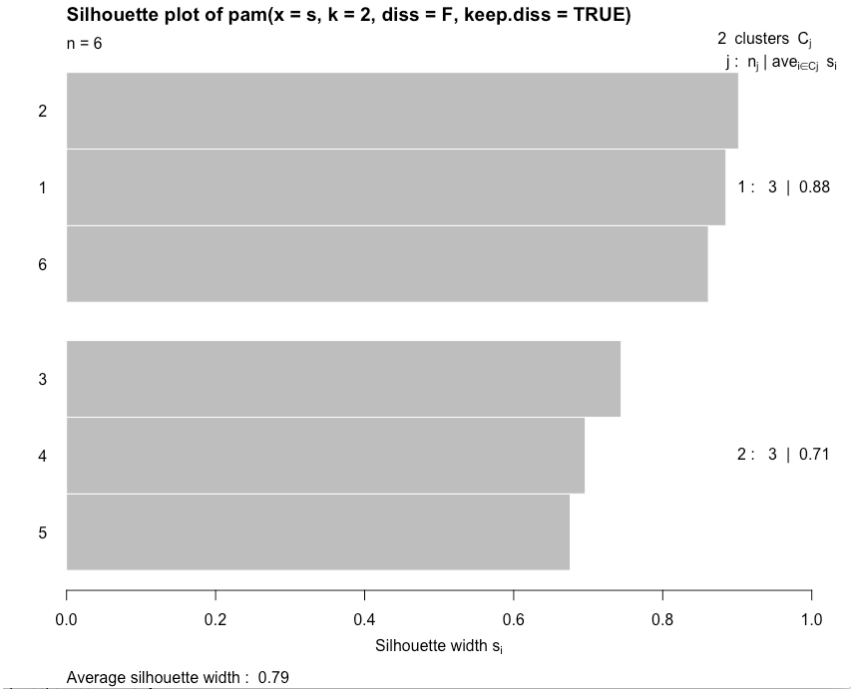



Figure 2.11: Example of a average silhouette width for two clusters

and Cooper, 1987).

1. Select the identities to be clustered, choosing a representative sample of elements.
2. Select the variables to be used in the cluster analysis, considering those containing "sufficient" information for the clustering.
3. Decide if the data should be standardized.
4. Select a measure of similarity, such as correlation, or a measure of dissimilarity such as distance.
5. Select the clustering method considering the different types of cluster structures.
6. Determine the number of clusters.
7. Interpret, test and replicate the resulting cluster analysis.

### 2.1.5 Feature selection

In the past few years, technological advances in data acquisition, data storage and ICT led to a explosion of data (Mayer-Schönberger and Cukier, 2013). Such huge amount of data is generated by different domains such as scientific (Guo et al., 2014), healthcare (Raghupathi and Raghupathi, 2014), game industry (El-Nasr et al., 2013) and agriculture (Ludena and Ahrary, 2013) (see Section 3). This scenario provides the opportunity for extracting valuable information and support decision-making process by means of data analysis and data mining (Hu et al., 2014).

However, the application of ML algorithms with tens or hundreds of thousands of variables is a challenge for algorithm performance (Yang and Pedersen, 1997b). Therefore, techniques to reduce dimensionality of data are required (Guyon and Elisseeff, 2003; Combarro et al., 2005) in order to identify irrelevant and redundant variables (Kohavi, 1995). Literature proposes a huge number of feature selection techniques that may be categorized as follows (Guyon and Elisseeff, 2003):

- **Feature subset selection**. This approach selects subsets of features as a pre-processing step, independently of the chosen predictor (Kohavi, 1995). Methods falling in this category use, for instance, decision trees (Cardie, 1993) or wrapper methods (Kohavi, 1995).
- **Feature ranking** (Bekkerman et al., 2003). It is a filter method (Guyon and Elisseeff, 2003) which assigns a weight to each feature according to their importance for the class providing a ranking of the relevance of each feature (Chang and Lin, 2008; Díaz et al., 2011). This approach includes IG (Hunt et al., 1966) and Correlation-based Feature Selection (CFS) (Hall and Smith, 1997). Both approaches are used in this thesis because its simplicity, scalability, and good empirical success (Guyon and Elisseeff, 2003).

As a difference between both approaches, feature subset selection chooses the best set of features whilst feature ranking selects individual features.

### Feature frequency

The frequency of occurrence of feature values within a class is a simple method used for feature ranking. The idea behind this method is that attribute values with low frequency are either non-informative for category prediction, or not influential in global information (Yang and Pedersen, 1997b).

### Information Gain

IG is a representative example of feature ranking (Yang and Pedersen, 1997b). It is studied in Díaz et al. (2004) takes into account either the presence of the attribute value in the category or its absence, and it can be defined Yang and Pedersen (1997a) by

$$IG(a,c) = P(a)P(c/a)\log(\frac{P(c/a)}{P(c)}) + P(\overline{a})P(c/\overline{a})\log(\frac{P(c/\overline{a})}{P(c)})$$

where $P(a)$ is the probability of the appearance of the attribute $a$ in an instance, $P(c/a)$ is the probability that an instance belongs to the category $c$ knowing that the attribute $a$ appears in it, $P(\overline{a})$ is the probability that the attribute $a$ does not appear in an instance and $P(c/\overline{a})$ is the probability that an instance belongs to the category $c$ if we know that the attribute $a$ does not occur in it.

Usually, these probabilities are estimated by means of the corresponding relative frequencies. *Expected cross entropy for attribute value* ($CET$) (Mladenic and Grobelnik (1999)) only takes into account the presence of the attribute value in the category. Its expression is

$$CET(a,c) = P(a)\cdot P(c/a)\cdot \lg\frac{P(c/a)}{P(c)}$$

There are more methods following the feature ranking approach. However, in general, IG performance is remarkable (Yang and Pedersen, 1997b) and it obtains good results.

### Correlation-based Feature Selection

CFS (Hall and Smith, 1997). CFS is a simple filter algorithm that ranks feature subsets according to a correlation based heuristic evaluation function. The bias of the evaluation function is toward subsets that contain features that are highly correlated with the class and uncorrelated with each other. Irrelevant features should be ignored because they will have low correlation with the class. Redundant features should be screened out as they will be highly correlated with one or more of the remaining features. The acceptance of a feature will depend on the extent to which it predicts classes in areas of the instance space not already predicted by other features. The evaluation function is

$$M_S = \frac{kr_{cf}}{\sqrt{k + k(k-1)r_{ff}}}, \tag{2.22}$$

where $M_S$ is the merit of a feature subset S containing $k$ features, $r_{cf}$ is the average feature-

class correlation for features $f$ in $S$ and $r_{ff}$ is the average feature-feature intercorrelation. The numerator represents how predictive the set of features is, while the denominator represents the amount of redundancy contained in subset $S$.

### 2.1.6 Performance evaluation

There are several approaches for evaluating the performance of ML algorithms. In particular, the evaluation of classifiers is measured, in general, through a confusion matrix (Sokolova and Lapalme, 2009) (see Table 2.3).

| Class | Classified as Positive | Classified as Negative |
|---|---|---|
| Positive | TP | FN |
| Negative | FP | TN |

Table 2.3: Confusion matrix for performance evaluation of ML algorithms

Where, True negatives (TN) are the elements correctly classified as not belonging to the class, False positives (FP) as the elements incorrectly classified as belonging to the class, False negatives (FN) as the elements incorrectly classified as not belonging to the class and True positives (TP) as elements correctly classified as belonging to the class.

From these values, it can be computed several performance measures (Jardine and van Rijsbergen, 1971) that are widely used when evaluating the behaviour of ML techniques (Sebastiani, 2002):

- **Precision**, related with the overall effectiveness of the classifier (Sokolova and Lapalme, 2009):

$$P = \frac{TP}{TP + FP} \tag{2.23}$$

- **Recall (Sensitivity)**, a measure of the effectiveness of a classifier to identify positive labels:

$$R = \frac{TP}{TP + FN} \tag{2.24}$$

- **Specificity**, measures how effectively a classifier identifies negative labels citep-sokolova2009systematic:

$$S = \frac{TN}{TN + FP} \tag{2.25}$$

Of course, it would desirable these values to be as close to 1 as possible, but it is well-known that there is a trade-off between them (Sebastiani, 2002; Powers, 2007). For that reason, these values are combined in more meaningful performance measures such as the F1 score (Jardine and van Rijsbergen, 1971; Powers, 2007), which is the harmonic mean of precision and sensitivity:

$$F1score = \frac{2TP}{2TP + FP + FN} \tag{2.26}$$

It should be notice that this approach is also applicable for evaluating cluster algorithms, whether it is possible to compare the clusters obtained from the clustering algorithm to some other previously defined.

Other popular way of presenting the overall behaviour of a classifier relies on the Receiver Operating Characteristic (ROC) (see Equ. 2.27) by means of the ROC curve (Fawcett, 2006) (see Figure 2.12). It plots the sensitivity against specificity as the discrimination threshold of the method is varied. Notice that the bigger the area under the ROC curve (AUC) (see Equ. 2.28), the better the classifier is.

$$ROC = \frac{P(x|positive)}{P(x|negative)}, \tag{2.27}$$

where $P(x|C)$ is the conditional probability that $x$ has the class $C$.

$$AUC = \frac{recall + specifity}{2} \tag{2.28}$$

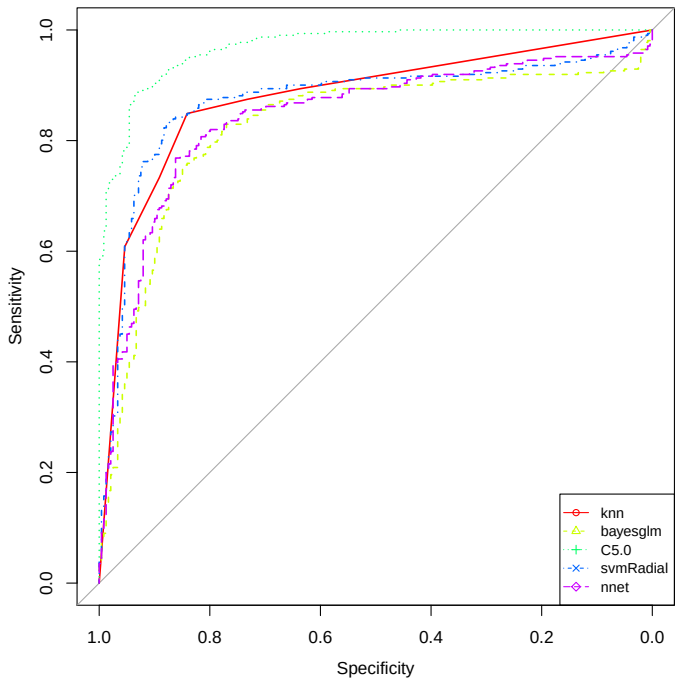



Figure 2.12: Example of ROC curves. Horizontal and vertical axis represents the false positive rate and true positive rate, respectively.

### Multi-class evaluation

Performance evaluation for classification with more than two classes is based on a generalization of the aforementioned measures. One approach is to compute the micro-averaging as follows:

- **Precision**:

$$P_\mu = \frac{\sum_{i=1}^{l} TP_i}{\sum_{i=1}^{l} (TP_i + FP_i)} \tag{2.29}$$

- **Recall (Sensitivity)**:

$$R_\mu = \frac{\sum_{i=1}^{l} TP_i}{\sum_{i=1}^{l} (TP_i + FN_i)} \tag{2.30}$$

- **Specificity**:

$$S_\mu = \frac{\sum_{i=1}^{l} TN_i}{\sum_{i=1}^{l} (TN_i + FP_i)} \tag{2.31}$$

- **F1score**:

$$F1score_\mu = \frac{\sum_{i=1}^{l} 2TP_i}{\sum_{i=1}^{l} (2TN_i + FP_i + FN_i)} \tag{2.32}$$

Where it is considered $C_i$ classes, $TP_i$ are true positive for $Ci$, $FP_i$ false positive, $FN_i$ false negative and $TN_i$ true negative. In order to compute them it can be used a confusion matrix where the rows are the classes and the column the predicted values.

For illustration proposes, let's consider the following flowers classified as shows the column Predicted:

| | Predicted | Class |
|---|---|---|
| Flower 1 | Setosa | Setosa |
| Flower 2 | Setosa | Setosa |
| Flower 3 | Setosa | Versicolor |
| Flower 4 | Versicolor | Versicolor |
| Flower 5 | Virginica | Virginica |
| Flower 6 | Virginica | Setosa |

Taking into account the predicted value and the class, the actual type of flower, the confusion matrix would be the following:

| Class | Classified as Setosa | Classified as Versicolor | Classified as Virginica |
|---|---|---|---|
| Setosa | 2 | 0 | 1 |
| Versicolor | 1 | 1 | 0 |
| Virginica | 0 | 0 | 1 |

Table 2.4: Confusion matrix for performance evaluation with multi-class classification

Regarding regression models, there numerous metrics that provide a measure of how well the prediction fits to the original data. In particular, the correlation coefficient (Pearson, 1920; Fisher, 1915; Lawrence and Lin, 1989) is a well-know metric (Lee Rodgers and Nicewander, 1988). It is defined as follows:

$$r^2 = b\ b', \tag{2.33}$$

where $b$ is:

$$b = \frac{n\sum x\ y - \sum x \sum y}{n\sum x^2 - (\sum x)^2} \tag{2.34}$$

and $b'$ is:

$$b' = \frac{n\sum x\ y - \sum x \sum y}{n\sum y^2 - (\sum y)^2} \tag{2.35}$$

**Cross-validation**

Cross-validation is an statistical method (Efron and Gong, 1983) for model selection in SML algorithms. It provides an estimation of the performance evaluation of the model with training data independent from the set of data for testing the performance. In this regard, it divides an input dataset into two mutually exclusive partitions: training set and test set; the former to train the algorithm and the test set for evaluating its accuracy (Kohavi et al., 1995). Such training and evaluation is repeated $k$ times with different partitions of the same dataset. As a result, the average of the performance measure is compute with all the measures obtained in each iteration (Shao, 1993).

There are several methods to decide how the partition is done. $k$-fold cross-validation splits the dataset into $k$ disjoint partitions (folds) of similar size (Kohavi et al., 1995). Each iteration takes a different partition for testing whilst the other $k$-1 folds are used for training. Leave-one-out cross-validation is particular case of $k$-fold cross-validation with $k = 1$ (Kohavi et al., 1995). Hence each iteration uses a test set with only one instance. In Stratified cross-validation (Kohavi et al., 1995) the partitions contains similar proportion of classes as the original dataset.

Regarding the performance measures, Kohavi et al. (1995) found that $k$-fold cross-validation with values of $k$ between 10 and 20 reduces the variance while increases the bias. However, according his study, ten-fold stratified cross-validation is the best method for model selection.

#### Bootstraping, bootstrap aggregating and boosting

Bootstraping (Efron, 1979) is a method that generates new training sets by sampling with replacement the instances of the input dataset in order to generate different classifiers. Such method is used for approaches intended to obtain a composite classifier combining multiple classifiers by voting (Quinlan, 1996a). Bootstrap aggregating (bagging) (Breiman, 1996) and boosting (Freund and Schapire, 1995) follow this approach. Although these methods generate multiple classifiers manipulating the input dataset, they differ in how the training set is generated and the voting system to form a composite classifier.

Generally, the use of these methods improve predictive accuracy (Quinlan, 1996a). However, boosting may produce degradation on some datasets (Quinlan, 1996a).

## 2.2 Related work

This section gathers literature related with the application of machine learning techniques in agricultural environments. Although the main focus of this thesis relays on automatic land delimitation, the identification of MZs and crop yield forecasting, other applications are included for better describing the framework.

### 2.2.1 Automatic land delimitation and land cover classification

Regarding the application of machine learning techniques in agricultural environments for land cover classification, there are many different approaches. Support Vector Machines (Gualtieri and Cromp, 1999; Huang et al., 2002; Pal and Mather, 2005; Marconcini et al., 2009), K-Nearest Neighbors Classifier (Zhu and Basir, 2005; Zhang et al., 2013) and Random forest (Gislason et al., 2006) are well-known examples of these techniques. Artificial Neural Networks were also used for land cover classification in (Kavzoglu and Mather, 2003) and it was concluded that accuracies can differ significantly depending on the selection of the network structure and parameter values.

On the other hand, Friedl and Brodley (1997) proposed the use of Decision trees and HAN et al. (2011) used a method based on C5.0 and Normalized Difference Vegetation Index (NDVI) to create a knowledge base of rules indicating that the texture could potentially be applied in object-oriented classification. Klein et al. (2012) developed a method also based in C5.0 to study the evolution of land cover using MODIS time-series. Ottinger et al. (2013) presented a supervised approach to classify Landsat 7 images to monitor changes in the Yellow River Delta.

However, most of these works do not use multitemporal imagery in the training process and the evaluation is done with only one or two images.

### 2.2.2 Identification of management zones

In the scientific literature, there are several works related to the automatic delimitation of land using different data sources, such as the classification of apparent soil electrical conductivity (Johnson et al., 2003; Peralta and Costa, 2013) or the analysis of yield maps (Blackmore et al.,

2003). In addition, Ortega and Santibáñez (2007) compared the results of the use of chemical properties of the soil with several techniques such as PCA and cluster analysis. Simbahan and Dobermann (2006) tested supervised classification algorithms with different datasets including soil maps, digital elevation models and apparent soil electrical conductivity. In the same line, other authors have considered soil properties (Moral et al., 2011; Fu et al., 2010) and also yield and crop quality (Aggelopooulou et al., 2013).

Schuster et al. (2011) identified homogeneous zones of a cotton field considering two datasets: the first one with two estimators of the yield and the second one considering geo-referenced field properties such as topographical characteristics and treatments applied to the field. The use of biophysical features such as annual moisture deficit/surplus and mean annual precipitation was explored by Liu and Samal (2002). The authors tested the same dataset with k-means and fuzzy algorithms concluding that a fuzzy approach generates more accurate delineations.

Kumar et al. (2011) studied the use of the k-means algorithm with the MODIS-based greenness index and the seasonal leaf area index (DAAC, 1990), developing a parallel implementation able to delimit 1,000 agroecozones in 700 seconds using 2,048 processors. On the other hand, Duro et al. (2012) studied the classification of agricultural landscapes by means of image analysis techniques with SPOT-5 (VITO, 1998) satellite imagery.

The major drawback of many of these approaches is that they require some kind of in-field exploration in order to measure the values of the features. On the other hand, biophysical features can be acquired from public sources such as agro-meteorological stations but the results will depend on the number and the geographical distribution of the stations. However, remotely-sensed imagery from satellites does not require in-field exploration and there are satellite programs such as Landsat (USGS, 1972) or the Sentinel missions (ESA, 2014) with global coverage and free distribution of their data products. Kumar et al. (2011) followed this approach but their approach for land delimitation only considers two features with a low spatial resolution for PA applications (250 m.). Even though Duro et al. (2012) have worked with high spatial resolution imagery (10 m.), the data products they used are not publicly available.

### 2.2.3 Crop yield forecasting and yield planning

Yu et al. (2010) proposed a crop yield forecasting model based on the combination of artificial neural networks (ANN). The ANN were trained with data related with the nutrient concentration of the soil and the amounts of fertilizer provided in the campaign. Although the results of this work are positives possibly may not be generalized because of some assumptions in the model such as the same type of crop and a set of fixed values for soil moisture and PH.

Zhou et al. (2007) applied the Grey-Markov forecasting model (Yidan, 1992) to yield prediction. The authors studied its accuracy comparing this model to the Grey Model GM(1,1) (Sun, 1991). The study used crop yield data from the National Bureau of Statistics of China and concluded that the Grey-Markov method can achieve better results than the Grey Model GM(1,1) for yield prediction.

Charvat and Pavel (2012) developed a service which is capable to generate yield planning

for next season maximizing the expected profit. The heterogeneous data sources processed by this service include precision agriculture information systems and Sensor Observation Services (SOS) databases.

On the other hand, Gonzalez-Sanchez et al. (2014) compared the predictive accuracy of ML and linear regression techniques for crop yield prediction in ten crop datasets. Multiple linear regression, M5-Prime model trees, Perceptron Multilayer Neural Networks, SVR and KNN methods are ranked. M5-Prime and KNN techniques obtain the lowest errors and the highest average correlation factors. M5-Prime, which achieves the largest number of crop yield models with the lowest errors, is considered a very suitable tool for massive crop yield prediction in agricultural planning. In addition, it is more interpretable than KNN. Other approaches combine partial least square models and spectral imaging technology (Ye et al., 2007).

The model tree technique (see, for example, Frank et al. (1998) or Samadi et al. (2014)) is based on combining decision trees with linear regression functions at the leaves. Among these learners, M5 builds multivariate linear tree-based models, producing a classification based on piecewise linear functions (Quinlan et al., 1992). Regression trees, have been demonstrated as suitable methods to crop yield prediction. The most common algorithms to build regression trees are CART (Breiman et al., 2001), M5 (Quinlan et al., 1992) and M5-Prime (Wang and Witten, 1996). The strategy to construct the tree is similar in all of them (ElGibreen and Aksoy, 2015).

### 2.2.4 Variable-rate fertilization

The estimation of the amount of fertilizer required by a specific crop area allows the use of variable-rate application (VRA) of fertilizer. This not only implies economic advantages but also may have a positive impact on environment if it entails a reduction on the use of fertilizers. In this regard, Panagos *et al.* in (Panagos et al., 2013) improved the estimations of Soil Organic Carbon (SOC) content by means of comparing the outcomes of the Organic Carbon Content In Topsoils In Europe (OCTOP) (Jones et al., 2005) model with the soil samples collected from 20,000 points in 23 UE countries and available on the LUCAS database (http://esdac.jrc.ec.europa.eu/projects/lucas).

Zhang *et al.* developed a model in (Zhang and Han, 2002) for the representation of crop yield in both the spatial and temporal dimensions dividing the crop land in small pieces of land or blocks. The model takes into account the yield from the previous years and calculates confidence levels for low, average and high crop production for each block. This provides support decision-making about VRA of fertilizers by means of comparing the confidence levels for the same block along the temporal dimension.

Seppelt and Voinov (2002) developed a methodology to optimize the use of the land. The method tries to maximize a specific crop yield goal considering multiple factors regarding the agricultural production such as the current price of fertilizer products or the results of soil analysis. The optimization is performed by means of simulation models considering different land-uses scenarios ranging from wheat and corn plantations. Finally, the method generate

land-use maps and fertilization maps.

### 2.2.5 Automatic plant identification

Automatic discrimination between soil and plants based on the dominant spectral component, green in the first case and red for the second one, may result inaccurate if plants are partially or totally covered for other materials on ground, for example, as a result of heavy rain conditions. To address this issue, in Guerrero et al. (2012) it is proposed a classification strategy by means of supervised learning algorithms. Specifically, SVM trained with vegetation colour indexes from RGB images and its corresponding black and white images. These last ones obtained from thresholding techniques and achieving a success rate of 93.1% for a test case with 40 images of corn crops.

On the other hand, neural networks and hyper spectral imagery were successfully used for clustering and crop identification (Seiffert et al., 2010). Whilst Meyer et al. (2004) studied the precision of fuzzy clustering algorithms on plants identification concluding that the Gustafson-Kessel (GK) algorithm (Gustafson and Kessel, 1978) is more precise than the fuzzy c-means (FCM) algorithm (Pal and Bezdek, 1995). Both algorithms were enhanced with Zadeh's (Z) fuzzy intensification technique (Zadeh, 1965).

# Chapter 3

# HETEROGENEOUS DATA SOURCES FOR PRECISION AGRICULTURE

Agriculture domain usually deals with multiple heterogeneous data sources such as networks of agro-meteorological stations, weather forecasts, soil properties, crop properties, satellite imagery or crop treatments. Although the potential of this information for providing solutions in the PA context is high, data should be properly collected, treated and analysed. However, such volume of data is not always available or easily accessible.

Hence, platforms are needed to gather, unify and analyse these data converting them into valuable information (Kitchen et al., 2002), not only for experts but also for farmers. This kind of platforms could provide specific and high-value applications and services for the support in the planning and decision-making processes of different stakeholders groups related to the agricultural and environmental domains. However, data processing may also become complex if the datasets are quickly updated, growing in a way that storage and analysis become challenging. Hence, we are also facing a Big Data Analysis problem that challenges the learning and the decision making processes.

On the other hand, remote-sensed imagery as those provided by satellites or drones is a valuable resource for PA (Ormeño Villajos et al., 2008). The possibility of obtaining free-of-charge satellite images, as those provided by the Landsat program (USGS, 1972), opens new paths that can be used to approach a more cost-effective solution to the PA. The images can be used for many purposes such as obtaining indicators related with the health and moisture of the crops (Kriegler et al., 1969; Qing Liu and Huete, 1995; Fensholt and Sandholt, 2003), inferring the properties of soil (Ludwig et al., 2008; Xie et al., 2011), land classification (Gualtieri and Cromp, 1999) or yield forecasting (Ye et al., 2007). However, the analysis of satellite imagery faces challenges related with the diversity of data sources, heterogeneous data formats and the massive volumes of data generated (see Section 3.1.5).

This chapter is organized as follows. Section 3.1 reviews open data satellite imagery with global coverage that may be used as data source for modelling smart agro-services. Section 3.2 describes climate and meteorological data sources. Finally, Section 3.2.5 shows agricultural data related with soil and crops.

## 3.1 Open data satellite imagery for precision agriculture services

Spectrometers, luxometers and multi-spectral cameras are some examples of measurement instruments which are equipped on small-size unmanned air vehicles and agricultural machinery. This kind of instruments combined with Global Positioning System (GPS) measure the radiation emitted and reflected by crops and allow to observe multiple geo-located variables which affect crops (Bingfang et al., 2010) such as chlorophyll or moisture. In this regard, the use of satellite imagery allows to observe multiple variables which affect crops (Bingfang et al., 2010). On-board satellite instruments related to environmental and Earth observation, measure electromagnetic radiation emitted and reflected by the observed objects. Based on these radiometric data it is possible to obtain valuable variables and indicators related with the agricultural domain (Ormeño Villajos et al., 2008) (see Section 3.3.6). These remotely sensed data may be used for the estimation of soil properties and the recognition of spatial patterns (Bhatti et al., 1991).

On the other hand, remote-sensed imagery acquired by satellite instruments can provide valuable information for planning and decision-making processes in the agriculture and environmental domains.

This section reviews satellite programmes related with land monitoring as potential data sources for modelling smart agro-services. In particular, those with global coverage and data products free to download.

### 3.1.1 NASA Land Processes Distributed Active Archive Center

NASA Land Processes Distributed Active Archive Center (LP DAAC) (NASA and USGS, 1990) provides access to historic data files from the MODIS instrument, operated from TERRA and AQUA satellites. MODIS offers valuable data related to agriculture such as the vegetation indices, land temperature or land cover characteristics (see Table 3.1 and Section A.1).

| Name | Data Product | Res. (m) | Frequency |
|---|---|---|---|
| MOD09GA | Surface Reflectance | 500m | Daily |
| MYD09GQ | Surface Reflectance | 250m | Daily |
| MOD11A1 | Land Surface Temperature and Emissivity | 1000m | Daily |
| MOD13Q1 | Vegetation Indices | 250m | 16 days |
| MOD15A2 | Leaf Area Index | 1000m | 8 days |
| MOD14A1 | Thermal Anomalies and Fire | 1000m | Daily |
| MOD44B | Vegetation Continuous Fields | 250m | Annual |

Table 3.1: MODIS data products related with agriculture

The data products are publicly available by HTTP or direct search on the Web (https://lpdaac.usgs.gov/get_data/data_pool)

### 3.1.2 Earthnet Online

The European Space Agency (ESA) manages Earthnet Online (Landgraf and Fusco, 1997). It is a web portal that provides a data pool of numerous satellite instruments, including MODIS and the Compact High Resolution Imaging Spectrometer (CHRIS) (Cutter et al., 2000) as well as satellites like Landsat 8.

CHRIS monitors the environment with resolutions of 18m and 36m and five different angles of view. The data products of this instrument are free available with HDF format. Its potential for agriculture includes:

- Monitoring the vegetation growth (Kneubuehler et al., 2006)
- Estimation of the chlorophyll content in leaves (Delegido et al., 2008)
- Automatic generation of land cover maps (Duca and Del Frate, 2008)

On the other hand, as a collaboration between the National Aeronautics and Space Administration (NASA) and the U.S. Geological Survey, the Landsat program (USGS, 1972) provides several operational satellites. The most recent one is the Landsat 8 that takes about 400 images every day using its on-board instruments: the Operational Land Imager (OLI) and the Thermal Infrared Sensor (TIRS). OLI collects data from 9 spectral bands with a spatial resolution of 30m, except for the band 8 with 15m. TIRS works on two thermal bands with a spatial resolution of 100m resampled to 50m. Both data products are distributed in GeoTIFF format.

### 3.1.3 Copernicus

This European Programme for the establishment of a European capacity for Earth Observation (http://www.copernicus.eu/), provides geographic information, land cover data and related variables such as vegetation health. Copernicus provide data, among other sources, from:

1. SENTINEL missions (ESA, 2014). Six satellite missions that are mainly intended for environmental monitoring and climate monitoring. Each mission has its own specific objectives

and is conformed for two or more satellites, reducing the time required for obtaining new data from the same location on Earth (revisit time). Currently, March of 2016, there are three operational missions:

- **SENTINEL-1**. The mission does not provide optical imagery. However, these satellites observe land and oceans taking radar data by means of a Synthetic Aperture Radar (SAR). The instrument uses a technique called radar interferometry that can detect changes between radar images of the same location. This kind of data could be beneficial in order to improve crop yield recommendations by means of predicting the growth development.
- **SENTINEL-2**. It provides two satellites equipped with high-resolution optical imagery focussed on vegetation, soil and coastal areas. The instrument responsible for the imagery acquisition is the Multispectral Imager (MSI) which collects 13 spectral bands (10m), 6 in the red and short-wave infrared spectrum (20m) and 3 bands for atmospheric correction (60m). Considering both satellites, the revisit time is 5 days and taking into account that the OLI instrument of the Landsat 8 covers 9 spectral bands (30m.) with a revisit time of 15 days, the SENTINEL-2 mission could lead remarkable improvements in those PA scientific works relying on open data satellite imagery.
- **SENTINEL-3**. The mission mainly focus on ocean monitoring providing optical imagery, thermal imagery and altimetry. The data products also provide land and atmospheric applications however the spatial resolution, more than 300m, does not improve the resolution of other instruments such as MODIS, OLI and TIRS.

The rest of the Sentinel missions, SENTINEL-4, SENTINEL-5 and SENTINEL-5P are dedicated to monitor the air quality .

2. Earth monitoring satellites included in the Contributing Missions. Currently, there are about 40 missions grouped in 5 categories or Mission Groups based on the type of mission. Some of them are atmospheric missions, optical High Resolution (HR) or optical Very High Resolution (VHR) missions. One example that falls into this last category is the satellite instrument VITO-SPOT-VEGETATION (VITO, 1998). This instrument offers the following family of data products related to PA, which are provided in Hierarchical Data Format (HDF) (Group et al., 2000) format:

   - **VGT-P**. Apparent reflectance perceived at the top of atmosphere. The temporal resolution of this data product is daily and the spatial resolution is 1km.
   - **VGT-S1**. The maximum NVDI (a vegetation index, see Section 3.3.6) computed from all the measurements taken for the same geographical area on the same day. The temporal resolution is daily and the spatial resolution is 1km.
   - **VGT-S10**. It is similar to VGT-S1 but the maximum values correspond to 10 days.

- **VGT-D10**. These data products are based on a bidirectional reflectance distribution function from the data obtained in 10 days. The spatial resolutions are: 1km, 4km o 8km. In contrast to VGT-S data products, these family of data products include indicators of sea/land, ice-snow/ice-snow absence and clouds/clear.

The mechanism to access to Copernicus satellite data product is the Copernicus Space Component Data Access system (CSCDA) and is restricted to:

- Institutions and Bodies of the EU
- Participants of a research project financed under the Union research programmes - Space
- Participants of a research project financed under the Union research programmes - Non-space
- Public Authorities
- International Organisations and Non-Governmental Organisations (NGOs)

### 3.1.4 The National Oceanic and Atmospheric Administration

The National Oceanic and Atmospheric Administration (NOAA) counts with satellites like the Geostationary Operational Environmental Satellites (GOES) which are mainly involved in climate models for meteorological forecasting. However NOAA also provides data products related with PA (*NOAA Products*, 2005) from the Advanced Very High Resolution Radiometer (AVHRR). For instance the Vegetation Health Product provide the following indices which fluctuate from 0 to 100, reflecting changes in vegetation conditions from extremely bad to optimal:

- **Vegetation Condition Index (VCI)** (Kogan, 2002), for estimation of cumulative moisture impacts on vegetation:

$$VCI = 100\frac{NDVI - NDVImin}{NDVImax - NDVImin}, \tag{3.1}$$

  where $NDVI$ is the Normalized Difference Vegetation Index (see Section 3.3.6) and $NDVImax$ and $NDVImin$ are their multi-year absolute maximum and minimum, respectively.

- **Temperature Condition Index (TCI)** (Kogan, 2002), for estimation of thermal impacts:

$$TCI = 100\frac{BTmax - BT}{BTmax - BTmin}, \tag{3.2}$$

  where $BT$ is the brightness temperature and $BTmax$ and $BTmin$ their multi-year absolute maximum and minimum, respectively.

- **Vegetation Health Index (VHI)** (Kogan, 2002), for estimation of combined conditions:

$$VHI = a \cdot VCI + (1 - a) \cdot TCI, \tag{3.3}$$

  where $a$ is determined by experience, currently, $a = 0.5$.

NOAA requires an online free registration for the download of the data products which are stored in a proprietary format called NOAA Level 1b (Oceanic and , NOAA).

### 3.1.5 Challenges of the analysis of satellite imagery

The use of satellite imagery involves challenges related to their acquisition, processing and analysis. Mainly, in this work it was identified a diversity of data sources providing massive volumes of data with different types of structures. The main challenges could be the following ones:

- **Diversity of data sources**. The acquisition of data may be considered as a challenge itself. First, there is not a common approach to seek the data. Satellite data is aggregated from different organisms such as ESA, NASA or NOAA. In addition, they are usually provided by web portals. However, each portal has its own functionality. Second, the download of data via File Transfer Protocol (FTP) or via web services is rarely supported. Therefore, it is difficult to schedule downloads of new data products and, frequently, human intervention is required for visiting the web portals and for the download of the data.

- **Heterogeneous data structures**. There is no a common structure of data for the storage of data products. Each satellite instrument could use its own structure. For instance, NOAA data products use a proprietary format whilst other providers use formats like HDF or Network Common Data Format (netCDF) (Melton et al., 1995). This diversity of formats requires different tools and processing strategies that hinders the process of data retrieving.

- **Massive volumes of data**. The volumes of data generated by each instrument depend on the characteristics and the operating mode of these sensors, among other factors, mainly:

  - **Temporal resolution** or frequency of observation of the same point in the surface of Earth: diary, each 8 days, each 16 days, etc. Furthermore, there are data product that summarize observable variables for a period of time: a month or a year, for example. The higher the temporal resolution the more data is produced.
  - **Spatial resolution** or the smallest object measured that can be distinguished. To continue with the MODIS example, it works with resolutions of 250m, 500m and 1000m. Considering the same temporal resolution, a data product from MODIS with the highest spatial resolution can produce around 30GB of data whilst with the lowest resolution generates about 700MB. The higher the resolution, the more data is produced.

- **Spectral resolution** or the number of bands captured for each sensor and its bandwidth. For instance, MODIS has 36 bands operating a different spatial resolution: 250 m bands 1-2 (250m), bands 3-7 (500m), bands 8-36 (1000m). The number of bands is also related with the number of data products that can be provided. MODIS, for example, has available seventy data products and a subset of only seven data products during a one year period was considered it would imply about 260000 files and 2.25 TB of data.

Hence, considering multiple satellite instruments and data products with both high spatial and temporal resolution might lead to massive volumes of data.

## 3.2 Climate and meteorological data sources

### 3.2.1 Public station networks

Networks of agro-meteorological stations are commonly used as sources of valuable data for farmers. In this regard, public administrations often own networks of agro-meteorological stations and publicly share their observations via Internet. Using this data from it is possible to develop, for example, disease early-warning systems (Neto et al., 2012). These systems often consider leaf moisture, air humidity, air temperature and time factors.

In particular for Spain there are several agro-meteorological networks with coverage in the main agricultural regions (see Fig.3.1). For example, the Sistema de Información Agroclimática para el Regadío (SIAR) (http://eportal.magrama.gob.es/websiar/Inicio.aspx) includes more than 350 stations in regions such as Andalucía, Canarias, Castilla-La Mancha, Castilla y León, Extremadura, Murcia and Comunidad Valenciana.

Other remarkable example is the Meteogalicia agro-meteorological network (http://www.meteogalicia.es/). This network covers the region of Galicia in the North-West of Spain and provides both historical and current data valuable for farmers as, for instance: hours of foliar humidity, soil temperature and wind speed. See Appendix B.1 for an example of the observations measured by agro-meteorological stations of Meteogalicia.

### 3.2.2 In-field private station networks

The density of public stations could be insufficient for PA purposes and frequently farmers have their own station network infrastructure. These agro-meteorological stations or motes, involve a cost of roughly 1000€ per station including the sensors. Hence, the study of how many stations are required for a agricultural exploitation is key for the viability of an PA project (McBratney et al., 2005). The location of the stations is also crucial as it is related with the spatial variability and the characteristics of the crop lands. Hence, optimizing the number of stations in an agricultural exploitation is a challenge problem both for economic and production purposes.

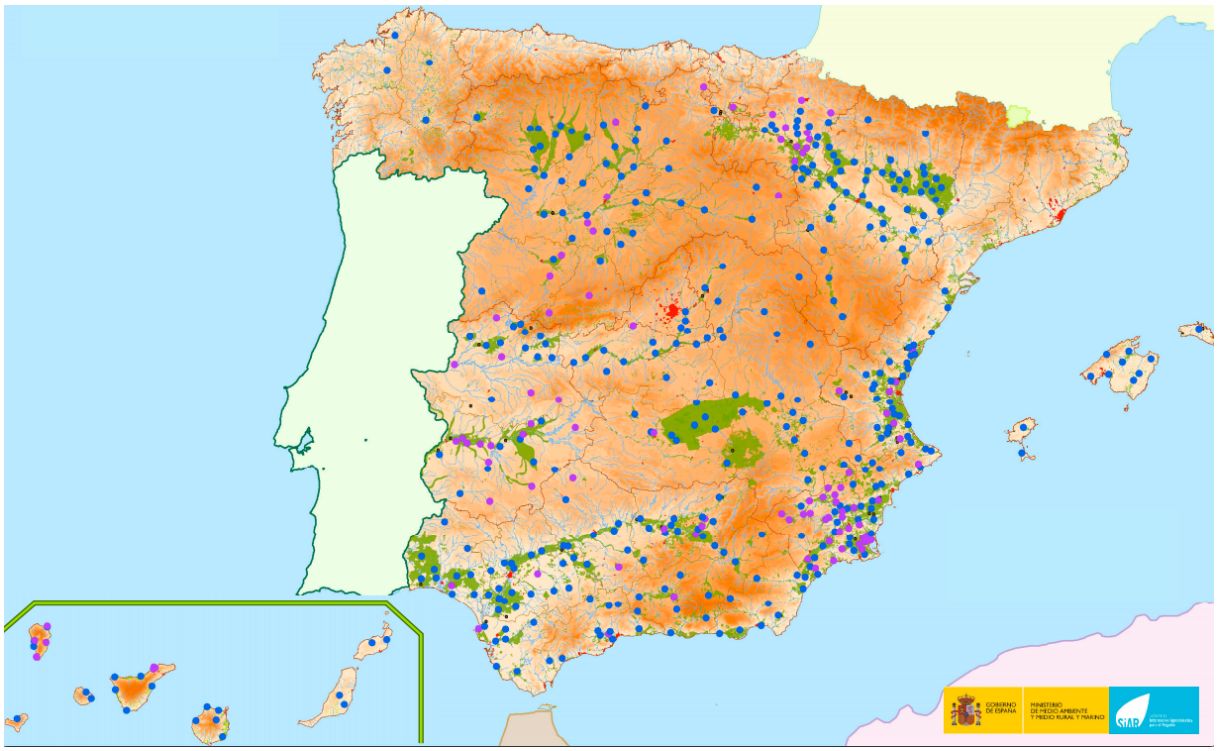


Figure 3.1: Location of the SIAR agro-meteorological network in Spain

The mote is a self contained system for the capture, processing and communication of agro-climate parameters. The mote often offers several ports to connect different sensors such as leaf wetness, leaf temperature, soil temperature and air temperature and humidity.

The system is solar powered and has an internal battery. To reduce power requirements, it is often used an strategy in which the microprocessor is in sleep mode and wakes up every certain number of seconds to take a measurement. Regarding communications, a typical configuration related with the Internet of Things (IoT) involves a M2M modem (3G) uploading the data to a server in the cloud using XML or GeoJSON files.

### 3.2.3 Meteorological radars

This kind of radars collects data about rain precipitation and its evolution measuring the raindrop size distribution and the radar reflectivity (see Fig.3.2). From these data, the meteorological radars calculate an indicator called Plan Position Indicator (PPI). According to Marshall-Palmer's formula (Marshall et al., 1947) (see Eq. 3.4), from PPI allows the following estimations:

- The type and intensity of the rainfall:
  - **Light rain**: intensity lower or equal than 2 mm/h.
  - **Moderate**: intensity greater than 2 mm/h and lower or equal than 15 mm/h.
  - **Heavy rain**: intensity greater than 15 mm/h and lower or equal than 30 mm/h.
  - **Very heavy rain**: intensity greater than 30 mm/hand lower or equal than 60 mm/h.

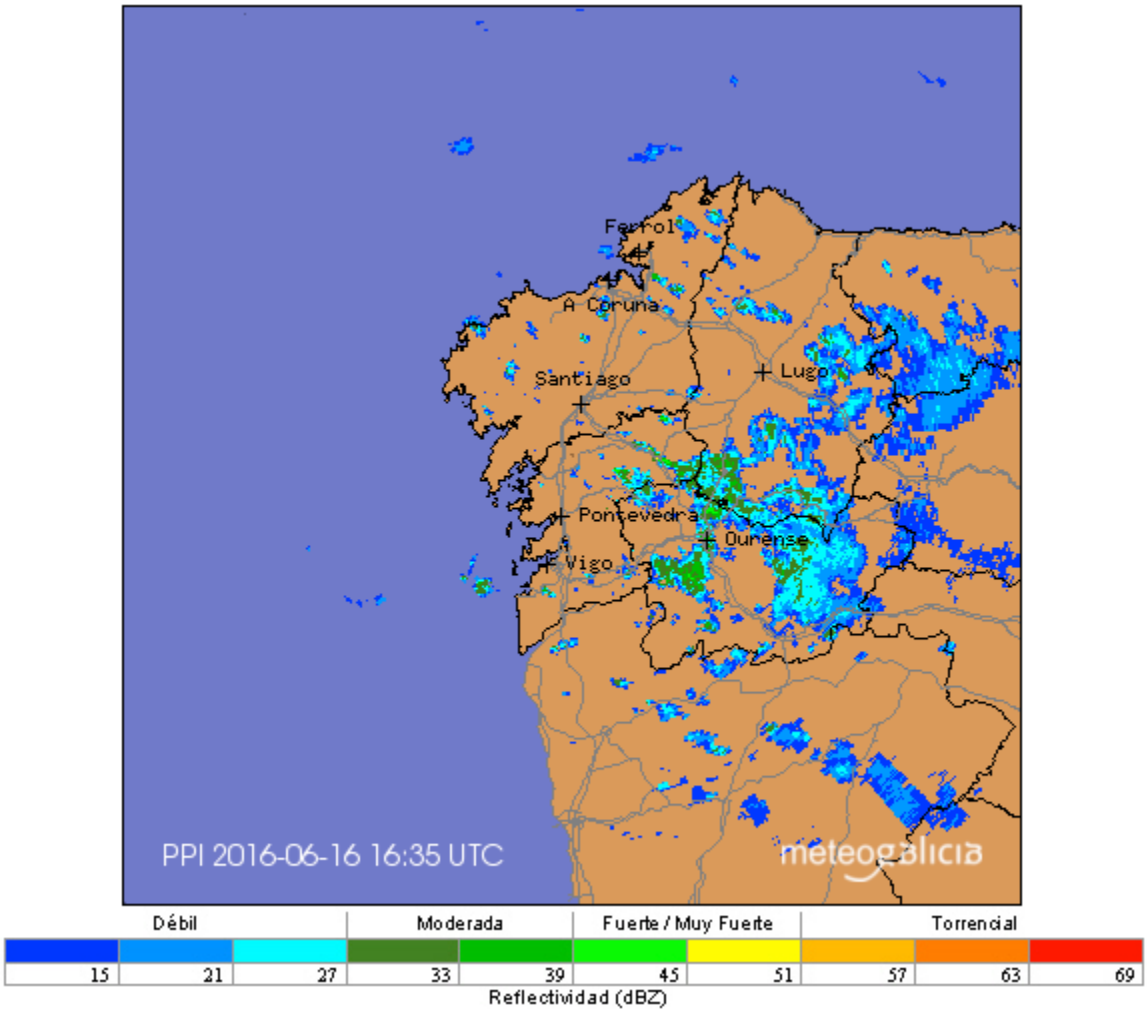



Figure 3.2: Radar image provided by Metogalicia for the Northwest region of Spain and Portugal

  - **Extreme rain**: intensity greater than 60 mm/h

- The rain accumulated in 6 hours.

$$Z = 200 \cdot R^1.6, \tag{3.4}$$

where $Z$ is the reflectivity measured by the radar expressed in $mm^6/m^3$ and $R$ is the precipitation rate expressed in $mm/hr$.

### 3.2.4 Weather forecast

Weather conditions represent an important factor for scheduling agricultural treatments and crop protection (Adams et al., 1995). Knowing in advance possible events of severe weather conditions such as floods, hail or strong wind gusts allow farmers to take preventive actions in order to minimize damage in crops. However, predicting regular weather conditions is also

important (Zhu et al., 2002). As an example, farmers could decide to apply a phytosanitary treatment earlier than planned because of a rainfall forecast, even if the impact of infection is low. Or they could delay spraying the fields because of the forecast of strong wind gust.

To make such predictions, the data from numeric forecast models such as the Weather Research Forecast (WRF) can be helpful (Cooter et al., 2012). This model provides the following information (with 1 Km of spatial resolution and 72 hours of temporal resolution):

- Sky state: sunny, high clouds, partly cloudy, overcast, cloudy, etc.
- Temperature
- Precipitation amount
- Wind direction
- Wind module
- Relative humidity
- Cloud area fraction
- Air pressure at sea level
- Snow level

### 3.2.5 Bioclimatic indices

Literature describes the use of viticultural climatic indices such as the Thermal Index of Winkle (Amerine and Winkler, 1944; Winkler, 1962) or the Heliothermal Index (Huglin, 1978), which are representative indicators of the variability of the viticultural regions worldwide (Tonietto and Carbonneau, 2004).

- Average temperature for the active period of vegetation:

$$GSTavg = \frac{\int_a^b T_{avg}}{n}, \tag{3.5}$$

  where a and b correspond to the March 1st and August 31th, respectively, $T_{avg}$ is the average of the daily temperature and $n$ is the number of days between $a$ and $b$.

- Average of maximum and minimum temperatures for the active period of vegetation:

$$GSTmax = \frac{\int_a^b T_{max}}{n} \tag{3.6}$$

$$GSTmin = \frac{\int_a^b T_{min}}{n}, \tag{3.7}$$

where a and b correspond to the March 1st and August 31th, respectively, $T_{max}$ is the daily maximum temperature, $T_{min}$ is the daily minimum temperature and $n$ is the number of days between $a$ and $b$.

- Frosts (FD) calculated as the number of days of the year with average temperature below 0°
- Number of days with maximum temperature above 25° and above 30° (ND25, ND30).
- Active Thermal Integral (Haba et al., 1997):

$$ATI = \int_a^b T_a, \tag{3.8}$$

where Ta is the active temperature, calculated by adding the daily mean temperatures above or equal to 10° and a and b correspond to the March 1st and August 31th, respectively.

- Thermal Index of Winkle is the sum of effective daily mean temperatures, calculated from the monthly average temperatures multiplied by days of each month during the growing season from April to October:

$$WI = \int_a^b T_e, \tag{3.9}$$

where the effective temperature $T_e$ is the active temperature $T_a$ minus 10° and $a$ and $b$ correspond to the March 1st and August 31th, respectively.

- Heliothermic product (Branas et al., 1946):

$$PH = 10^{-6} \cdot \int_a^b T_a \cdot \int_a^b H, \tag{3.10}$$

where $a$ and $b$ correspond to the March 1st and August 31th, respectively, $T_a$ is the active temperature and $H$ the daily hours of light.

- Heliothermic index (P) (Branas et al., 1946):

$$P = \int_a^b Tm_{avg} \cdot Rm_{acc}, \tag{3.11}$$

where $a$ and $b$ correspond to March and August, respectively, $Tm_{avg}$ is the monthly average temperature and $Rm_{acc}$ the monthly accumulate rainfall.

- Huglin index of helio-thermal aptitude (Huglin, 1978):

$$IH = \frac{\int_a^b ((T_a - 10°) + (T_m - 10°))K}{2}, \tag{3.12}$$

where $T_a$ is the daily average temperature, $T_m$ is the maximum daily temperature during the active period of vegetation, and $K$ is the length ratio of days varying from 1.02 to 1.06 between 40 and 50 degrees of latitude.

- Average temperature in April, May, June and July.
- Average rainfall (mm) in April, May, June and July.
- Total annual rainfall:

$$Pannual = \int_a^b r, \tag{3.13}$$

where $r$ is daily rainfall in mm and $a$ and $b$ correspond to the January 1st and December 31th, respectively.

- Annual rainfall for the active period of vegetation:

$$Pgs = \int_a^b r, \tag{3.14}$$

where $r$ is daily rainfall in mm and $a$ and $b$ correspond to the March 1st and August 31th, respectively.

- Maximum rainfall (Pmax) calculated as the maximum daily rainfall of the year.

## 3.3 Agricultural data sources

### 3.3.1 Soil analysis

The knowledge about soil characterization by means of electrical conductivity and soil analysis could enhance decisions regarding how to delimit the MZs (Franzen et al., 2002) (see Section 1.1.4). On the other hand, the estimation of soil nutrients from geo-located soil samples could enhance decisions regarding to the fertilization plan (Cambardella and Karlen, 1999) considering the levels of parameters such as:

- pH-$H_2O$. pH in water.
- pH-KCl. pH in potassium chloride.
- % Organic matter.
- P-content. Phosphor content.
- K-content. Potassium content.
- Exchangeable Mg. Magnesium ions.
- Electric conductivity. It is a measure of soil salinity (Corwin and Lesch, 2005)

### 3.3.2 Crop treatments

The geo-location of all the agricultural interventions that involves the crop management could be valuable for taking site-specific decisions. However, it requires some kind of registration process, manual or automatic by means of the signals emitted by the agricultural vehicles whilst applying the treatment to the crops.

As an example, the following data could be gathered regarding the treatment application:

- **Management zone**. The ID of the MZ where the treatment is applied.
- **Treatment plan ID**. The treatment plan related with the intervention.
- **Date/hour**. The date and hour of the treatment.
- **Litres**. The amount of litres used in the treatment.
- **Machinery**. The reference to the machinery used in the treatment (e.g., tractor and sprayer)
- **Pressure**. The pressure in bars used.
- **Speed**. The speed of the vehicle.
- **Number of nozzles**. The number of nozzles used by the machinery.
- **Notes**. Annotations about the intervention.
- **Geometry**. The geometry corresponding to the specific area of the intervention if it is not the same as the whole MZ.

### 3.3.3 Crop yield

Crop production is the outcome as a result of the agriculture labours, soil characterization, weather conditions and other factors. Regarding PA, crop yield might be considered in the followings tasks:

- **MZ identification**. For example, by means of yield maps and clustering those zones with similar rate production per unit of surface.
- **Annual fertilization plan**. The annual plan of fertilization for next campaign might consider, among other factors, yield production and soil analysis.
- **Yield forecasting**. Historical yield data combined with other data sources as, for instance, vegetation indices and weather variables, might be used for yield crop forecasting. However, the yield could present high variability between regions and between years and this fact increases the difficulties of this task. As an example, between the years 2011 and 2012, the Designation of Origin Rias Baixas had a variation of production of a 42.9%.

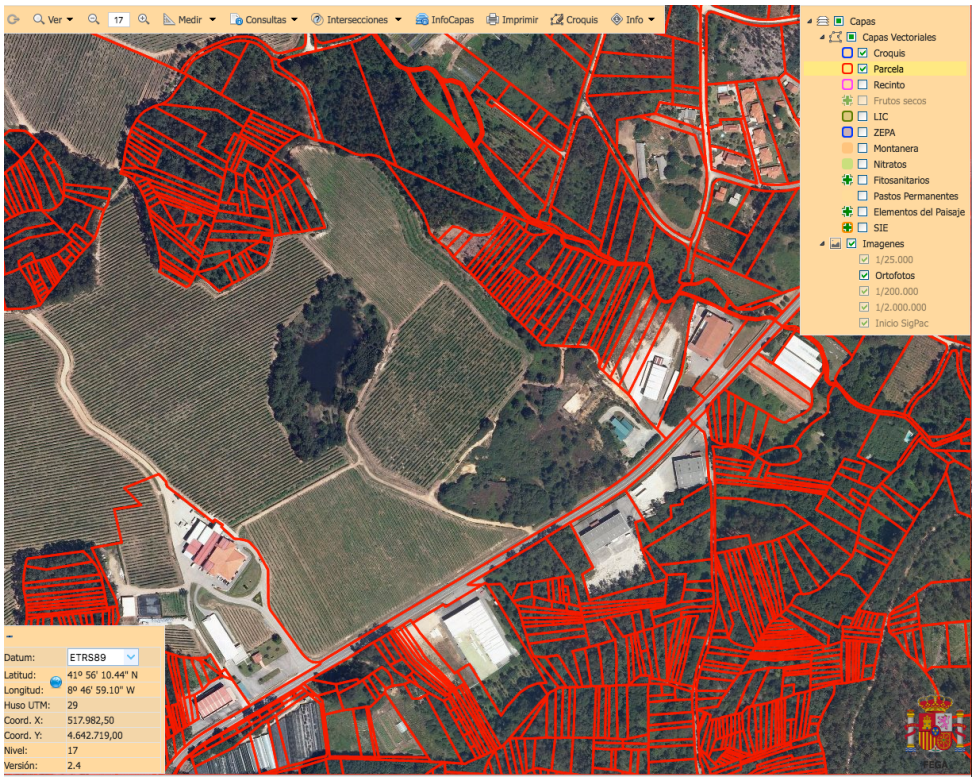



Figure 3.3: Example of use of SIGPAC to find the boundaries of parcels

### 3.3.4 Land parcel identification systems (LPIS)

As a mechanism to register and consult the census of the land parcels. Generally, LPIS offers public Web Mapping Services (WMS) for retrieving the maps providing the field parcel boundary lines, ownership and land use of farm areas. In Spain, the Ministry of Agriculture, Food and Environment (MAGRAMA) provides a GIS application called SIGPAC (see Fig.3.3) which provides this kind of information (http://sigpac.mapa.es/fega/visor/) However, the deployment of these measurement systems is not essential for PA.

### 3.3.5 Phenological stages

Phenological growth stages of plants describe their developmental stages. In particular, for grapevines there are some methods for the representation of those stages such as Baggliolini (1952); Eichhorn et al. (1977). Table 3.2 includes the major phenological stages of grapevine (Boso et al., 2008) following the method of Eichhorn et al. (1977) and revised from Coombe (1995).

| Stage | Description |
|---|---|
| 04 | Bud burst. Leaf tips visible |
| 12 | Leaves separated. Shoots about 10 cm long; inflorescence clear |
| 15 | Inflorescence elongating. Flowers closely pressed together |
| 19 | Flowering begins. About 16 leaves separated; beginning of flowering (first flower caps loosening) |
| 23 | Full flowering. 17-20 leaves separated; 50% caps off |
| 27 | Setting. Young berries enlarging (> 2 mm diam.), bunch at right angles to stem |
| 31 | Berries pea-sized. Bunches hang |
| 32 | Beginning of bunch closure, berries touching (if bunches are tight) |
| 34 | Berries begin to soften. Sugar starts increasing |
| 35 | Beginning of berry ripening. Beginning of loss of green colour *veraison* |
| 38 | Berries ripe for harvest |

Table 3.2: Representation of the major phenological stages of grapevine (Boso et al., 2008) following the method of Eichhorn et al. (1977) and revised from Coombe (1995)

### 3.3.6 Vegetation and moisture indices

Sensors and on-board satellite instruments related to environmental and Earth observation, measure electromagnetic radiation emitted and reflected by observed objects. In fact, plants absorb and reflect the solar radiation following an unique pattern to them according to their cell structures, leaf and surface widths, amount of water in their bodies and their positions in their natural environments (see Fig.3.4). In particular, plants absorb light in the wavelength of 0.4-0.7 $\mu$m which is called the visible light region (see Fig.3.5), absorbing a small region of the infrared spectrum and reflecting almost the whole of this spectrum.

Taking into account these patterns of reflectivity, some authors propose the calculation of indices based on arithmetic combinations of spectral bands for monitoring health, moisture and water stress of the crops (Jackson and Huete, 1991). Hence, from such radiometric data it is possible to obtain valuable variables and indicators from the agriculture perspective Villajos et al. (2008) such as moisture, soil temperature and vegetation health.

The potential of these indices for agriculture is very high (Basso et al., 2004). Such indices can be categorized as follows based on the potential for the estimation of vegetation health and moisture conditions:

#### Vegetation indices

Vegetation indices are based on chlorophyll spectral reflectance (Jordan, 1969) and specifically, reflectance differences of the chlorophyll on the Red and Near infrared (NIR) spectrum. This last one is divided on two: Short Wavelength Infrared (SWIR1) from 1.560nm to 1,660nm and Short Wavelength Infrared 2 (SWIR2) from 2,100nm to 2,300nm. There are numerous vegetation indices. The following ones are probably among the most known:

- **Ratio Vegetation Index (RVI)** (Jordan, 1969). It is probably one of the first index to be defined (Jackson and Huete, 1991) and is an indicator of density and vigour in green vegetation:

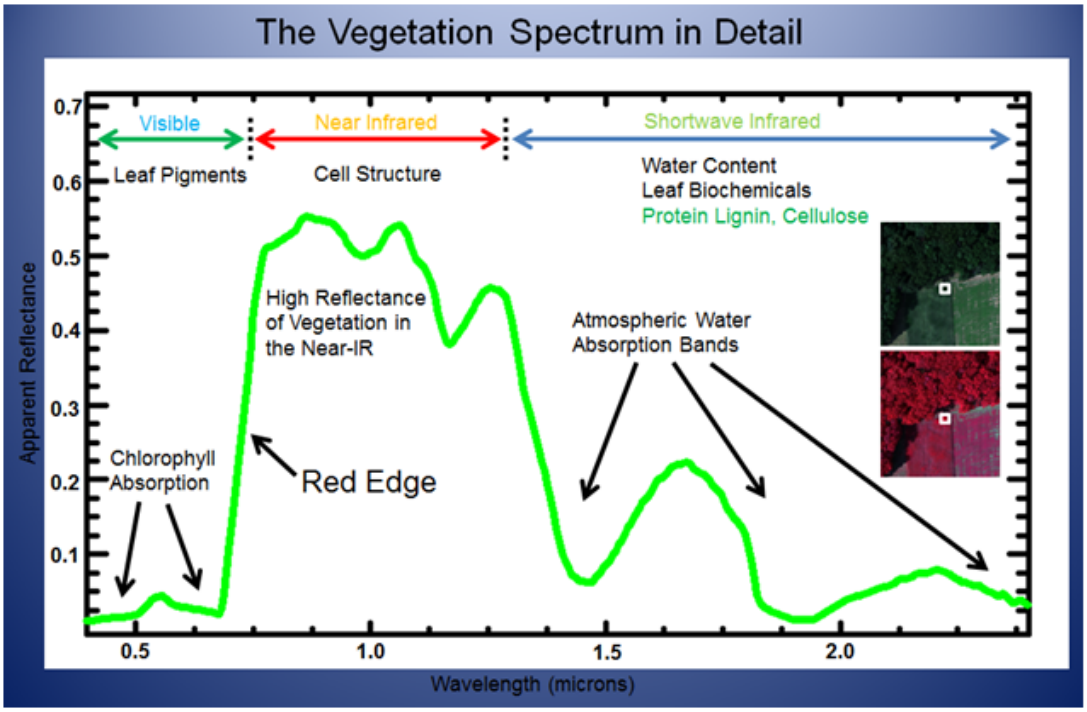



Figure 3.4: Electromagnetic spectrum related with the cell structure of vegetation and with their water content. Source: Elowitz, Mark R. (www.markelowitz.com/Hyperspectral.html)

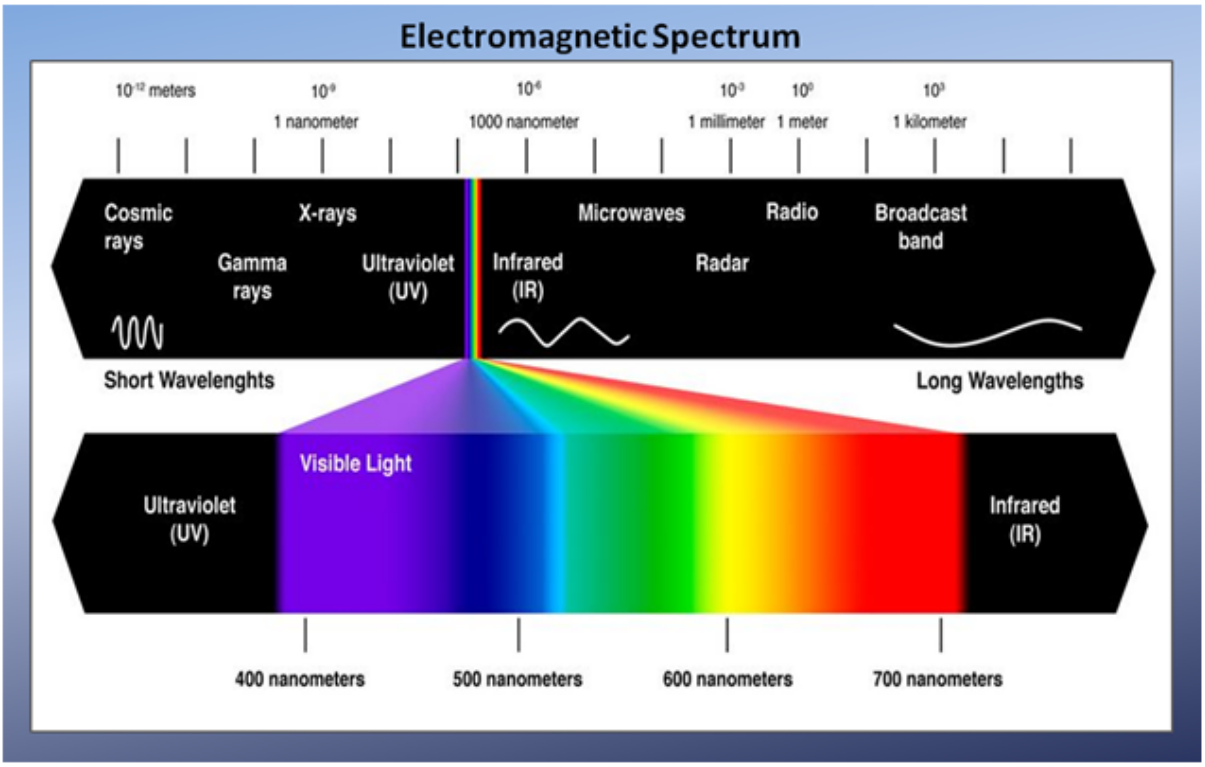



Figure 3.5: The Electromagnetic Spectrum. Source: Zami, Zuly. (www.zulyzami.com/The+Electromagnetic+Spectrum)

$$RVI = \frac{NIR}{Red} \tag{3.15}$$

- **Normalized Difference Vegetation Index (NDVI)** (Kriegler et al., 1969; Rouse Jr et al., 1974). Although it is similar to RVI, this index is more sensitive to sparse vegetation than RVI (Jackson and Huete, 1991). Negative values correspond to water, clouds or snow since their reflectance in the visible spectrum is greater than the corresponding in near infrared, whilst soil and rocks have values near zero. Ranges between 0.1 and 0.6 are indicator of vegetation. Values above 0.6 correspond to dense vegetation canopy:

$$NDVI = \frac{NIR - Red}{NIR + Red} \tag{3.16}$$

- **Enhanced Vegetation Index (EVI)** (Qing Liu and Huete, 1995). It is an indicator more responsive to canopy structural variations, including canopy type, plant physiognomy, and canopy architecture:

$$EVI = G\ \frac{(NIR - Red)}{NIR + (C1 \cdot Red - C2 \cdot Blue) + L}, \tag{3.17}$$

  where $G = 2.5$, $C1 = 6.0$, $C2 = 7.5$ and $L = 1$ (Huete et al., 1997).

- **Global Environmental Monitoring Index (GEMI)** (Pinty and Verstraete, 1992). It is related with presence or absence of vegetation. It takes values between 0 and 1 and is less sensible to atmospheric conditions than NDVI. However, it is not suitable to poor vegetation zones:

$$GEMI = eta\ (1 - 0.25 \cdot eta) - \frac{Red - 0.125}{1 - Red}, \tag{3.18}$$

  where

$$eta = \frac{2(NIR^2 - Red^2) + 1.5 \cdot NIR + 0.5 \cdot Red}{NIR + Red + 0.5} \tag{3.19}$$

- **Vegetation Condition Index (VCI)** (Kogan, 1990). This normalized index is used as an impact indicator of climate conditions on vegetation (Kogan et al., 2003). The values of the index ranges from 0 (extremely poor condition) to 100 (excellent):

$$VCI = 100 \frac{NDVI - NDVI_{min}}{NDVI_{max} - NDVI_{min}} \tag{3.20}$$

- **Leaf Area Index (LAI)** (Chen and Black, 1992). It is related to plant canopies. Considering the leaf morphology, authors propose different definitions. Chen and Black (1992) defines LAI as the ratio of total vegetation canopy divided by the surface area of the land on which the vegetation grows and Myneni et al. (2002) as the projected needle leaf area in

coniferous canopies. There are wide range of methods to obtain this index. For instance, MODIS uses an algorithm (Knyazikhin et al., 1998) based on canopy reflectance.

**Moisture and thermal indices**

These indices are indicators of canopy water content calculated mainly from NIR and SWIR spectrum. The following vegetation indices are probably among the most known:

- **Normalized Difference Water Index 7 (NDI7)** (Rubio et al., 2006; Trombetti et al., 2008):

$$NDI7 = \frac{(NIR - SWIR2)}{(NIR + SWIR2)} \tag{3.21}$$

- **Shortwave Infrared Water Stress Index (SIWSI)** (Fensholt and Sandholt, 2003):

$$SIWSI = \frac{(NIR - SWIR1)}{(NIR + SWIR1)} \tag{3.22}$$

- **Shortwave Infrared Ratio (SWIRR)** (Trombetti et al., 2008):

$$SWIRR = \frac{SWIR1}{SWIR2} \tag{3.23}$$

- **Moisture Stress Index (MSI)** (Hunt Jr and Rock, 1989):

$$MSI = \frac{SWIR1}{NIR} \tag{3.24}$$

- **Moisture Stress Index 7 (MSI7)** (Trombetti et al., 2008):

$$MSI7 = \frac{SWIR2}{NIR} \tag{3.25}$$

- **Global Vegetation Moisture Index (GVMI)** (Ceccato et al., 2002):

$$GVMI = \frac{(NIR + 1) - (SWIR1 + 0.02)}{(NIR + 1) + (SWIR1 + 0.02)} \tag{3.26}$$

# Chapter 4

# IMPACT OF TEMPORAL RESOLUTION FOR AUTOMATIC LAND DELIMITATION

One of the first steps in the application of PA techniques for a particular geographical zone is the task of land delimitation, that is, to identify which regions share soil properties and can be treated in the same way. In particular, automatic land delimitation is focused on providing delimitations using different data sources as those from satellite or sensors (Arango et al., 2015).

Satellite data products are characterized by the spatial resolution and the temporal resolution. The last one corresponds to the revisit time and, generally, the higher the spatial resolution, the lower the temporal one. So there is a trade off between both resolutions.

Hence, taking into account that PA requires high spatial resolution, this chapter studies how the temporal resolution affects the results of the automatic delimitation of land using clustering algorithms and satellite reflectivity as input. In order to check this delimitation, two different clustering paradigms are applied to data collected from Terras Gauda vineyards.

The results are promising in the sense that obtained clusters are consistent to the current land organization. In addition, it is shown that the lower temporal resolution, the more compact the clusters. The Chapter is organized as follows. Section 4.1 explains the satellite data products used in the clustering process. Section 4.2 enumerates the proposed clustering algorithm. Finally, Section 4.3 shows the main conclusions obtained about time-series reflectance.

## 4.1 Input data

Valuable data related to PA such as vegetation indices, land surface temperature or surface reflectance are collected by MODIS. Around 70 data products are provided by this instrument operated from TERRA and AQUA satellites (see Section 3.1.1). These data products are publicly available by HTTP, FTP and at the NASA Land Processes Distributed Active Archive

Center (see Appendix A.1.1).

In order to generate the data sets for the MZ identification process, daily surface reflectance data products were considered. Specifically MOD09GQ at 250m of spatial resolution (see Section A.1). It includes data about surface reflectance for spectral bands 1 (B1) and 2 (B2) and other variables to measure the quality of the observations and their coverage.

The general procedure followed to obtain the data is:

- Identify the spatial coordinates for the geographic zones of interest, the parcels of Terras Gauda. This task was done taking as a input a shape file with the geometry of all the polygons of the parcels of Terras Gauda, as registered on the Sistema de Información Geográfica de Parcelas Agrícolas (SIGPAC) (http://sigpac.mapa.es/fega/visor/)

- Download the MOD09GQ data products corresponding to the identified extension between January and May of 2014, both inclusive. The installation of the Geospatial Data Abstraction Library (Bivand et al., 2013) and MODIS Reprojection Tool (Dwyer and Schmidt, 2006) was required.

- Obtain the reflectance measures and other quality variables from the data product and for the Terras Gauda's vineyard.

- Calculate NDVI and NDVI scaled using the reflectance values (see Section A.1).

- Identify the MODIS GeoTIFF pixels belonging to Terras Gauda parcels, using the overlay geospatial operation with the shape file and the GeoTIFF.

- Generate files with the geolocated reflectivity values for the MOD09GQ data product (see table 4.1) and its calculated values. These files were used as input for the clustering algorithms in order to identify MZ.

| Column | Description |
|---|---|
| x | Coordinate x of the data point in the UTM 29 CRS |
| y | Coordinate y of the data point in the UTM 29 CRS |
| date | Year + Day number of the year in the format YYYYddd |
| refl_b01 | Reflectivity values from MOD09GQ band 1 |
| refl_b02 | Reflectivity values from MOD09GQ band 2 |
| num_observations | The number of observations |
| QC_250m | A byte about the quality of the measure |
| NDVI | Normalized Difference Vegetation Index |
| NDVI_scaled | Scaled NDVI |

Table 4.1: Columns of the dataset for the MOD09GQ data product. Spatial resolution: 250 m. Temporal resolution: daily. CRS: UTM 29

## 4.2 Clustering algorithms

As it was described in Section 4.1, the data used in this approach include information about surface reflectance for spectral bands 1 and 2 with a spatial resolution of 250m obtained from MODIS and the NVDI and NVDI-scaled indices.

To obtain the automatic delimitation of the vineyard zones two different approaches were used, The first one is based on a partition clustering algorithm called PAM. The second one is based on hierarchical clustering paradigm. Both algorithms and their characteristics are explained on Section 2.1.3.

Regarding PAM, in order to properly select the best clustering distribution and since there is no reference to external information, the method selected to validate the clustering was the Silhouette coefficient (see Section 2.1.4). Therefore, to select the optimum $k$ according to that method, is followed the procedure described in Algorithm 2.2 of Section 2.1.4. Once $k$ is computed, the cluster algorithm is executed and the cluster assignment is retrieved, providing the land delimitation.

The other approach followed is based on hierarchical methods. Among all the hierarchical algorithms, *pvclust* algorithm (Suzuki and Shimodaira, 2006) was selected because it is a combination of the standard hierarchical algorithm *hclust* (Mahdavi et al., 2008) and bootstrap resampling (see Section 2.1.6), so the groups it obtains have higher confidence. According to the semantic of the problem being solved here, the most adequate linkage method is Ward's minimum variance (see hierarchical clustering, Section 2.1.3). The output of this method is a dendrogram, showing not only the groups but also the strength of the connection between them. However, it is not clear how strongly a cluster is supported by the data. To check the certainty of the existence of a cluster, the approximately unbiased probability values ($p$-values, Westfall and Young (1993)) are computed with multiscale bootstrap resampling. If the $p$-value is less than a certain threshold (usually a number close to 1), the cluster is rejected (Suzuki and Shimodaira, 2013).

## 4.3 Evaluation of time-series of reflectance

The Terras Gauda vineyard is divided into three separated parcels (hereinafter called $p_1$, $p_2$, $p_3$). According to the spatial resolution provided by the MODIS satellite (250m.), $p_1$ and $p_2$ are represented using 30 points and $p_3$ by 16 points. Each point is characterised by 4 variables per day (surface reflectance for spectral bands 1 and 2, NVDI and NVDI-scaled indices). As data were extracted for 90 days, each point $x$ is represented by 360 values.

The three parcels were clustered using the Manhattan distance (see Table 2.1) as dissimilarity metric. Other metrics where tested however the best result were obtained with this the Manhattan distance.

The performance of the values obtained from the MODIS satellite was tested for the automatic land delimitation and also for the effect of different temporal resolutions. To test the temporal resolution, all the attributes every day, every 2 days, ... until every 10 days, were

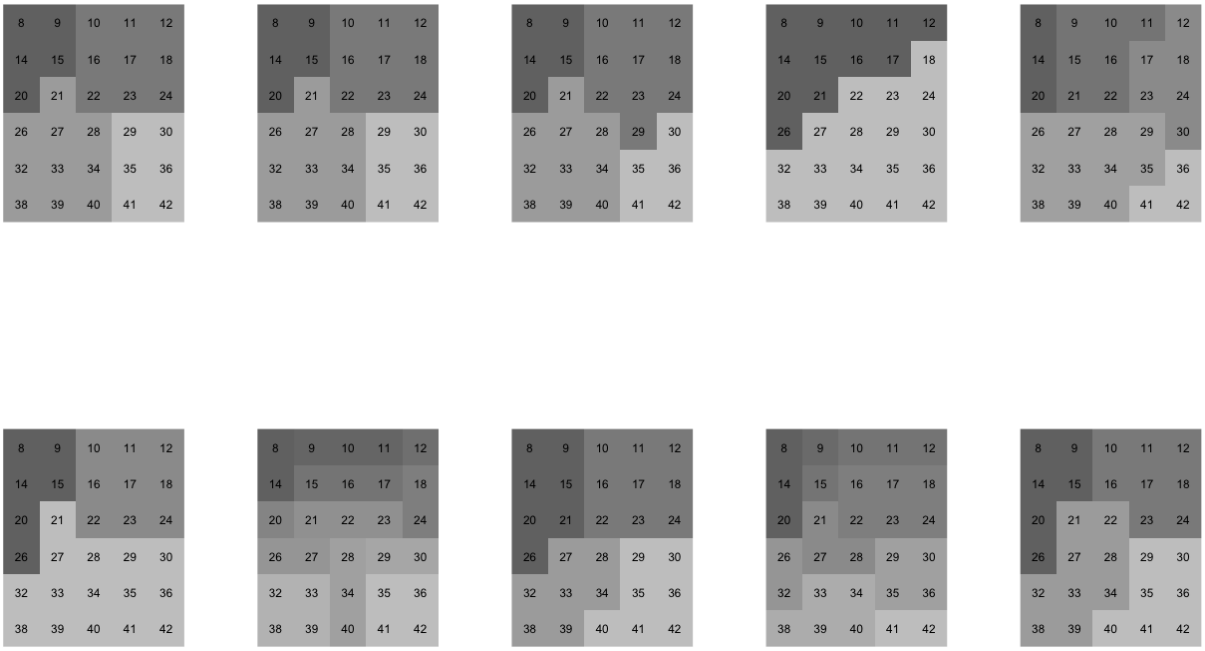


Figure 4.1: Land delimitation for parcel $p_1$.

considered.

To test the effect of the four attributes, the clustering procedure was performed considering each possible combination of the four attributes obtained per day. However, since the results are similar, only the clusterings obtained when the four attributes afore defined are taken together will be considered in what follows.

Figures 4.1 to 4.3 show the land delimitation for parcels $p_1$, $p_2$ and $p_3$, respectively. The structure of each figure is the following: It contains 10 squares divided into smaller squares according to the the spatial resolution provided by MODIS. Considering each figure as a matrix with 2 rows and 5 columns, the square at position $[1, j]$ contains the clusters obtained when temporal resolution is $j$. The square at position $[2, j]$ contains the clusters obtained when temporal resolution is $j + 5$. Therefore the top-left square represents clustering results for daily resolution and the bottom-right square represents clustering results for a ten days resolution.

The question now is how to decide which land delimitation is the best. As it was stated in Section 4.2, the criterion used was the Silhouette coefficient. $X$-axis of Figure 4.4 represents the temporal resolution (from 1 day to 10 days) while $Y$-axis represents the value of the Silhouette coefficient. As it can be seen in Figure 4.4 the lower the temporal resolution, the higher the Silhouette coefficient. Therefore, the main result we can extract is that there is no need to include daily information in order to keep the performance of the clusters. In fact, the clusters are more compact when the temporal resolution is lower.

### 4.3.1 Clustering aggregation

In order to obtain a better land delimitation from the ones obtained in the previous subsection with different temporal resolutions, it was been explored the possibility of aggregating the dif-

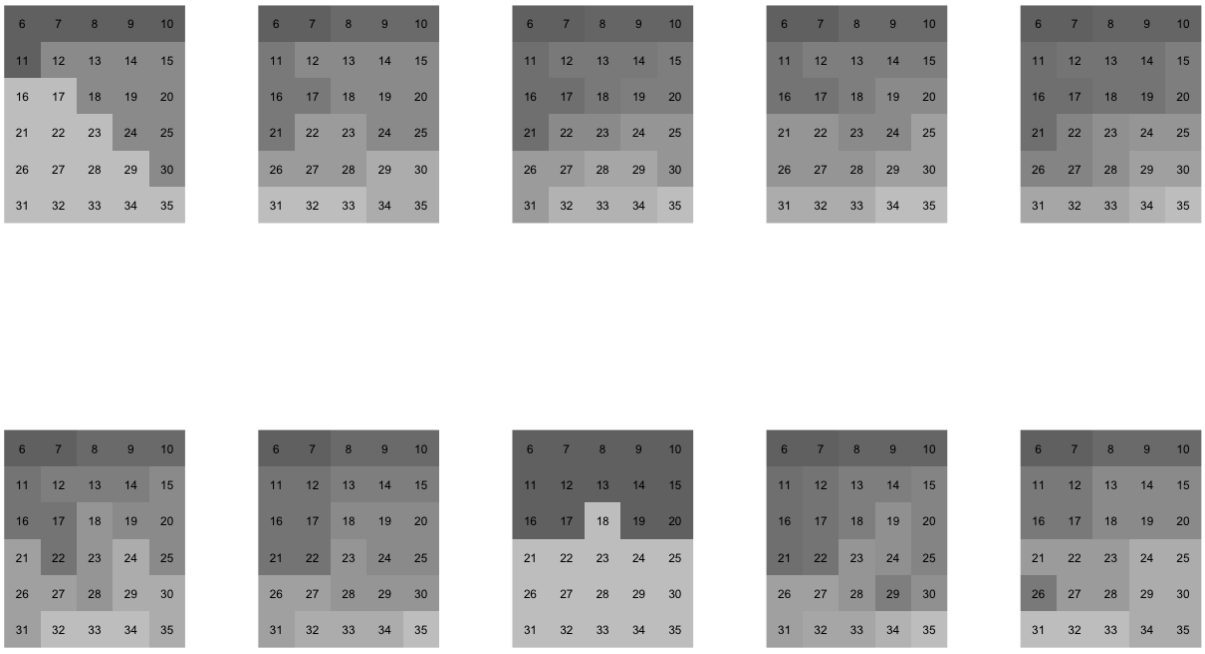


Figure 4.2: Land delimitation for parcel $p_2$.

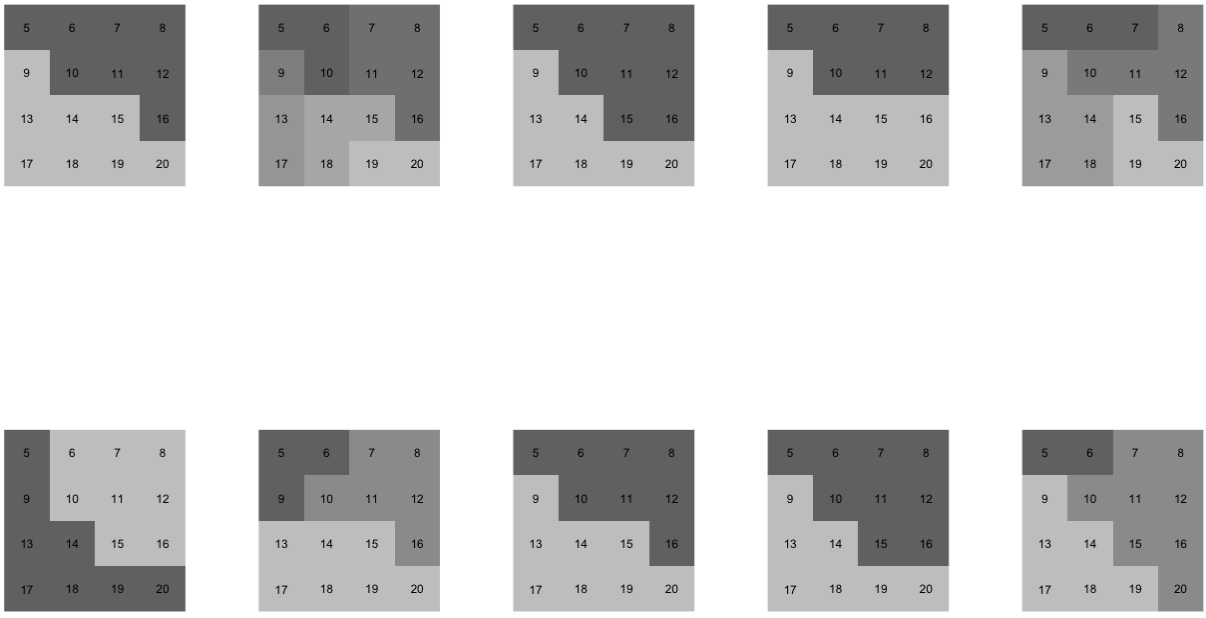


Figure 4.3: Land delimitation for parcel $p_3$.

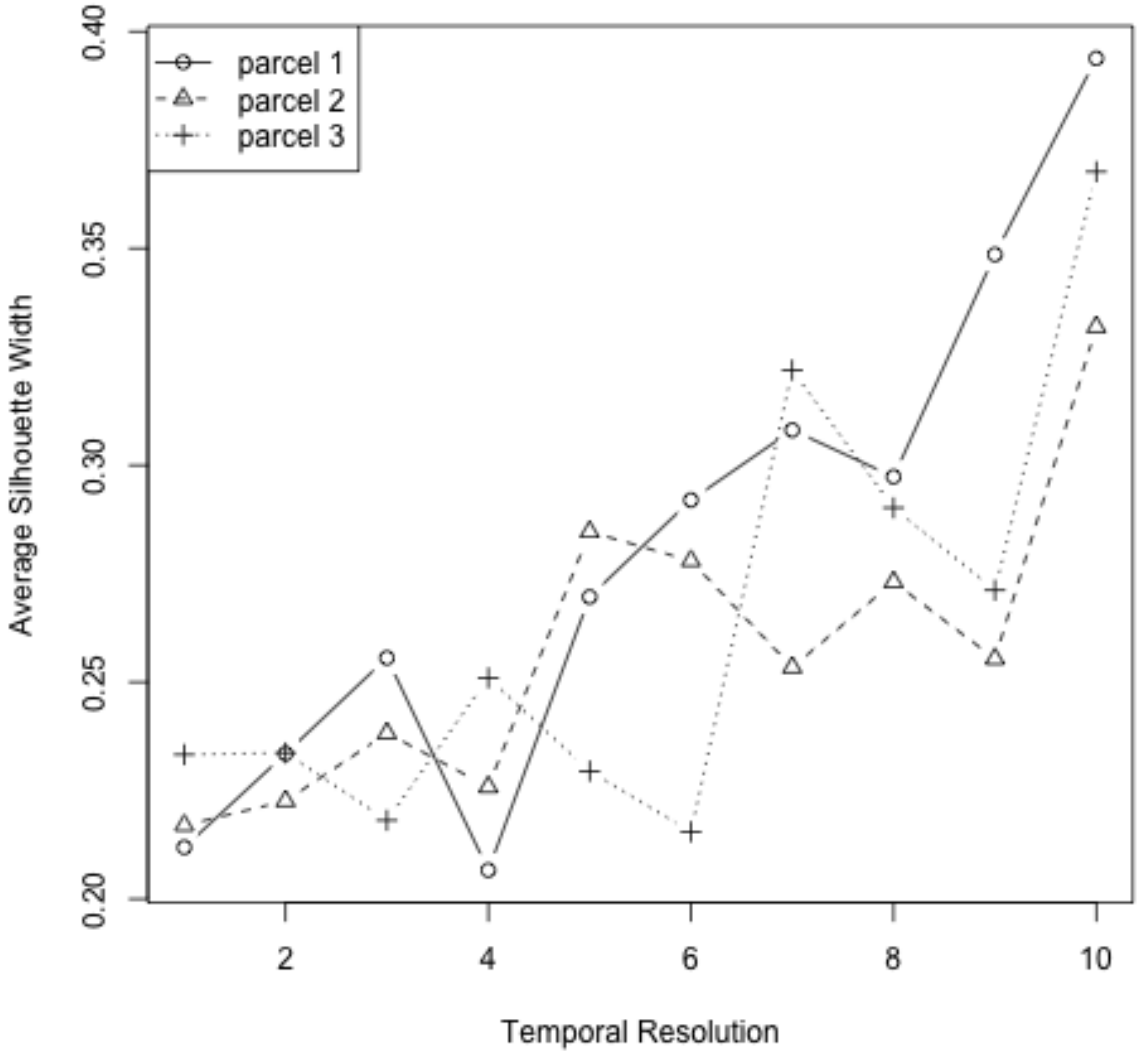



Figure 4.4: Silhouette coefficient obtained by the clustering algorithm

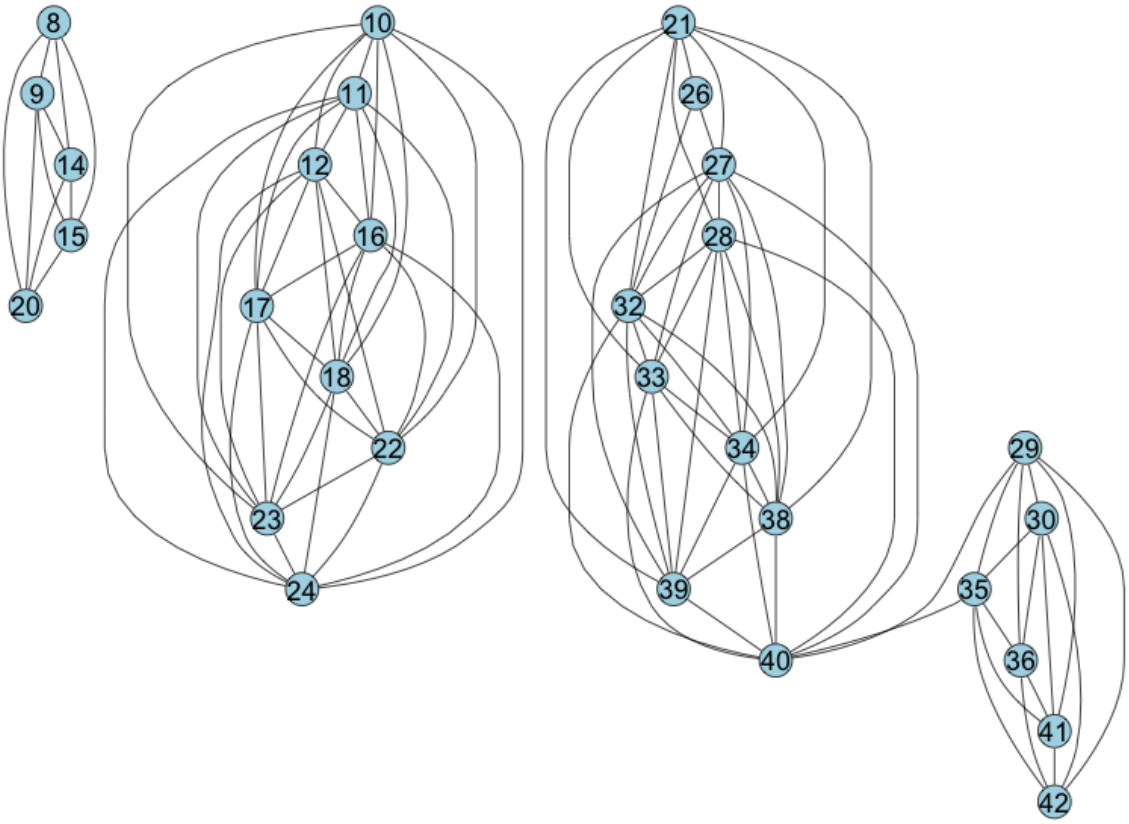



Figure 4.5: Connected components for parcel $p_1$ with $h = 5$

ferent clusterings. To this extent, it was considered a series of binary relations $R_h$ defined as follows: two points $a_1$ and $a_2$ are related by $R_h$ (in symbols, $a_1 R_h a_2$) if and only if they appear together in at least $h$ of the clusterings considered in the previous subsection.

Thus, there are ten different binary relations $R_1, R_2, \ldots, R_{10}$ each one contained in the previous one (i.e., if $i < j$ then $a_1 R_j a_2$ implies $a_1 R_i a_2$). It was considered, then, the graph representation of these binary relations and calculate their connected components. Each of these connected components contains points that are usually clustered together (with a higher probability the higher the value of $h$ is) and, in consequence, may be considered as a new, aggregated clustering.

Figures 4.5 to 4.7 show some of these graphs for parcels $p_1$ to $p_3$. In the case of the two first parcels we have set $h = 5$ while $h$ is 6 for the third case.

As can be seen, in these cases 3 areas were obtained for parcels $p_1$ and $p_2$ and 2 for $p_3$ which may be considered as *consensus* areas between all the clusterings (cf. Figures 4.1 to 4.3).

### 4.3.2 Using hierarchical clustering

In this section it is studied the problem of land delimitation using a hierarchical clustering method. As in the previous case, it just shown the results obtained when using the four input

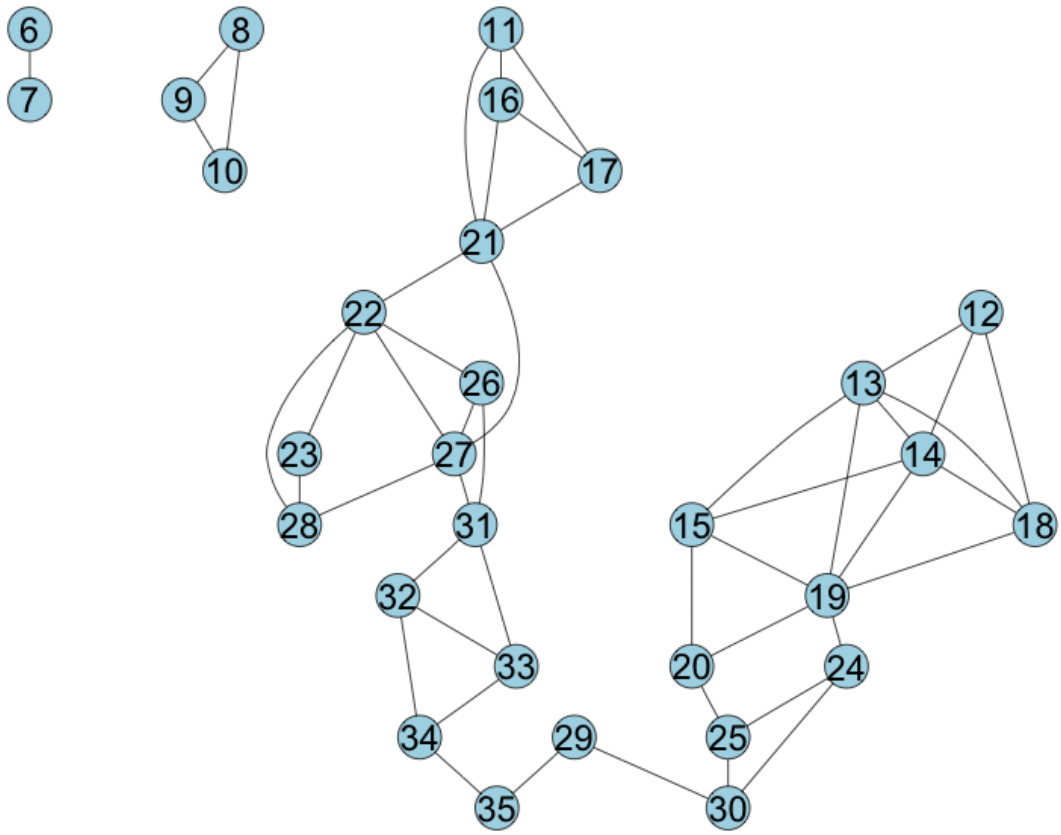



Figure 4.6: Connected components for parcel $p_2$ with $h = 5$

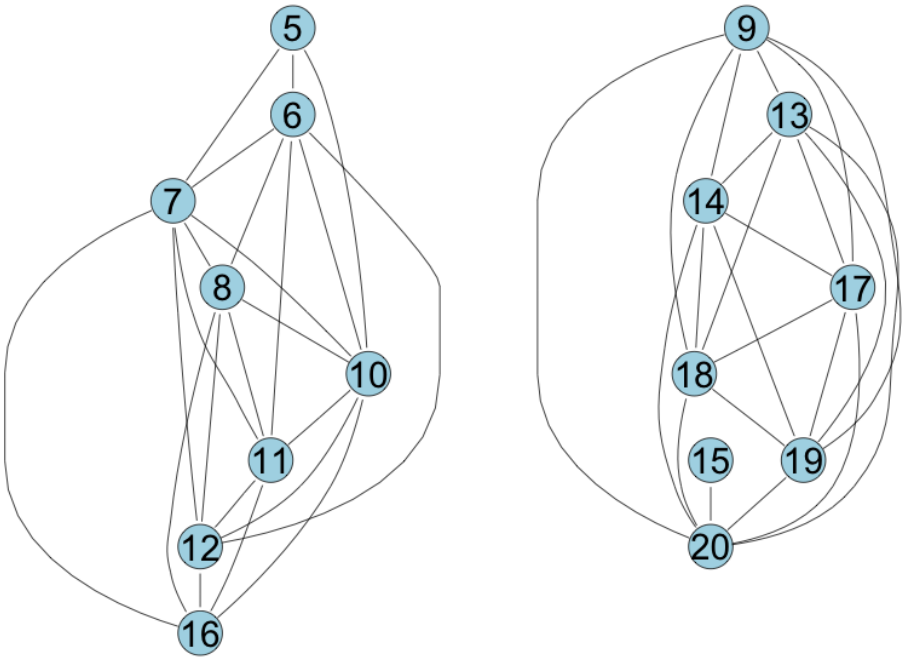



Figure 4.7: Connected components for parcel $p_3$ with $h = 6$

variables, testing the influence of temporal resolution.

As it was detailed in Section 4.2, this method provides a parameter $p$ to check the certainty of the existence of a cluster. Therefore, this parameter is used to select the best temporal resolution. Figures 4.8 to 4.10 represent $p$-values against standard error for the best temporal resolution for parcels $p_1$, $p_2$ and $p_3$ respectively.

Note that the close $p$-values to 1, the better. In addition, the lower the standard error, the better. According to these premises, the best resolutions are 9 for parcel $p_1$, 7 for parcel $p_2$ and 10 for parcel $p_3$. Figures 4.11 to 4.13 show the dendrograms associated to the selected resolutions obtained by *pvclust* for the three parcels. Groups with a certainty above 90% are highlighted. It is important to note that in general the lower the temporal resolution, the higher the $p$-values. This fact allows to conclude that when temporal resolution is lower, the performance of the clusterings is at least the same that the one obtained with all data. Therefore, there is no need to use daily data.

With regard to the groups themselves, note that for parcel $p_1$ the system is not able to group with a high confidence only four points (9, 12, 20 and 21) , five for parcel $p_2$ and two for parcel $p_3$.

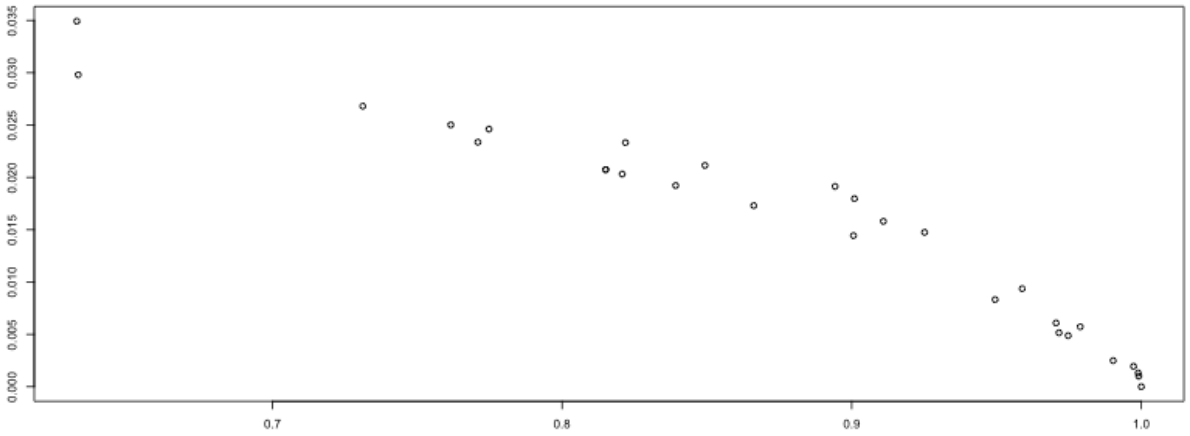



Figure 4.8: $p$-value against standard error for parcel $p_1$ and data obtained each 9 days

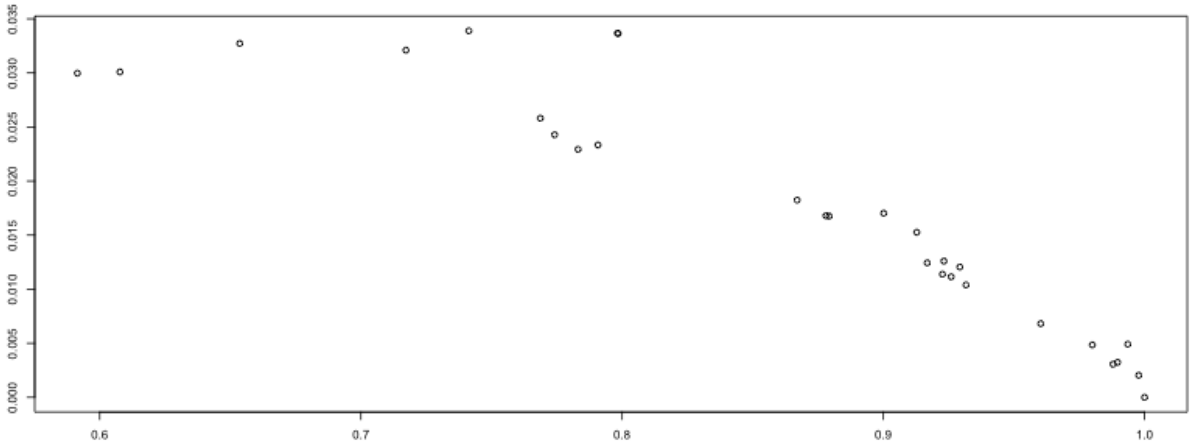



Figure 4.9: $p$-value against standard error for parcel $p_2$ and data obtained each 7 days

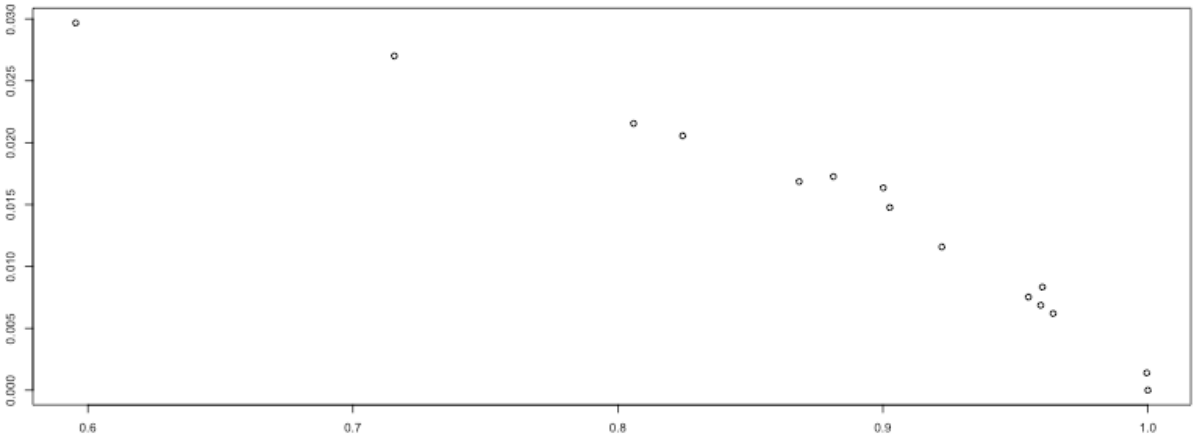



Figure 4.10: $p$-value against standard error for parcel $p_3$ and data obtained each 10 days

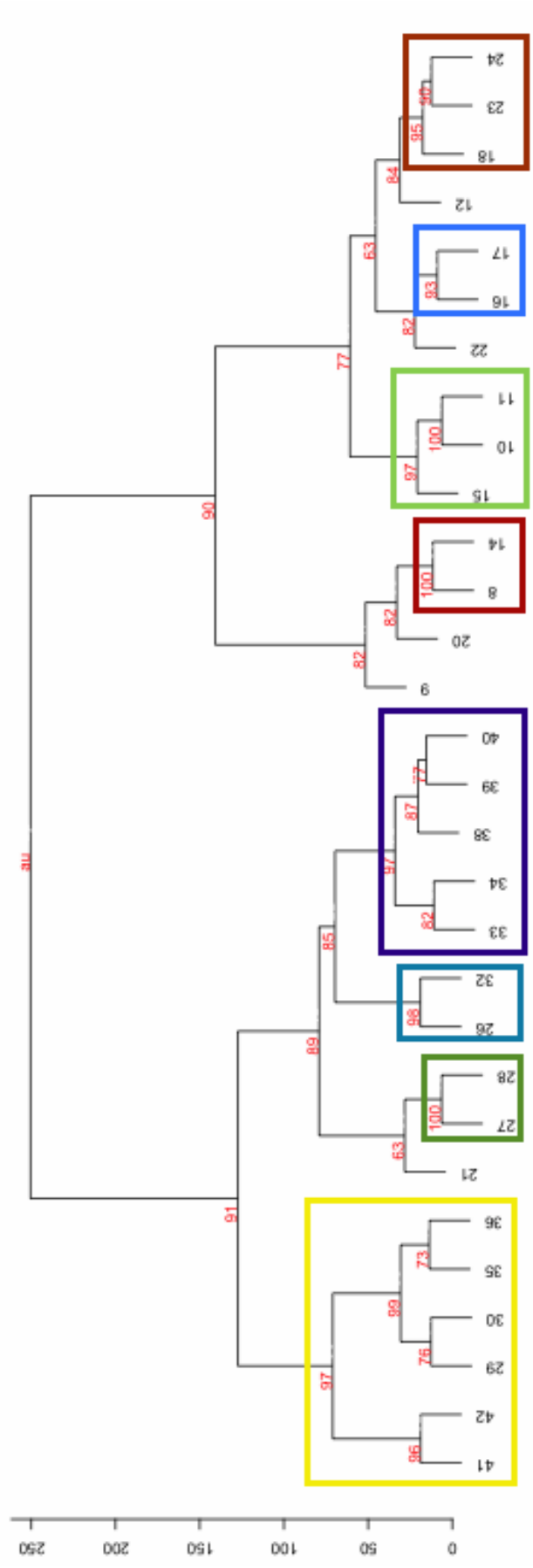



Figure 4.11: Dendrogram representing parcel $p_1$ clustering. Approximately unbiased $p$-values are presented at branch connections as percentages, and cluster labels are presented below the branches.

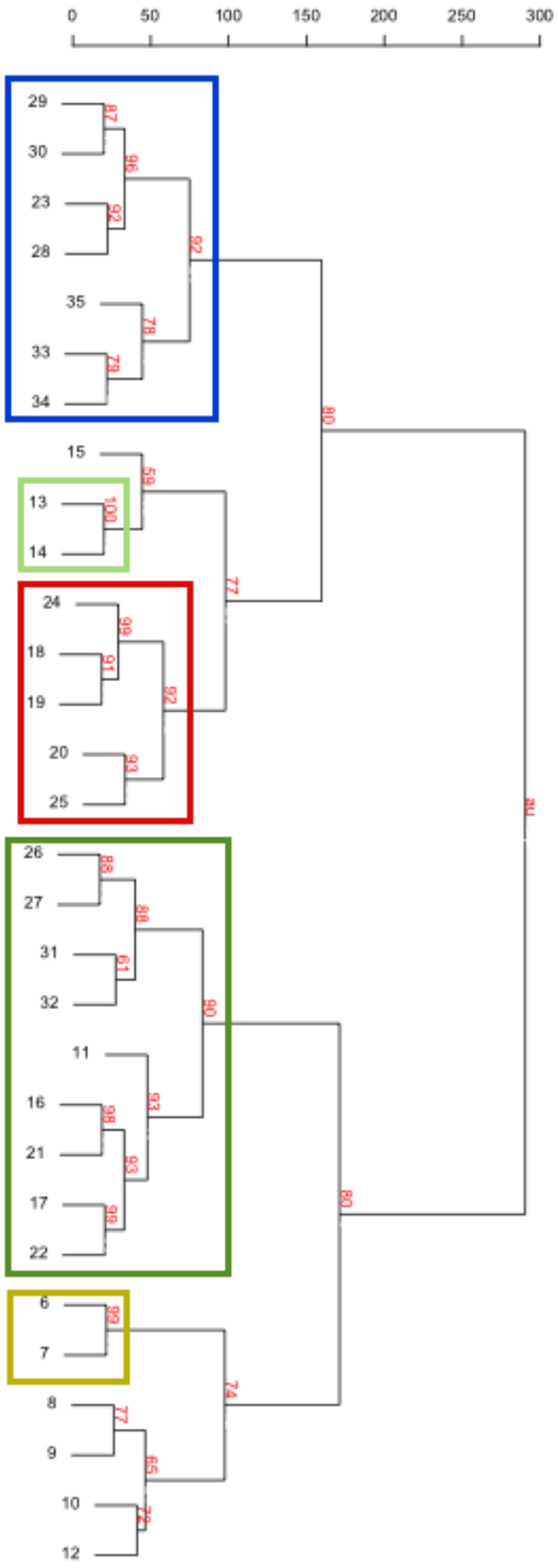



Figure 4.12: Dendrogram representing parcel $p_2$ clustering. Approximately unbiased $p$-values are presented at branch connections as percentages, and cluster labels are presented below the branches.

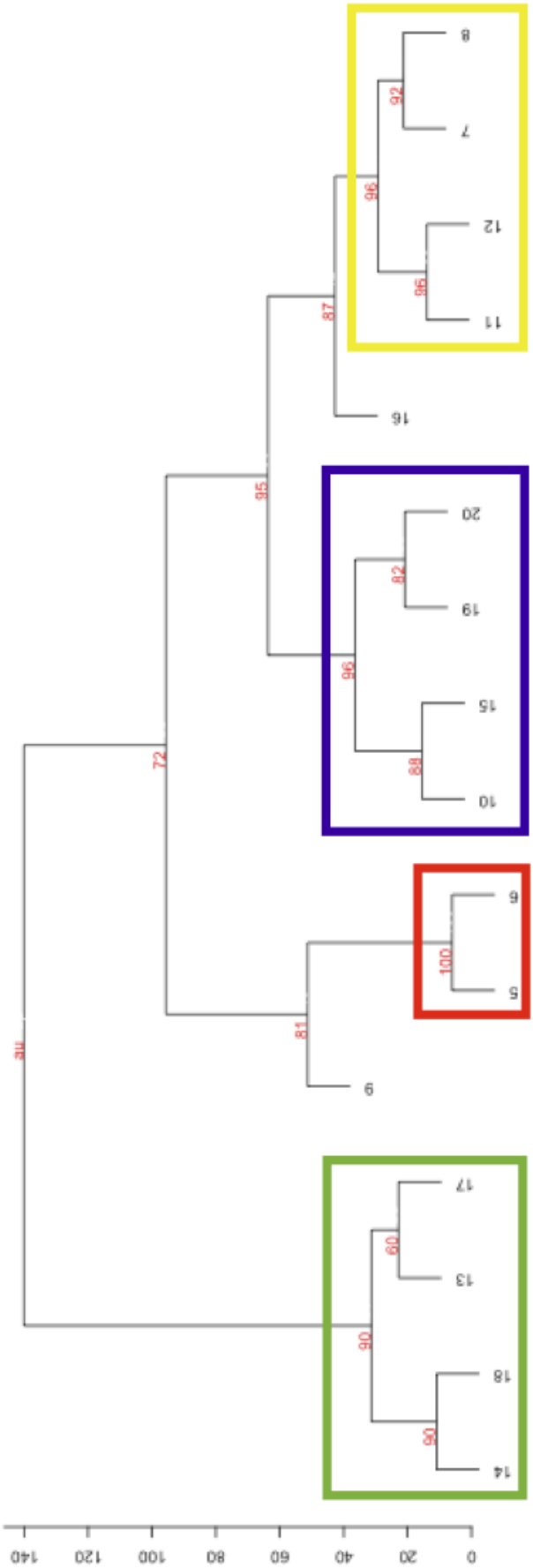



Figure 4.13: Dendrogram representing parcel $p_3$ clustering. Approximately unbiased $p$-values are presented at branch connections as percentages, and cluster labels are presented below the branches.

# Chapter 5

# MAPPING CULTIVABLE LAND WITH MACHINE LEARNING

In PA one of the basic tasks is the classification of land zones in either cultivable or non-cultivable land. Several works have proposed different approaches using data obtained from soil analysis or local exploration of the parcels (Van Alphen, 2002; Godwin and Miller, 2003). However, sometimes only data from satellite images are available and then the problem becomes not only more challenging but also more interesting because it is much more cost-effective.

The possibilities offered satellite images as those provided by Landsat 8 open new paths to obtain a more cost-effective solution to the problem, since soil attributes such as organic matter (Ludwig et al., 2008) or nitrogen content (Xie et al., 2011) may be inferred by means of hyperspectral remote sensing (Luoa et al., 2008).

Thus, this Chapter explores the use of several ML techniques for the task of automatic land classification. In particular, the land able to be used for farming, cultivable land (Fawcett, 1930), using as input satellite remotely sensed data (Kumar et al., 2011) and conducting experiments with different sets of parameters and feature reduction techniques. In addition, it was also studied the performance of spectral and thermal bands provided by Landsat 8 satellite in detecting cultivable land. Finally, a methodology for automatic cultivable land classification is proposed (Arango et al., 2016).

The Chapter is organized as follows. Section 5.1 addresses the supervised learning techniques for automatic cultivable land classification proposed in this work. Section 5.2 shows the results of this supervised approach. Section 5.3 proposes a methodology based on unsupervised ML. Finally, Section 5.4 shows the results of this unsupervised approach.

## 5.1 Automatic cultivable land detection with supervised machine learning

This Section describes the methodology followed in order to obtain land classification as cultivable or non-cultivable zones. Different spectral and thermal bands are considered from the Landsat 8 satellite images corresponding to the vineyard used as case study (see Section 1.3). A range of supervised Machine Learning methods to classify are applied for classify different land zones. The goal is to check the performance of these methods in solving these kind of problems. First, it is detailed the input data and how to prepare it. Then, it is explained the algorithms used to process the information. These methods were tested with the vineyards of Terras Gauda, a wine producer from Galicia, Spain. The study considers one vineyard parcel with topographical dissimilarities and compares the results of cultivable land delimitation with different proposed configurations. The results show that an adequate choice of the algorithm parameters together with feature selection techniques can yield a classification that is both highly effective and efficient.

### 5.1.1 Retrieving input data

Landsat 8 satellite provides radiometric data valuable for precision agriculture (Ormeño Villajos et al., 2008). It takes about 400 images every day using the instruments OLI and TIRS. The first one collects data from 8 spectral bands with a spatial resolution of 30 meters and from a panchromatic band of 15 m. The second one offers two thermal bands of 100 m. resampled to 30 m. All the data collected by these instruments are publicly available in GeoTIFF format via web portals such as Earthnet Online (*EarthOnline*, 2000).

The dataset considered in this study (see Table 5.1) contains the raw reflectivity values of bands 2 through 7 (OLI) and raw thermal infrared values of bands 10 and 11 (TIRS). The OLI bands B1 (Coastal/aerosol) and B9 (Cirrus) are not considered in the dataset because the information of these bands is not related to soil or vegetation. The B8 (Panchromatic) band is not included either because is related with the visible spectrum.

In order to study the relevance of vegetation and moisture indices (see Section 3.3.6), eight widely-used indicators were also calculated and included in the dataset.

In order to obtain the data shown in Table 5.1, which are the data input for the algorithm, it is necessary:

1. Download the Landsat 8 TIRS data product corresponding to the region of study and for the whole winter season of, at least, one crop year. Using EarthExplorer, for instance, will allow to provide search criteria for both the bounding box coordinates of the plots and the data range. However, downloading the data from Amazon S3 will require to know the name of the files and the directory which are located (Amazon, 2015). Both, the name of the files and the directory structure, are based on the path and row numbers of the Worldwide Reference System (WRS) (*Landsat Programme*, 1972) for identifying the Earth's location of a Landsat image.

2. Process the GeoTIFF file (see Section A.2), getting the raw values of the thermal bands and vegetation indices corresponding to the spatial data points of the region of study.

3. Generate a dataset with the following format with the data extracted in the previous task:

$$\begin{pmatrix} f^1_{1,1} & f^1_{2,1} & \cdots & f^1_{n,1} & f^1_{1,2} & f^1_{2,2} & \cdots & f^1_{n,2} & \cdots & f^1_{1,m} & f^1_{2,m} & \cdots & f^1_{n,m} \\ \vdots & & & & & & & & & & & & \vdots \\ f^i_{1,1} & f^i_{2,1} & \cdots & f^i_{n,1} & f^i_{1,2} & f^i_{2,2} & \cdots & f^i_{n,2} & \cdots & f^i_{1,m} & f^i_{2,m} & \cdots & f^i_{n,m} \\ \vdots & & & & & & & & & & & & \vdots \\ f^r_{1,1} & f^r_{2,1} & \cdots & f^r_{n,1} & f^r_{1,2} & f^r_{2,2} & \cdots & f^r_{n,2} & \cdots & f^r_{1,m} & f^r_{2,m} & \cdots & f^r_{n,m} \end{pmatrix}.$$

Each $f^k_{i,j}$ represents the value of band $i$ taken day $j$ for pixel $k$.

In addition, each point is labeled as either being arable or non-arable land considering the shape files from the Terras Gauda's plots and the data of the Sistema de Información Geográfica de Parcelas Agrícolas (SIGPAC) (*Sistema de Información Geográfica de Parcelas Agrícolas*, n.d.), an official database for the identification of agricultural plots in Spain. The labeling process is done by spatial operators with the data points of the region under study and the polygons identified as arable land according to SIGPAC.

| Column | Description |
|---|---|
| x | Coordinate x of the data point in the UTM 29 CRS |
| y | Coordinate y of the data point in the UTM 29 CRS |
| date | Year + Day number of the year in the format YYYYddd |
| B2 | OLI band 2 |
| B3 | OLI band 3 |
| B4 | OLI band 4 |
| B5 | OLI band 5 |
| B6 | OLI band 6 |
| B7 | OLI band 7 |
| B10 | TIRS band 10 |
| B11 | TIRS band 11 |
| NDVI | Normalized Difference Vegetation Index |
| EVI | Enhanced Vegetation Index |
| NDI7 | Normalized Difference Water Index 7 |
| SIWSI | Shortwave Infrared Water Stress Index |
| SWIRR | Shortwave Infrared Ratio |
| MSI | Moisture Stress Index |
| MSI7 | Moisture Stress Index 7 |
| GVMI | Global Vegetation Moisture Index |

Table 5.1: Columns of the dataset for the Landsat 8 data product. Spatial resolution: 30m. Temporal resolution: 16 days.

### 5.1.2 Supervised learning

Once the data have been preprocessed as detailed in the previous sections, it can be applied any SML algorithm in order to obtain a classifier able to determine if a point corresponds to an cultivable land zone. In this work it has been selected some of the most popular SML methods, taking special care in selecting at least from each of the principal approaches. The algorithms applied in the experiments of this work were already explained on Section 2.1.1: C5.0, Naive Bayes, k-NN, SVM with the Radial Basis Function Kernel, which is one of the most commonly used with this method (Chang et al., 2010); and it was fitted a feed-forward single-hidden-layer NN. The implementations used were those provided from the Caret R Package (Kuhn, 2008).

### 5.1.3 Feature selection

The algorithms enumerated in Section 5.1.2 can be applied on all the input data and on subsets that only include those bands and indices selected after performing a process of feature reduction (see Section 2.1.5). To select the best set of attributes it was measured their relevance by means of their IG and then identified the features which have a significantly higher importance (as implemented in the FSelector R Package (*FSelector R Package*, n.d.)). It was also performed a process of redundancy elimination using CFS.

## 5.2 Results of automatic cultivable land detection with supervised machine learning

The experiments conducted in order to test if the methods enumerated in Section 5.1.2 yield effective results in automatic arable land detection are now detailed. First, the settings of the experiments are described and then the results obtained according to that settings are shown.

### 5.2.1 Settings of the experiments

The dataset is obtained by downloading OLI and TIRS data with EarthExplorer, for the extension of the plots of Terras Gauda (see Fig. 1.8). The data collected correspond to the dormant stage in the vines (the winter season of 2013-14).

| Method | Precision | Recall | F1 | AUC |
|---|---|---|---|---|
| k-NN | 0.8112 | 0.8145 | 0.8129 | 0.8768 |
| Naive Bayes | 0.7350 | 0.7944 | 0.7636 | 0.8244 |
| **C5.0** | **0.9178** | **0.8852** | **0.9013** | **0.9692** |
| SVM Radial | 0.8056 | 0.8185 | 0.8120 | 0.8714 |
| Neural Networks | 0.7659 | 0.7782 | 0.7720 | 0.8432 |

Table 5.2: Results when no feature selection is performed

The data points prepared as explained in Section 5.1.1 are divided into a training set and a test set using stratified cross validation (Kohavi, 1995) with a proportion of 0.75 and 10 repetitions. Then, on the training set two feature selection approaches are performed: in the first one, we compute the IG of each attribute and then select those which are maximally separated from the rest; the second one is the application of CFS. Thus, we have three sets of training values: those corresponding to features selected via IG, those selected with CFS and the whole set, which includes all features.

These sets are the input of the methods of Section 5.1.2. The optimal parameters for the algorithms are determined using cross validation on the training data. Finally, the classifiers are tested on the values that were not used for training.

### 5.2.2 Performance evaluation

The different classifiers are evaluated through a confusion matrix with the predictions of the different methods and the current values of the test points. For computing the matrix it was used the same assumptions than those in Section 5.4.2. The metrics obtained were those described in Section 2.1.6. In addition, it was study the overall behaviour of the classifiers by means of ROC curves and the study of AUC.

### 5.2.3 Results

The results obtained when using the optimised methods (see subsection 5.2.3) without feature selection are shown in Table 5.2. Figure 5.1 shows the ROC curves associated to the five used algorithms. All the values are the average of the 10 cross validation repetitions. As it can be seen, almost all the methods obtain fairly good results, but C5.0 clearly outperform the others, exceeding in more than 10 points the other methods with regard to all the evaluation metrics. These results are reinforced by the ROC curve behaviour (shown in Figure 5.1), where the area under the curve obtained by C4.5 is the closest to one.

Table 5.3 and Figure 5.2 present the values and ROC curves for the methods when feature selection is performed with IG. In this case, only three attributes were selected (namely, $B5.2$, $B5.6$ and $B5.9$). Again, the results are averaged over all cross validation repetitions. In this case, the results are very similar one to another (and, for some of them, lower than those obtained without performing feature selection), probably because the number of features that were taken into account is very low. In addition, the values obtained for both precision, recall and $F_1$ are far from those obtained without performing feature selection. Note also that all the features selected with IG are related to the same band in 3 different days (the suffix $.n$ in the variable name means the day the value was taken). That means this feature selection method only considers the information of band B5. From this perspective, the results with only one band are quite successful.

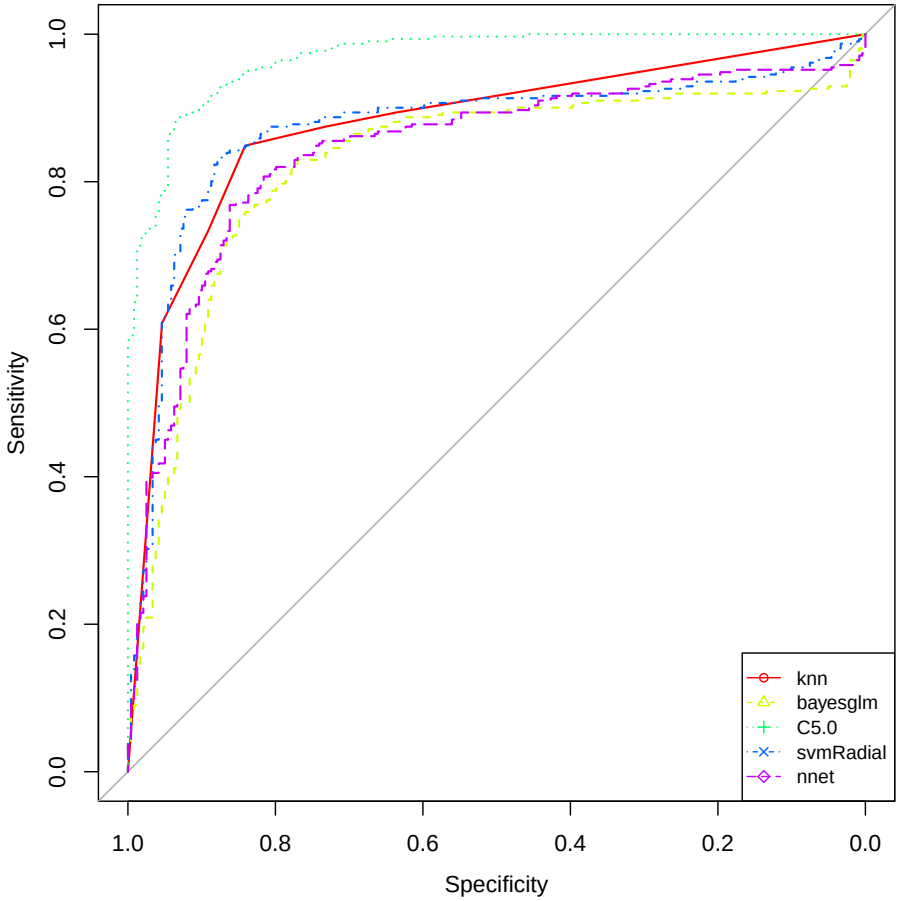



Figure 5.1: ROC curves when no feature selection is performed

| Method | Precision | Recall | F1 | AUC |
|---|---|---|---|---|
| k-NN | 0.8224 | 0.7213 | 0.7686 | 0.8707 |
| Naive Bayes | 0.8365 | 0.7268 | 0.7778 | 0.8507 |
| C5.0 | 0.8228 | 0.7104 | 0.7625 | 0.8625 |
| SVM Radial | 0.8207 | 0.7377 | 0.7770 | 0.8602 |
| Neural Networks | 0.8190 | 0.7295 | 0.7717 | **0.8781** |

Table 5.3: Results when features are selected according to IG

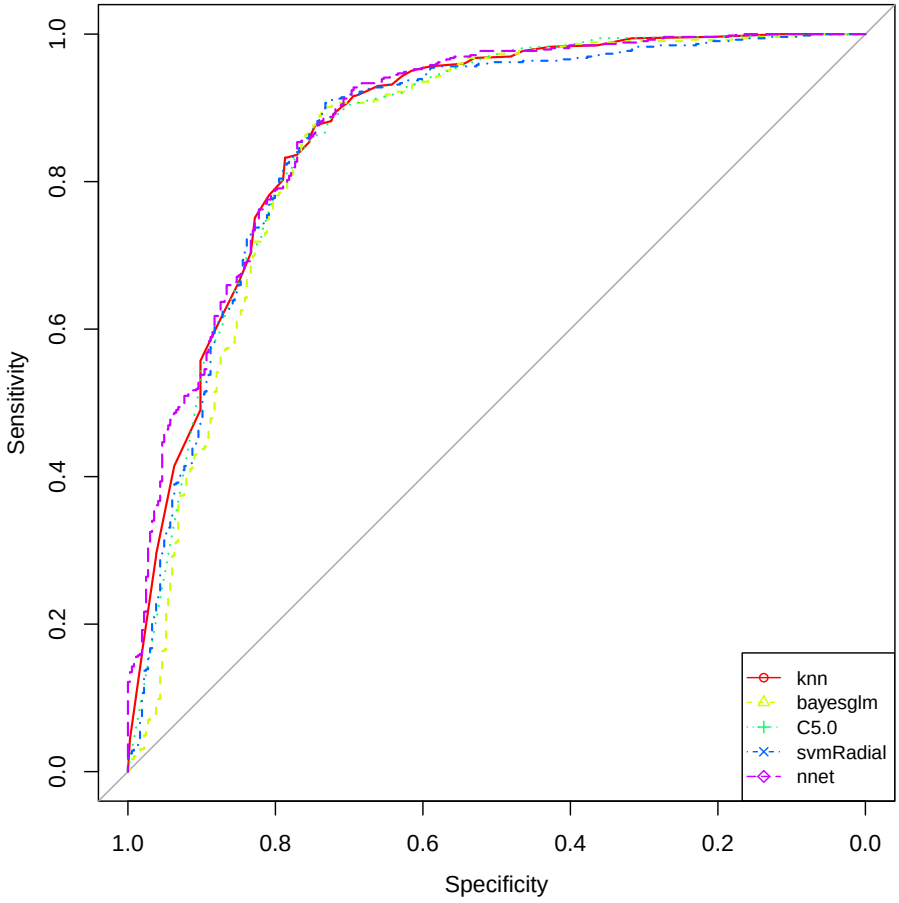



Figure 5.2: ROC curves when features are selected according to IG

Finally, Table 5.4 and Figure 5.3 show the values and ROC curves of the methods when CFS was used to select the most relevant features (in this case, 19 of the 170 total attributes were selected). According to Precision, Recall,F1 and AUC values, $K-NN$, SVM Radial and Neural Networks based methods combined with CFS feature selection method performs better than both $K-NN$ without feature selection or with feature selection via IG. Regarding feature selection methods associated to Naive Bayes, CFS shows higher Precision, F1 and AUC than no feature selection. It also improves Recall, F1 and AUC when comparing it to IG. Finally, the

best results in terms of Precision, Recall, F1 and AUC for C5.0 are obtained when no feature selection is applied, although the results obtained filtering with CFS are quite close to those obtained without feature selection.

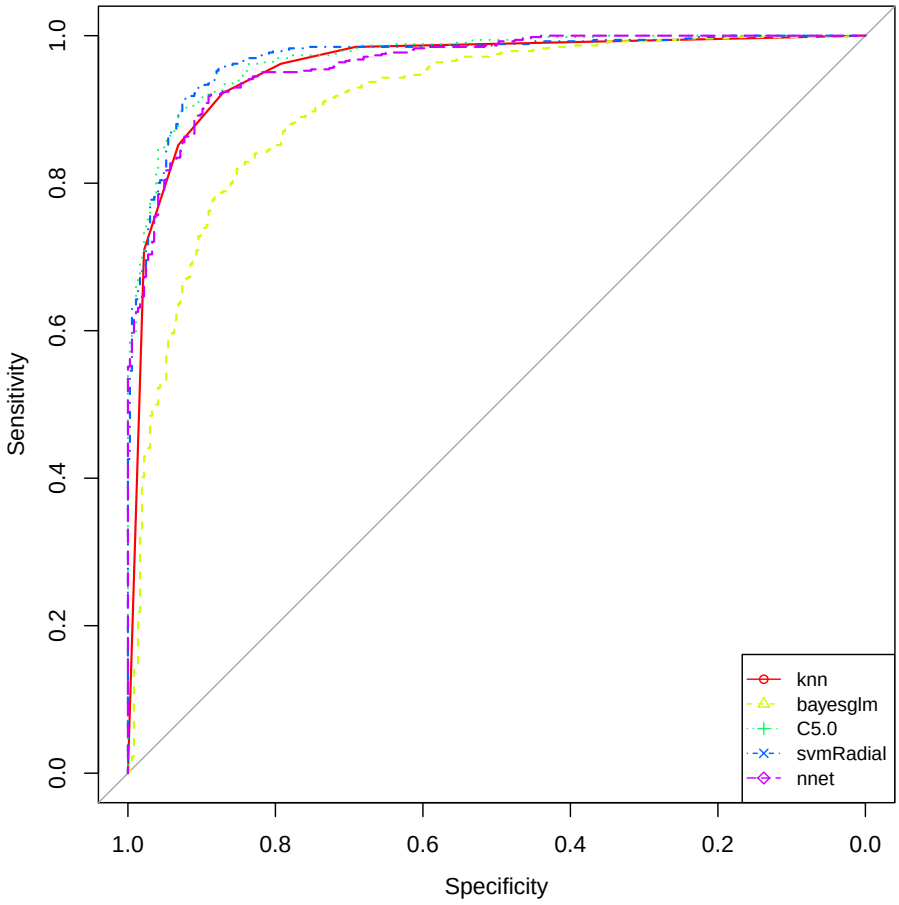



Figure 5.3: ROC curves when features are selected according to CFS

| Method | Precision | Recall | F1 | AUC |
|---|---|---|---|---|
| k-NN | 0.8861 | 0.8716 | 0.8788 | 0.9564 |
| Naive Bayes | 0.8279 | 0.7623 | 0.7937 | 0.9071 |
| C5.0 | 0.9008 | 0.8689 | 0.8846 | **0.9698** |
| SVM Radial | 0.9078 | 0.8880 | 0.8978 | 0.9693 |
| Neural Networks | 0.8740 | 0.8907 | 0.8823 | 0.9610 |

Table 5.4: Results when features are selected according to CFS

Thus, most methods show a significant improvement in effectiveness when compared to the case with all the features and, with the exception of Naive Bayes, all of them obtain AUCs over 0.95. Also, this setting is very efficient because it uses less than 12% of the features.

However, CFS negatively affects C4.5, obtaining a performance slightly under that obtained without feature selection for precision, recall and F1.

On the other hand, the features selected by CFS were $B2.3$, $B2.4$, $B3.2$, $B3.5$, $B3.6$, $B3.9$, $B5.2$, $B5.6$, $B5.9$, $B10.2$ , $B10.3$, $B11.7$, $NDVI.2$, $NDVI.7$, $NDVI.8$, $NDVI.9$, $EVI.5$, $EVI.9$, $SWIRR.9$. Note that those features selected by IG are contained in the features selected by CFS. In addition, CFS does not select any information related to OLI bands $B4$, $B6$, $B7$. Regarding vegetation indexes, it only considers $NVDI$, $EVI$ and $SWIRR$. It is also important to remark that there is only one day not considered at all by CFS (day 1).

Considering both performance in terms of Precision, Recall, F1 and AUC, and also in terms of model complexity, it is clear that the recommended method is to perform feature selection method combined with C5.0 or SVM Radial.

**Optimization of parameter methods**

Some of the machine learning methods were optimized according to certain parameters. k-NN was optimized with regard to the number of elements in the vicinity. Figure 5.4 shows the behaviour of parameter $k$ (ranging from 2 to 50) when the classifier was trained using the whole feature set (left), the features selected by CFS (middle) or by IG (right). Note that the behaviour of $knn$ is more or less the same when no filtering or when feature selection is performed using CFS, but it is almost the opposite when features are selected according to IG.

This behaviour is shared by SVM with radial kernel (Figure 5.5). In this case, the parameter to be optimised was the cost (ranging from 0 to 70). Finally, Figure 5.6 shows the optimization of neural networks regarding the number of hidden units (from 1 unit to 19 units) and the weight decay (from 0 to 0.1) (see artificial NN in Section 2.1.1). In this case, when all the features are considered, the ROC value is above 0.8 only when 1, 17 or 19 hidden units are considered. These numbers of hidden units are the ones performing better also when CFS is applied. Again, when IG filtering is considered the behaviour of the parameters is completely different.

Naive Bayes does not provide any parameter to be optimised. Regarding C5.0, the three parameters to optimize are model (set to "tree"), winnow (set to FALSE) and trials (ranging from 1 to 100). The optimisation according to trials is more or less the same for the three feature sets (see Kuhn (2008)).

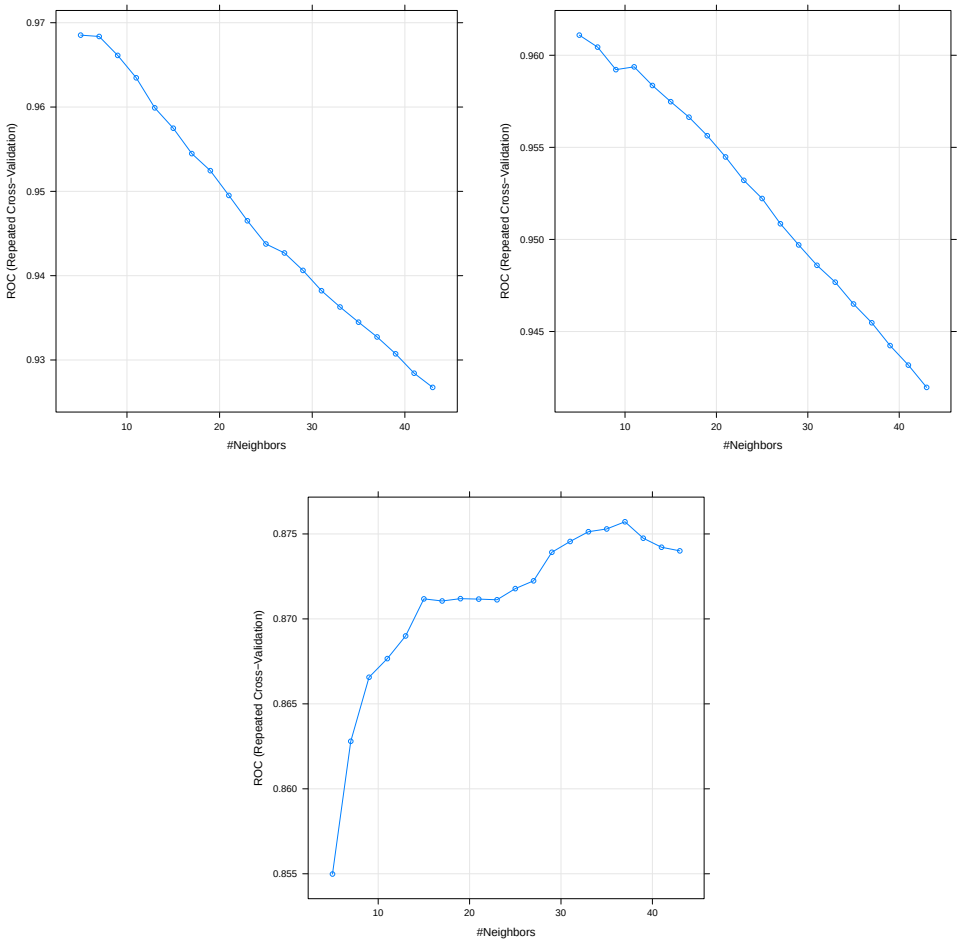



Figure 5.4: Optimum k. Without Feature Selection (top-left), CFS (top-right), IG (bottom)

## 5.3 Mapping cultivable land from satellite imagery with clustering algorithms

In this Section is proposed a methodology for the automatic delimitation of land able to be used for farming, cultivable land (Fawcett, 1930), using clustering algorithms with publicly available satellite data. The method uses a partition clustering algorithm called PAM (see Section 2.1.3) and considers the quality of the clusters obtained for each satellite band in order to evaluate which one better identifies cultivable land.

In addition, it was also studied the performance of spectral and thermal bands provided by Landsat 8 satellite in detecting cultivable land. The approach developed was tested and applied to the case study (see Section 1.3). The study considered three plots with topographical dissimilarities and compared the results of cultivable land delimitation from clusters obtained using different spectral and thermal bands. The experimental results show the great potential of this method for cultivable land monitoring from remote-sensed multispectral imagery.

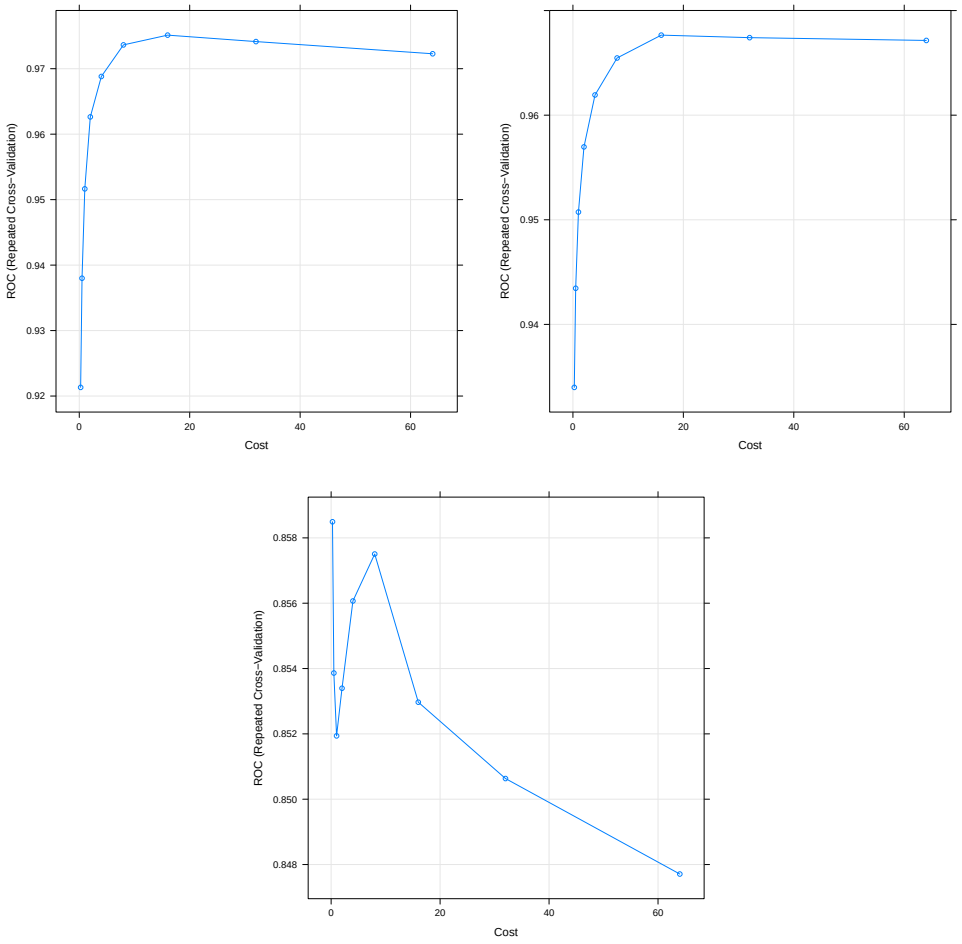



Figure 5.5: Optimum Cost. Without Feature Selection (top-left), CFS (top-right), IG (bottom)

### 5.3.1 Cultivable land delimitation methodology

The proposed method groups the pixels of multispectral images acquired by on-board satellite instruments for a desired land zone and period of time, in two main clusters: cultivable and non-cultivable land. Each element of the clustering represents one pixel of the image and is assigned to a group by means of a dissimilarity metric. The method tries to group as well the pixels in three, four and five clusters, with the aim of selecting the better clustering configuration. In this regard, the scientific literature proposes the calculation of indices (Milligan and Cooper, 1985) that measure the quality of the clustering. Subsection 5.3.4 explains the index used by the method. If as a result the clustering has more than two groups then the method generates two metaclusters merging the clusters.

Note also that one of the objectives of this work is the study of the performance of Landsat 8 satellite in detecting cultivable land. Thus, the clustering algorithm is applied as many times as the considered bands and taking as input one satellite band each time. The main steps are the following:

1. Download and process the Landsat 8 data products corresponding to the region under

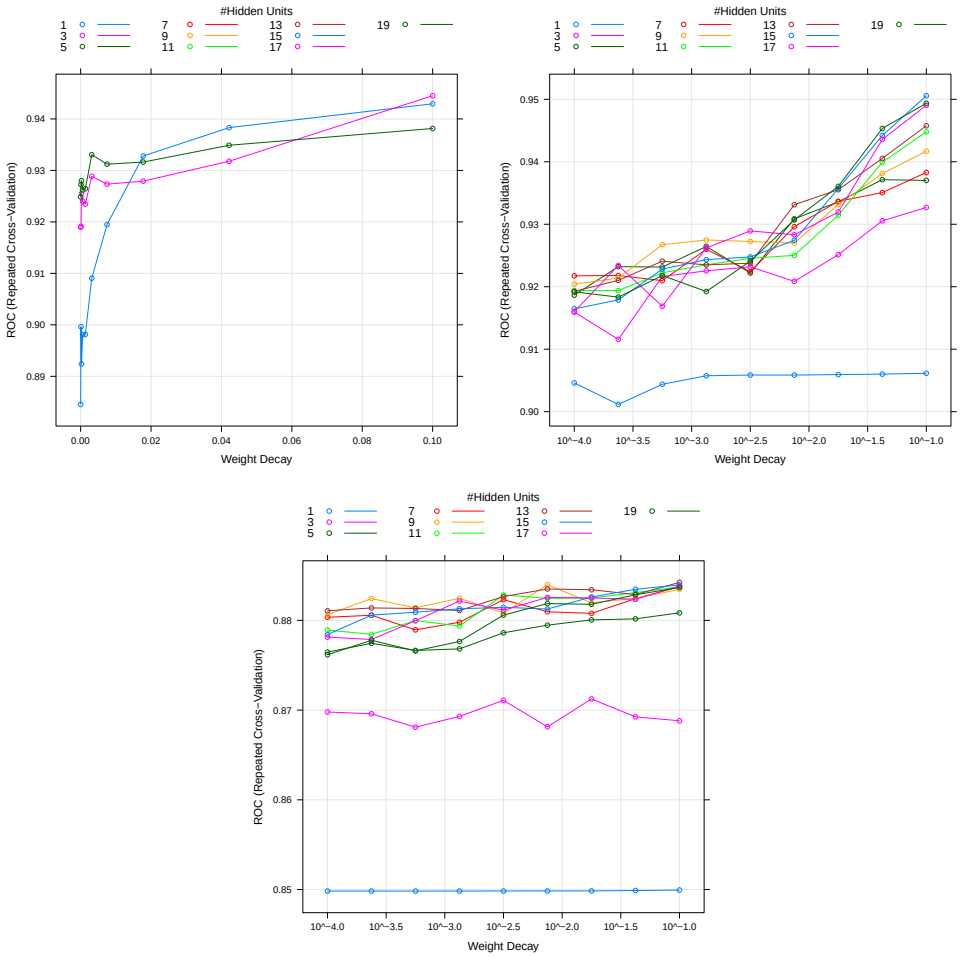



Figure 5.6: Optimum number of hidden layers with respect to weight decay. Without Feature Selection (top-left), CFS (top-right), IG (bottom)

study, getting the raw values of the spectral and thermal bands.

2. For each band $B_i$:

   - Perform the clustering with a number of clusters $k$ varying from 2 to 5.
   - Select the number of clusters $k$ maximizing a quality clustering index (see Subsection 5.3.4).
   - If $k > 2$, merge clusters in 2 groups (associated to cultivable and not cultivable land) by computing distances between each pair of cluster representatives.

As output of the method, each pixel from the considered image is labeled as cultivable or non-cultivable land. In this point it is possible to generate, for instance, a new layer for the raster image with the aim of producing a map of cultivable/non-cultivable land (see Fig. 5.7).

In the following sections, all the steps involved in the methodology are described in detail.

Figure 5.7: Mapping cultivable land cluster for the land parcel 2 with the Band 5 (NIR). The zeroes represent the land correctly classified as non-cultivable. The ones represent the the land incorrectly classified as non-cultivable. The zones without numbers belong to the cultivable-land cluster.

### 5.3.2 Input data

The input data used by this SML approach is the same as the describe in Section 5.1.1 with the exception of the class of each data point, which is not considered because is the class to predict.

### 5.3.3 Clustering algorithm

As it was described in Section 5.3.2, the data clustered in this approach include spectral and thermal infrared data collected by satellite from the area of land intended to separate cultivable land from non-cultivable land. For this purpose, is used the partition clustering algorithm called PAM, already explained in Section 2.1.3. This algorithm was chosen instead of the well-known

k-means because PAM is more robust since it minimizes a sum of dissimilarities instead of a sum of squared euclidean distances. The main difference between this algorithm and the classical k-means method is that PAM uses medoids as centers of the clusters and these medoids are selected among the objects to be clustered.

The objective function used is the sum of the distances from each object to the closest medoid. The distance proposed in this work is the Manhattan distance. Once the number of clusters is computed, the cluster assignment is retrieved, providing the land delimitation.

### 5.3.4 Selection of the clusters

In the case of the identification of cultivable land, it used a quality index in order to select those clusters maximizing the index as candidates for the cultivable land identification. Initially, the top 5 indices in Milligan and Cooper study (Milligan and Cooper, 1985) were considered, including both Silhouette coefficient (Rousseeuw, 1987) and Calinski-Harabasz index (Caliński and Harabasz, 1974). It was selected one of them taking into account the computational cost. Specifically, the Calinski-Harabasz index was selected to validate the clustering (see Section 2.1.4).

As the goal of this work is to identify those fractions of land which can be considered as cultivable, if the optimal number of clusters is larger than two, they are reduced to two clusters just by computing the distances between cluster representatives and merging the $k$ clusters into two groups according to these distances. The procedure can be summarized in the following steps.

- Select the representative $x_i$, from each one of the $k$ clusters.
- Run PAM to cluster $\{x_i, i = 1, \ldots, k\}$ into two clusters considering as input the matrix of distances between $x_i$ and $x_j$, $\forall i, j = 1, \ldots, k, i \neq j$.
- Assign each point to the final clusters.

## 5.4 Results

In this section experimental results are shown in order to evaluate: (a) which satellite band better identifies cultivable land and (b) the proposed methodology for mapping cultivable land from satellite imagery using clustering algorithms. To this aim, a case study was carried out with the vineyards of Terras Gauda (see Section 1.3).

Regarding the input data, the dataset is obtained by downloading OLI and TIRS EarthExplorer data for the extension of the land parcels of Terras Gauda. The data collected correspond to the dormant stage of the vines (the winter season of 2013-14).

### 5.4.1 Automatic land identification

According to the proposed method (see Section 5.3.1) it were considered as candidates for the automatic land identification the clusters with highest values for the Calinski-Harabasz quality index.

Fig. 5.8 to Fig. 5.10 show the performance of the clustering algorithm with regard to the Calinsky-Harabasz (CH) index for the three different land parcels. Each line represents the results obtained using as input a certain band (B2, B3, B4, B5, B6, B7, B10 and B11). The X-axis represents the number of clusters (ranging from 2 to 5) and the Y-axis contains the value of the CH index.

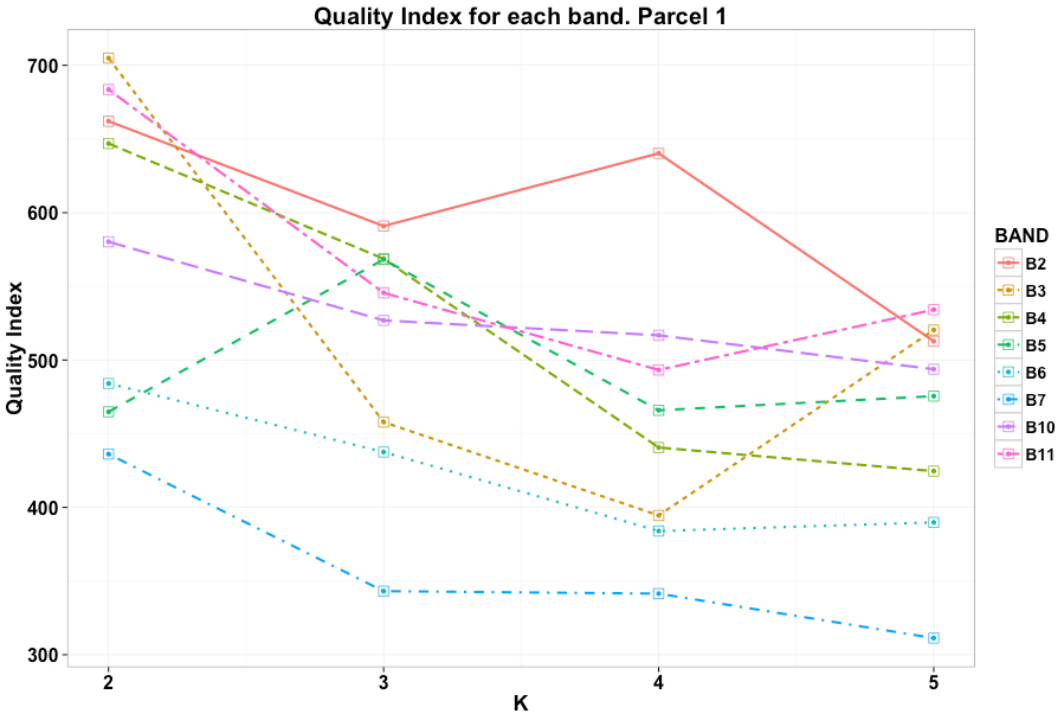



Figure 5.8: Performance of different bands for clustering the land parcel 1

Fig. 5.8 shows the behavior of the different bands when the number of clusters ranges from 2 to 5 for parcel 1. The highest CH value is obtained when $k = 2$ and B3 as input data. Note that bands B2 and B11 also obtains high CH for k=2. However, when $k$ increases the behavior of the clustering using B3 as input data steeply falls. As B2 stays more or less the same, this band is selected.

Fig. 5.9 shows the behavior of the different bands when the number of clusters ranges from 2 to 5 for parcel 2. In this case, the band with best overall performance is B5, which also achieves the highest CH value when $k = 2$.

The behavior of the different bands and $k$ for parcel 3 is shown in Fig. 5.10. Bands B3, B6 and B7 are the ones performing better regarding the CH index. However, as it can be seen in Fig. 5.10 CH for B6 and B7 gradually go down when $k$ increases. In contrast, band B3, which also obtains high CH values is more stable. Thus, in addition to $k = 2$ and B7 it seems that one could select B3 and $k = 2, 3$ as optimal configuration.

In this case it is possible to validate the performance of the method as the real boundaries of the cultivable land are known. Therefore we can check if the configurations obtaining the highest CH also obtain the best performance in classification. The following section is devoted

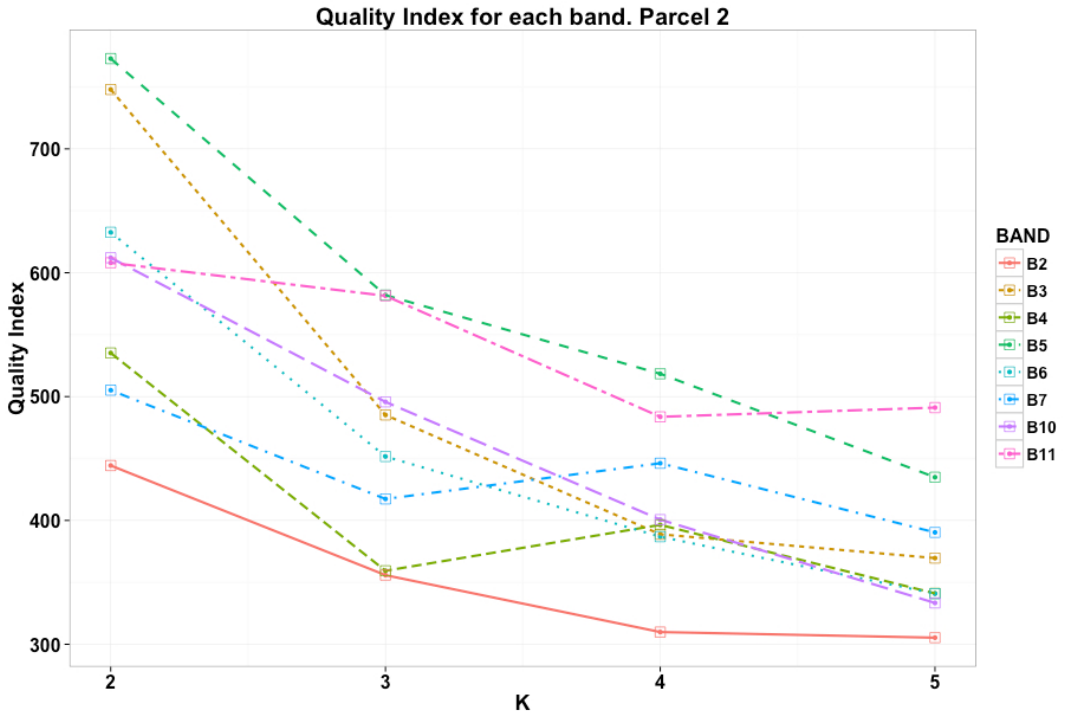



Figure 5.9: Performance of different bands for clustering the land parcel 2

to this comparison.

### 5.4.2 Validation of cultivable land identification

When it is possible to compare the clusters obtained from the clustering algorithm to some other previously defined, the performance of our method can be measured through a confusion matrix and the metrics describe in Section 2.1.6. Considering for each cluster, True negatives ($a$) as the points correctly classified as not belonging to the cluster, False positives ($b$) as the points incorrectly classified as belonging to the cluster, False negatives ($c$) as the points incorrectly classified as not belonging to the cluster and True positives ($d$) as spatial data points correctly classified as belonging to the cluster.

According to the clustering and identification of cultivable land, Fig. 5.11 presents the metrics for the bands of the instruments OLI and TIRS considered in this study. The best $F_1$ is obtained for B2 while the results obtained by B3 are not good. Note that the behavior of B3 regarding CH index was zig-zagged. The results shown in Fig. 5.11 corresponds to $k = 4$. The values obtained for Precision, Recall and $F_1$ when $k = 2$ are worse. Note that the shape of this parcel is a bit irregular and the zone of cultivable land is not connected so the difficulty in detecting borders is higher.

Fig. 5.12 shows the Precision, Recall and $F_1$ for the clustering with the different bands for parcel 2. In this case $k = 2$ because the highest values associated to each band are obtained when $k = 2$ and then all of them decreases. The best $F_1$ is obtained for B5. Note that B5 is the band obtaining the highest CH values (see Fig. 5.9).

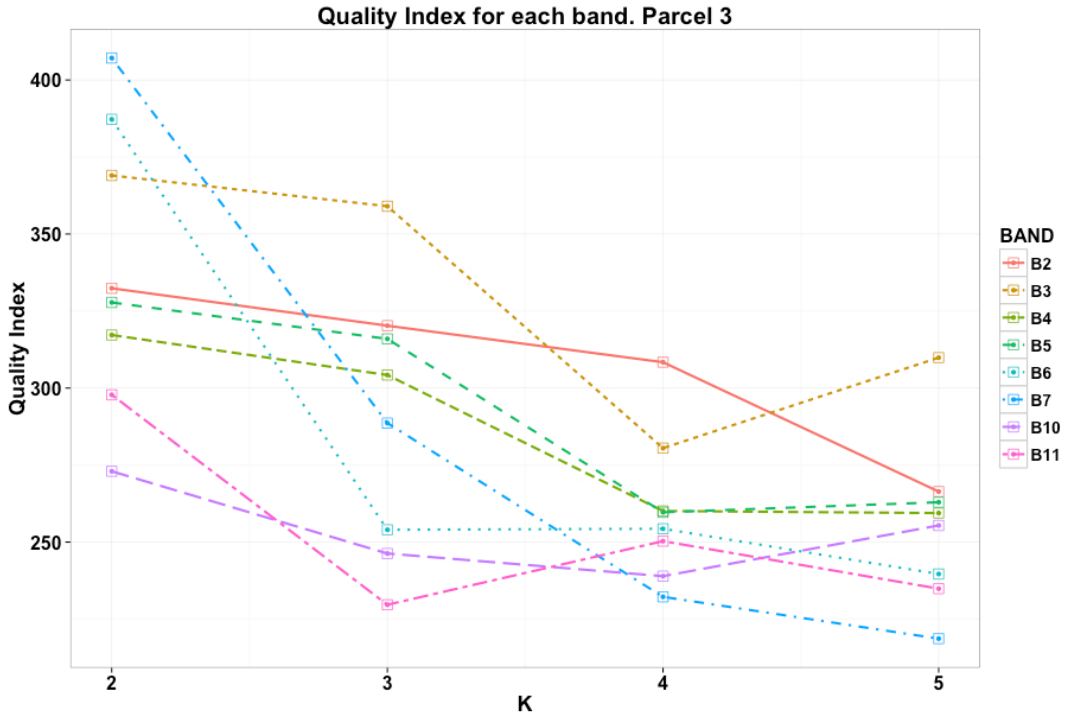



Figure 5.10: Performance of different bands for clustering the land parcel 3

Finally, Fig. 5.13 shows the Precision, Recall and $F_1$ for the clustering with the different bands for parcel 3. In this case $k = 2$ represents again the number of clusters obtaining the highest CH values and for this reason Fig. 5.13 represents the performance of the clustering when $k = 2$ across the different bands. Note that the bands reaching the highest CH (B3, B6 and B7) also obtain the highest Precision, Recall and $F_1$.

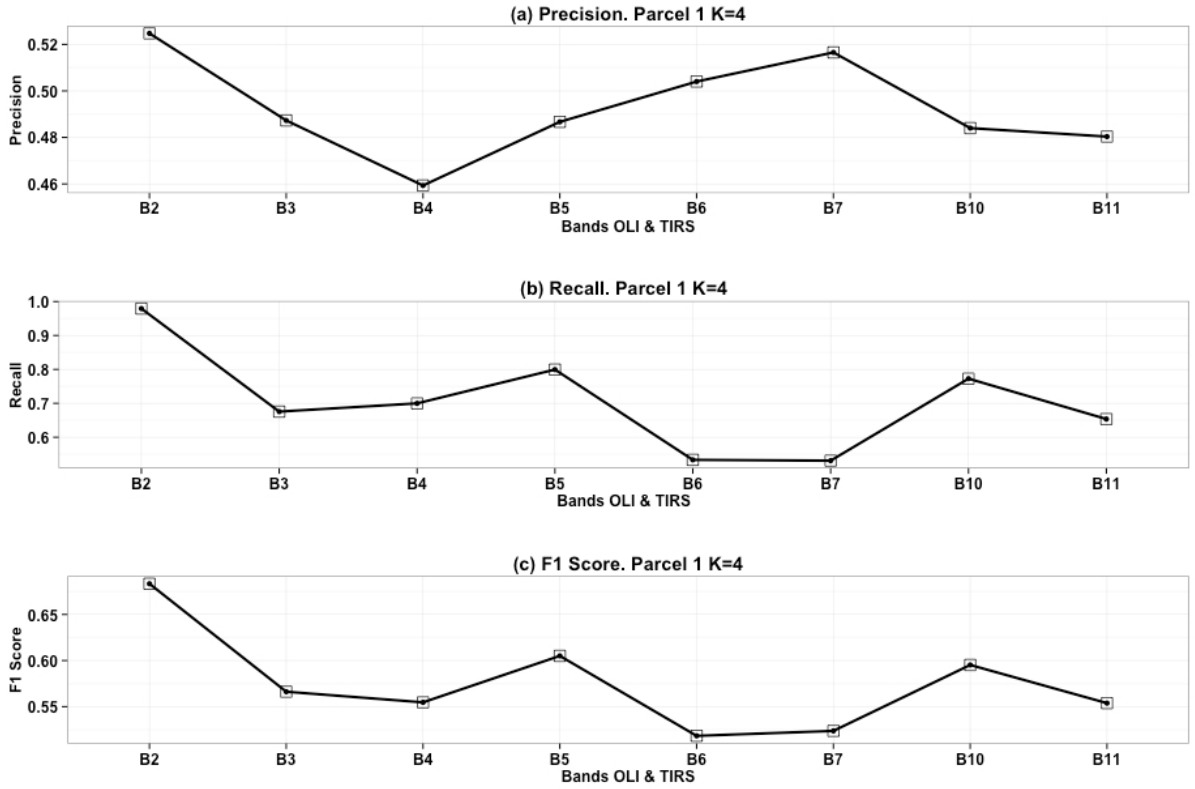



Figure 5.11: Metrics related with the automatic delimitation of the land parcel 1. Band 2 (Blue) seems to be a good estimator to cluster correctly cultivable land

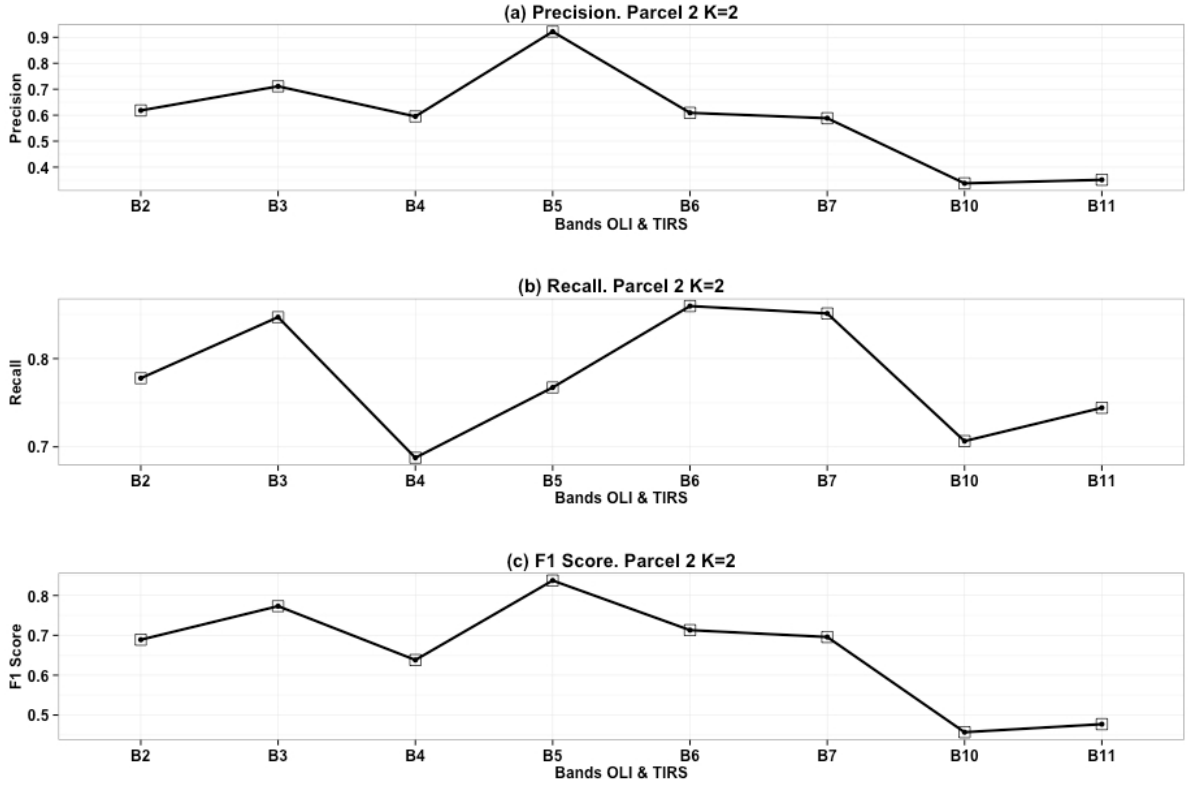



Figure 5.12: Metrics related with the automatic delimitation of the land parcel 2. Band 5 (NIR) seems to be a good estimator to cluster correctly cultivable land

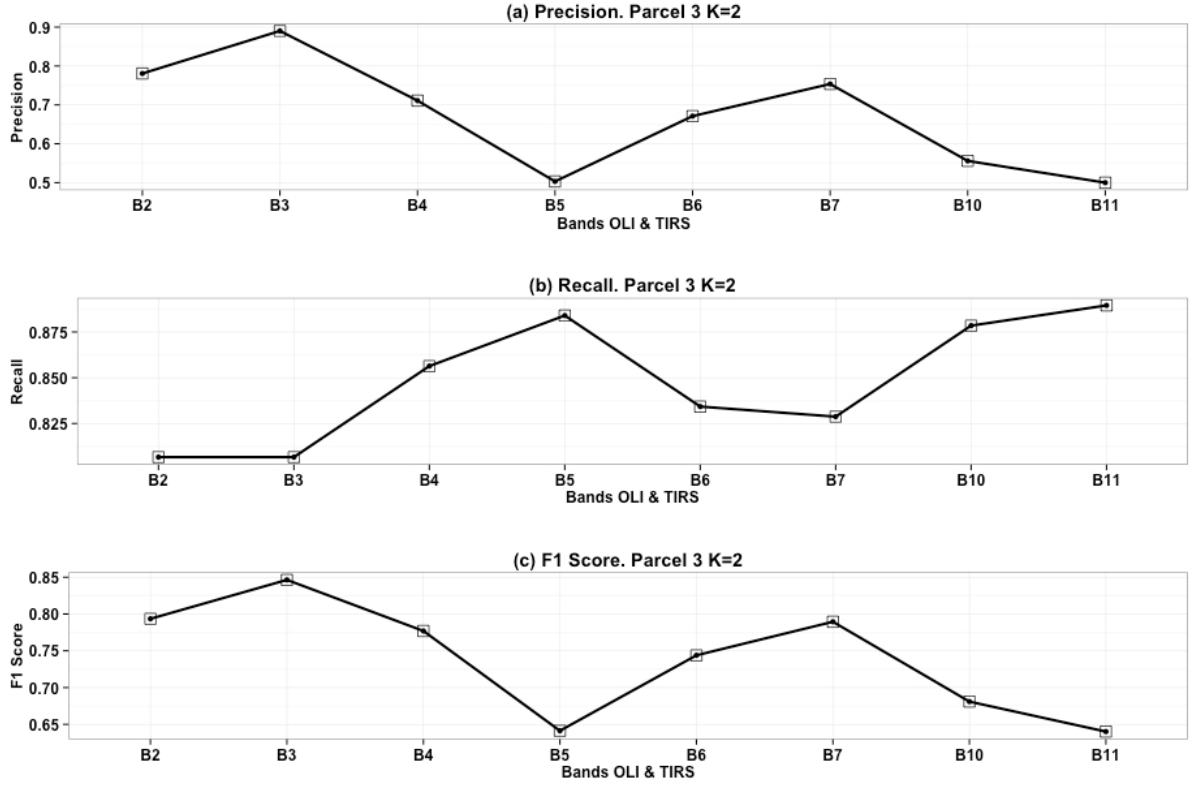



Figure 5.13: Metrics related with the automatic delimitation of the land parcel 3. Band 7 (SWIR2) seems to be a good estimator to cluster correctly cultivable land

# Chapter 6

# IDENTIFICATION OF AGRICULTURAL MANAGEMENT ZONES WITH UNSUPERVISED LEARNING ALGORTIHMS

Other important task in PA entails the appropriate management of the inherent variability of soil and crops, resulting in an increase of economic benefits and a reduction of environmental impact. However, site-specific treatments require maps of the soil variability to identify areas of land that share similar properties. Although the tacit knowledge of the farmer about crops and soil could be a starting point for MZs identification, other systematic approaches are required.

In order to produce these maps of MZs, it is proposed a cost-efficient method that combines clustering algorithms with publicly available satellite imagery. The method does not require exploring the parcels with any special equipment neither taking soil samples for laboratory analysis.

The proposed method was tested in a case study for three vineyard parcels with topographical dissimilarities. The study compares different spectral and thermal bands from the Landsat 8 satellite as well as vegetation and moisture indices to determine which one produces the best clustering. The experimental results were evaluated according to a previous study at this location for the delimitation of MZs.

As it is shown in Section 6.2.3 the results seem promising for identification of agricultural management zones. The findings suggest that thermal bands produce better clustering than those based on the NDVI index.

The Chapter is organized as follows. Section 6.1 explains the tasks involving the methodology proposed in this work for the automatic identification of MZs. Section 6.2 focuses on the results of the method applied to the case of study.

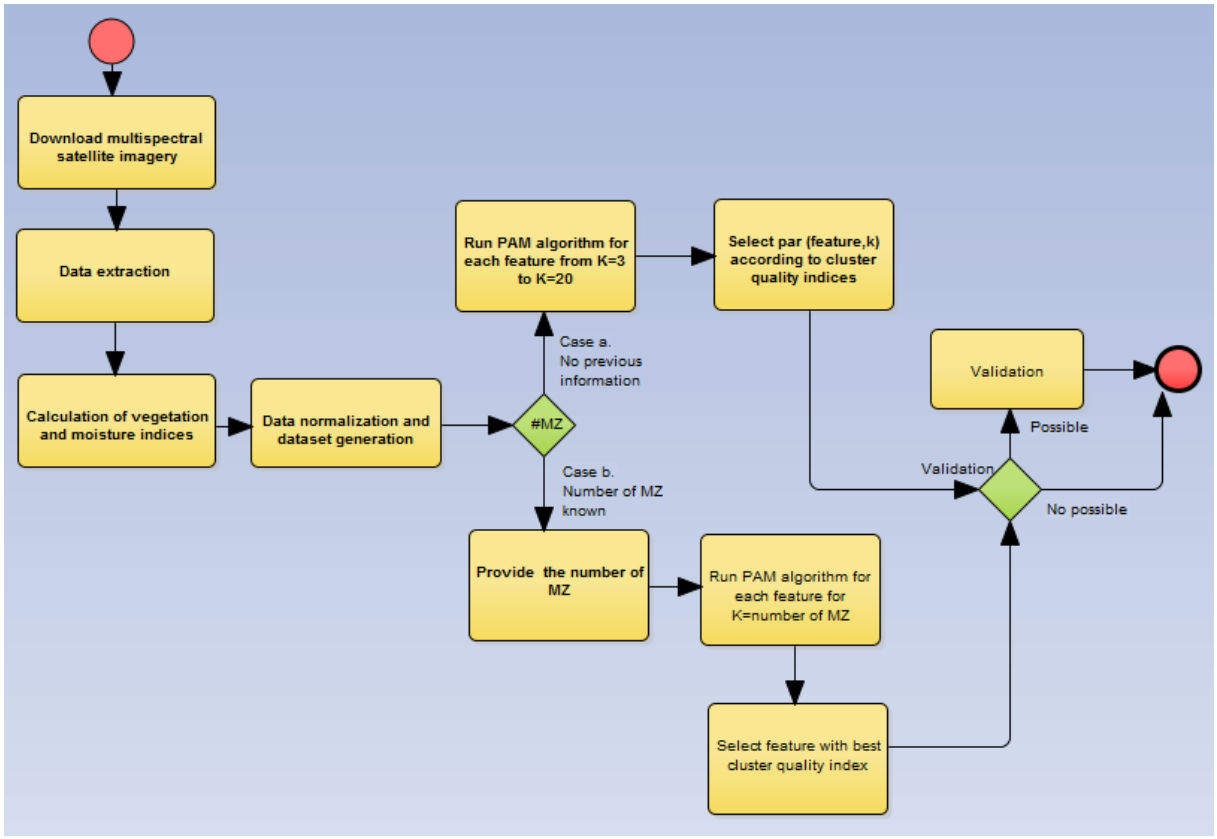



Figure 6.1: Tasks involving the proposed method for the MZs identification based on clustering remote-sensed spectral and thermal infrared data from satellite.

## 6.1 Management zone identification methodology

This Section describes the methodology proposed for MZ identification by means of clustering algorithms using thermal and multispectral imagery from satellite as input. The proposed method considers two scenarios or use cases. The first one (Case a) considers the number $k$ of MZs as unknown and then the method looks for the optimum $(feature, k)$ pair according to cluster quality. The second one (Case b) identifies the best feature for a specific number of MZs, fixed by an agronomist expert. The method involves the following tasks (see Fig. 6.1):

1. Data collection, transformation and dataset generation.

   (a) Multispectral and thermal imagery acquisition by downloading the satellite data products corresponding to the region under study.

   (b) Data extraction by processing the imagery files. The raw values of the spectral and thermal bands are obtained after this step.

   (c) Calculation of vegetation and moisture indices from raw values of thermal and multispectral bands.

   (d) Data normalisation and dataset generation.

2. Clustering and MZ identification. In this step the performance of each feature (bands and indices) in detecting MZs is tested using a clustering algorithm.

   (a) Case a. No previous information about the number of MZs $k$.

      i. PAM algorithm execution for each feature of the dataset and for each $k$ clusters from $k = 3$ to $k = 20$.

      ii. Selection of the pair $(feature, k)$ providing the best clusters according to the quality indices.

   (b) Case b. The number of MZs $k$ is known in advance.

      i. PAM algorithm execution for each feature of the dataset and $k$ clusters.

      ii. Selection of the $feature$ providing the best clusters according to the quality indices.

3. Clustering validation if possible.

The following subsections describe each step of the process in detail.

### 6.1.1 Data collection, transformation and dataset generation

The method for the automatic identification of MZs requires the following tasks in order to produce the dataset used as clustering algorithm input.

#### Multispectral and thermal imagery acquisition

The clustering process considers surface reflectivity and thermal infrared data corresponding to the region of study during a period of time of at least three years to allow reflectivity patterns in the sample. Landsat 8 and Sentinel-2 are well-known examples of satellites that provide multispectral imagery publicly available (see Section 3.1) and covering almost the entire surface of the Earth.

In the case of Sentinel-2, the MultiSpectral Instrument (MSI) collects 13 spectral bands between latitudes 56° south and 84° north with a revisit time of ten days (or five days when operating the second Sentinel-2). It provides high resolution imagery, from 10 to 60 meters, and its spectrum range from visible and the near infrared to the shortwave infrared. The data products are available in GeoTIFF format at the Sentinels Scientific Data Hub (https://scihub.esa.int/)

#### Data extraction

In this task GeoTIFF files are processed, getting the raw values of the layers corresponding to the spatial data points of the region under study. To that end, there are analysis and processing tools such as the Sentinel-2 Toolbox (https://sentinel.esa.int/web/sentinel/toolboxes) that also supports third party data as, for instance, MODIS and Landsat. Other common approaches are the use of scripts in Python or R. In this work, R scripts were used (see Section A.2).

### 6.1.2 Data normalisation and dataset generation

From the previously extracted information, a dataset is generated with the format explained on Section 5.1.1 with the exception of the class of each data point, which is not considered because is the class to predict.

### 6.1.3 Clustering and management zone identification

Once the data are processed, the MZ identification is obtained using clustering algorithms as described Section 2.1.3. As it was stated previously, the MZs identification method proposed here relays on a clustering algorithm. Specifically, the partition clustering algorithm PAM is applied. Other cluster algorithms such as K-means were tested, but the experimental results were not as good as those obtained with PAM.

Regarding the selection of $k$, the number of clusters, in the case of the identification of MZs will be the number of MZs. This number, as we aforementioned and the beginning of this section, may be establish by the technical expert on agriculture (Case b) or, if it is not possible, it may be estimated (Case a). This subsection studies how to estimate $k$ when no input from agriculture experts about the number of MZs is provided.

Initially, the top-5 indices in Milligan and Cooper study (Milligan and Cooper, 1985) were considered, including Silhouette coefficient (Rousseeuw, 1987) and Calinski-Harabasz index (Caliński and Harabasz, 1974). However there were no significant differences among the optimum values of $k$ experimentally provided by these indexes. Hence, it was selected one of them taking into account the computational cost. Specifically, the Silhouette coefficient was selected to validate the clustering as described in Algorithm 2.1.4 of Section 2.1.4.

As a brief remainder, the Silhouette coefficient is based on the comparison of cluster tightness and separation. It shows which objects lie well within their cluster and which ones are merely somewhere in between clusters. The average silhouette width provides an evaluation of the clustering validity and can be used to select an "appropriate" number of clusters.

For selecting the optimum $k$ according to the Silhouette coefficient, is followed the procedure

### 6.1.4 Clustering validation

When it is possible to compare the clusters obtained from the clustering algorithm to some other previously defined, the performance of our method can be measured through a confusion matrix and the metrics described in Section 2.1.6.

In many situations, it is possible to take advantage of some valuable information, as for example the provided by agro-meteorological stations or the real land delimitation. If agro-meteorological stations are available in each representative zone, it is possible to use their geolocation to identify the extension of each MZ using a Voronoi partition (Voronoï, 1908) and then compare the results of the tessellation with the results of the clustering by means of a confusion matrix (Kohavi and Provost, 1998).

## 6.2 Results of identification of agricultural management zones

In this section it is evaluated which feature (satellite band or vegetation index) better identifies MZs for the vineyards of Terras Gauda considered for the study case (see Section 1.3).

### 6.2.1 Data collection, transformation and dataset generation

The satellite data were obtained from EarthExplorer downloading the OLI and TIRS data products of Landsat 8 for the region of the vineyards of Terras Gauda. The data collected correspond to April 2013 to September 2015.

To process, extract and transform the data from the GeoTIFF files, it was followed the process described on Appendix A.2 and a dataset was generated with the structure explained on Table 5.1.

### 6.2.2 Clustering and management zone identification

PAM algorithm was ran for each plot and feature of the dataset, as described in Subsection 6.1.3. A different land clustering is obtained depending on the feature used as input. Regarding the number of clusters, no previous information about the number of MZs is considered (Case a).

According to the quality index, the Silhouette, the thermal infrared bands B10 and B11 produce the best clusters for the three vineyard plots (see Fig.6.2 and Fig.6.3). However, the behaviour of these two bands decreases more sharply (with regard to the number of clusters) in Plot 3 than in the other two.

On the other hand, the vegetation and moisture indices for the three parcels do not reach the quality of the clusters produced by B10 and B11 bands. In particular, NDVI does not show any quality improvement compared to the other features. Fig.6.4 shows the performance of land delimitation when B10 and B11 and NVDI (which is the most used index in this framework). It is easy to check that the clustering produced by both bands clearly outperforms the obtained with NVDI.

Regarding the selection of the pair $(feature, k)$, Fig.6.4 shows peaks corresponding to $(NDVI, 4)$, $(B10, 4)$, $(B10, 8)$ and $(B10, 10)$ for Plot 1. Band 11 has also peaks values for $k = 4, 8$ and both bands have the maximum value in $k = 8$. Taking this value of k, the cluster distribution generated by B10 and B11 is similar as it is shown in Fig 6.5 (top part).

The quality indices of the thermal bands for Plot 2 (see figure at middle of Fig.6.4) have it maximum value for $k = 15$ whilst for NDVI the maximum is reached when $k = 7$. Other peaks correspond to $(B10, 9)$ and $(B11, 6)$. Regarding the MZs delineation for $k = 15$, Fig.6.5 (bottom left part) shows different cluster distribution for B11 and for B12 with the exception of the MZs labelled as 12 to 15.

Finally, for Plot 3, Fig.6.4 (bottom part) shows higher quality values for $(NDVI, 4)$, $(B10, 5)$ and $(B11, 4)$. Considering the thermal bands, although the band B10 considers one MZ more than the B11, Fig.6.5 (bottom right part) shows that the delineations of the MZs are similar to

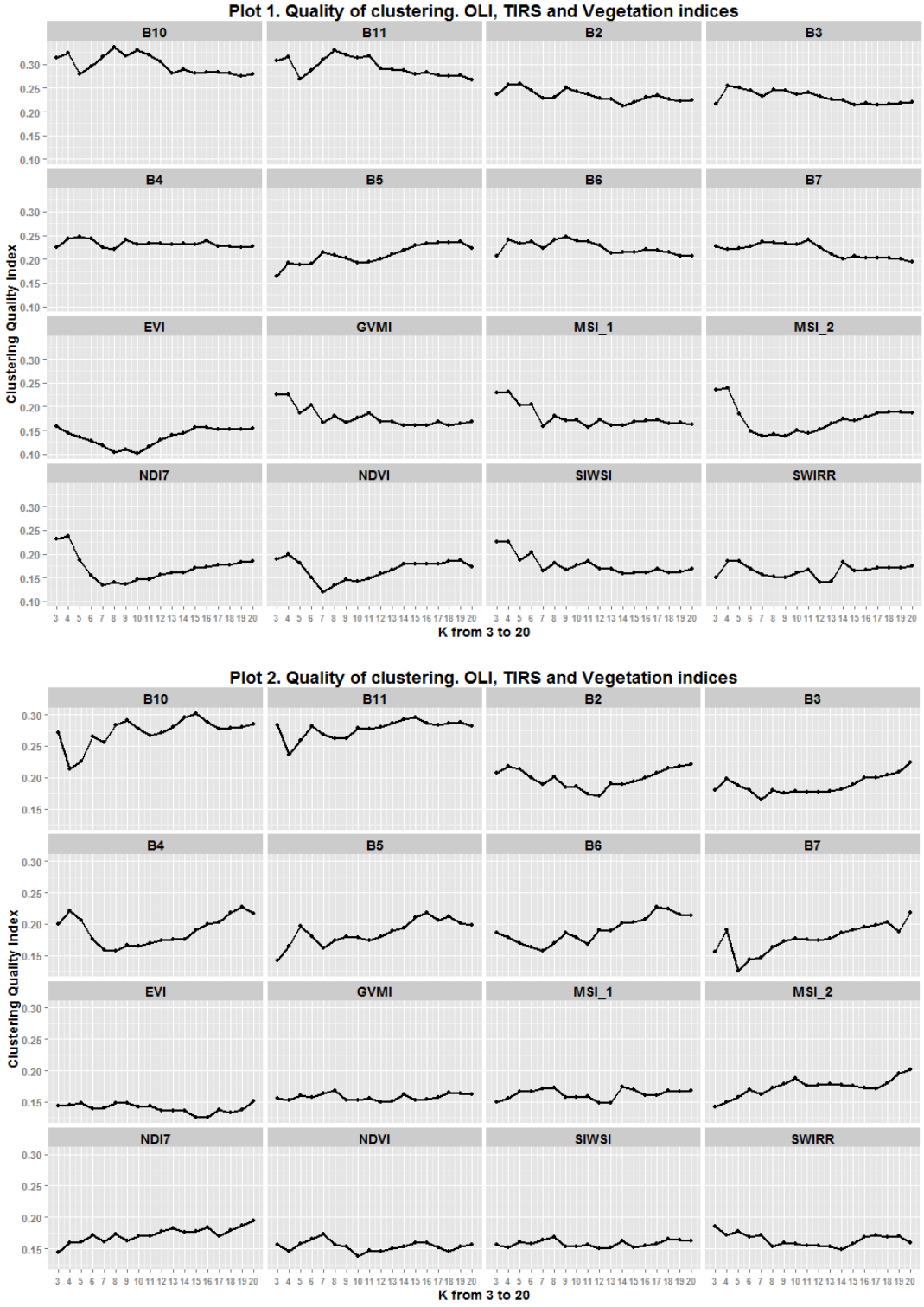



Figure 6.2: Silhouette index for Plot 1 (top) and for Plot 2 (bottom). Each graph represents the behaviour of the clustering associated to each band and vegetation index when the number of clusters ranges from 3 to 20. Thermal infrared bands $B10$ and $B11$ produce clusters with better quality than the other ones

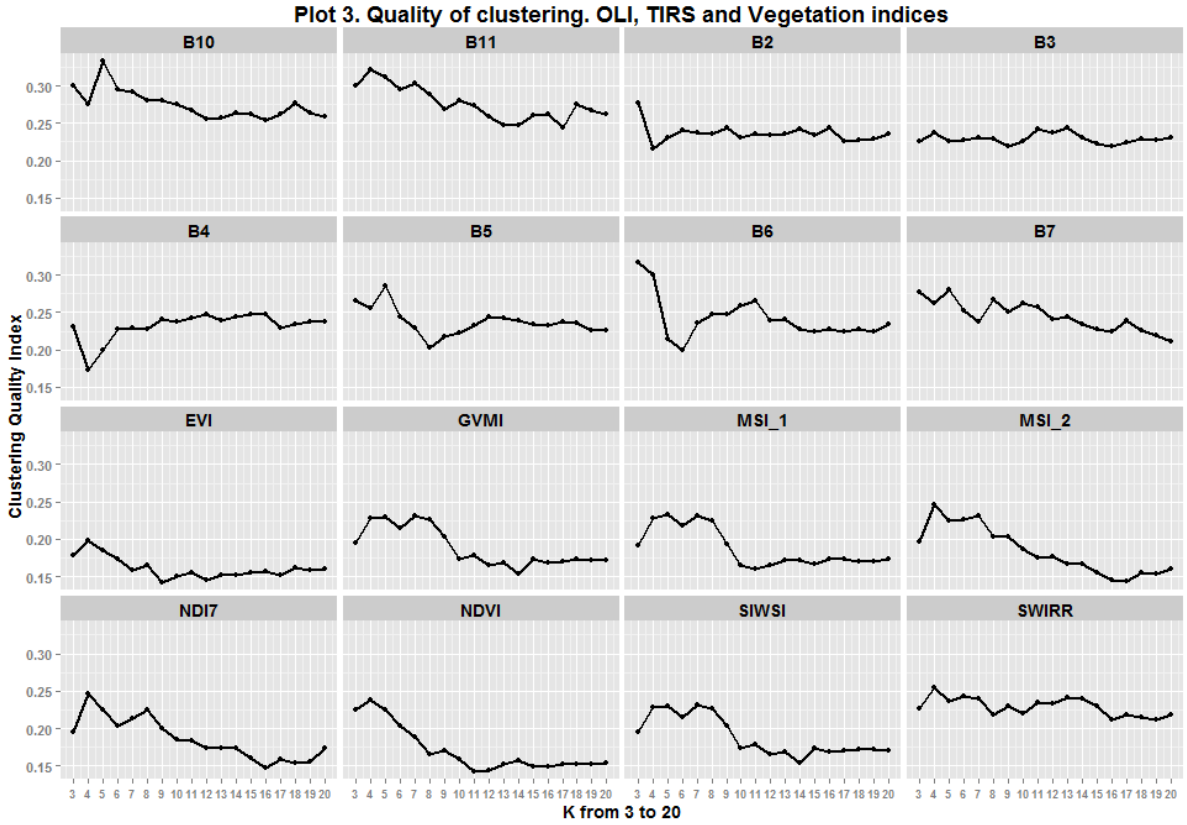



Figure 6.3: The silhouette index for Plot 3 shows that Thermal infrared bands $B10$ and $B11$ produce clusters with better quality than those from spectral bands and vegetation indices.

each other.

### 6.2.3 Validation of management zones

In order to verify the results for the MZ delimitation (called $MG_C$) it is considered a manual zone identification in the Terras Gauda's vineyard where zones are identified with the aim of deploying a station in each zone. Each station would be able to register the temperature of the air, the relative humidity and the hours of leaf wetness. The total number of stations and their location are based on the experience and knowledge of the technical director of Terras Gauda (this MZs delimitation is called $MG_{TG}$ from now on).

Thus, the number of MZs proposed by $MG_C$ method for each plot is compared to the total number of stations provided by $MG_{TG}$ delimitation for the same plot. Second, the MZs delimitation for each plot is obtained by means of the clustering of thermal bands B10 and B11, considering k the number of stations. Finally, the geographic delimitation of the MZs provided by the thermal bands ($MG_C$) and the delimitation of the MZs provided by Terras Gauda expert $MG_{TG}$ are compared by means of a confusion matrix (Kohavi and Provost, 1998). To address such comparison, from the geolocation of the agro-meteorological stations the delineation of each MZ is obtained by means of a Voronoi partition (Voronoï, 1908).

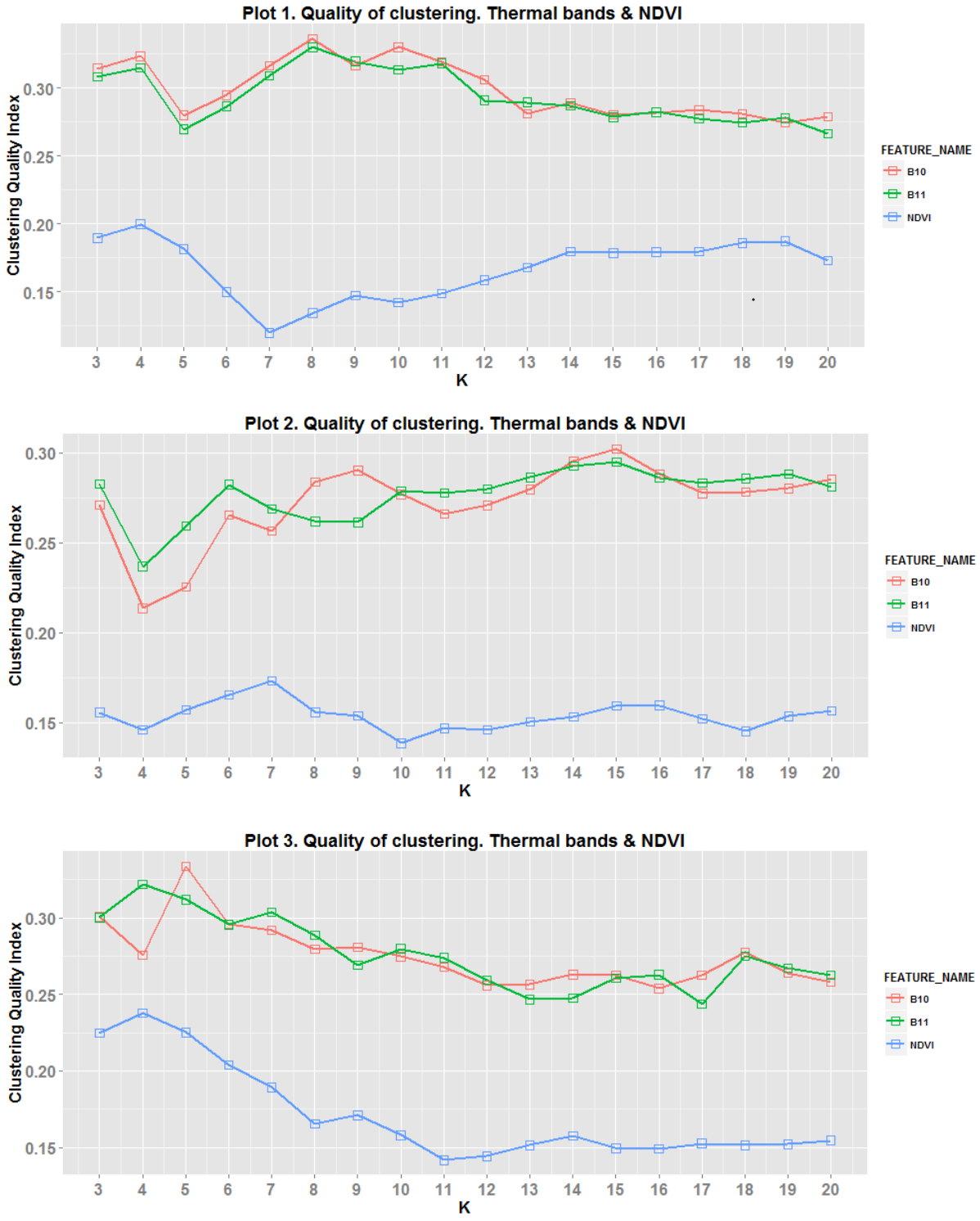



Figure 6.4: Value of Silhouette for Plot 1 (top), Plot 2 (middle) and Plot 3 (bottom). $X$ axis represents the number of clusters 3 to 20. $Y$ axis contains the value of Silhouette index.

### Validation of the number of management zones

Fig.6.6 shows the comparison between $MG_C$ and $MG_{TG}$. $MG_C$ method identifies 8 MZs whilst the number of stations considered with $MG_{TG}$ is 15, approximately one station per each 1.5

hectares. The increase in this number of stations can be associated with the high incidence of the Plasmopara viticola disease and the need of more stations for the early detection of diseases in strategic locations.

Regarding Plot 2, the number of stations considered in $MG_{TG}$ is 10 but $MG_C$ clustering method identifies 15 MZs. Fig.6.8 shows that the MZs obtained by $MG_{TG}$ are also identified by $MG_C$ clustering method. However, in some cases, $MG_C$ is able to recognize two or three MZs meanwhile $MG_{TG}$ considers only one.

Finally, the number MZs estimated for Plot 3, $(B10, 5)$ and $(B11, 4)$ with $MG_C$ is close to the 6 stations considered by $MG_{TG}$.

**Validation of the geographic delimitation**

As it was previously mentioned, the clustering produced by $MG_C$ and $MG_{TG}$ is compared. On the one hand, we considered the geographic delimitation of the MZs proposed by the clustering method ($MG_C$), taking $k$ as the number of the stations. On the other hand, the MZ delimitation from the Voronoi partition based on the geolocation of the stations ($MG_{TG}$). Fig.6.7, Fig.6.8 and Fig.6.9 respectively show the Voronoi tessellation for plots 1, 2 and 3, including the MZ delimitation generated by $MG_C$. In order to compare the results of the tessellation ($MG_{TG}$) with the results of the clustering ($MG_C$) the confusion matrix and Precision, Recall and Accuracy are computed (see Subsection 6.2.3).

Fig.6.10 shows the box plots with the metrics related to the three parcels considered in this study. The results show high values for Accuracy with medians above the 87% for both thermal bands. Precision has the values concentrated around the 60% and 80% and Recall around the 70%.

In general, the medians of the metrics for B10 and B11 are around the same values for all the plots with the exception of the precision in Plot 3, a 5% higher for B10.

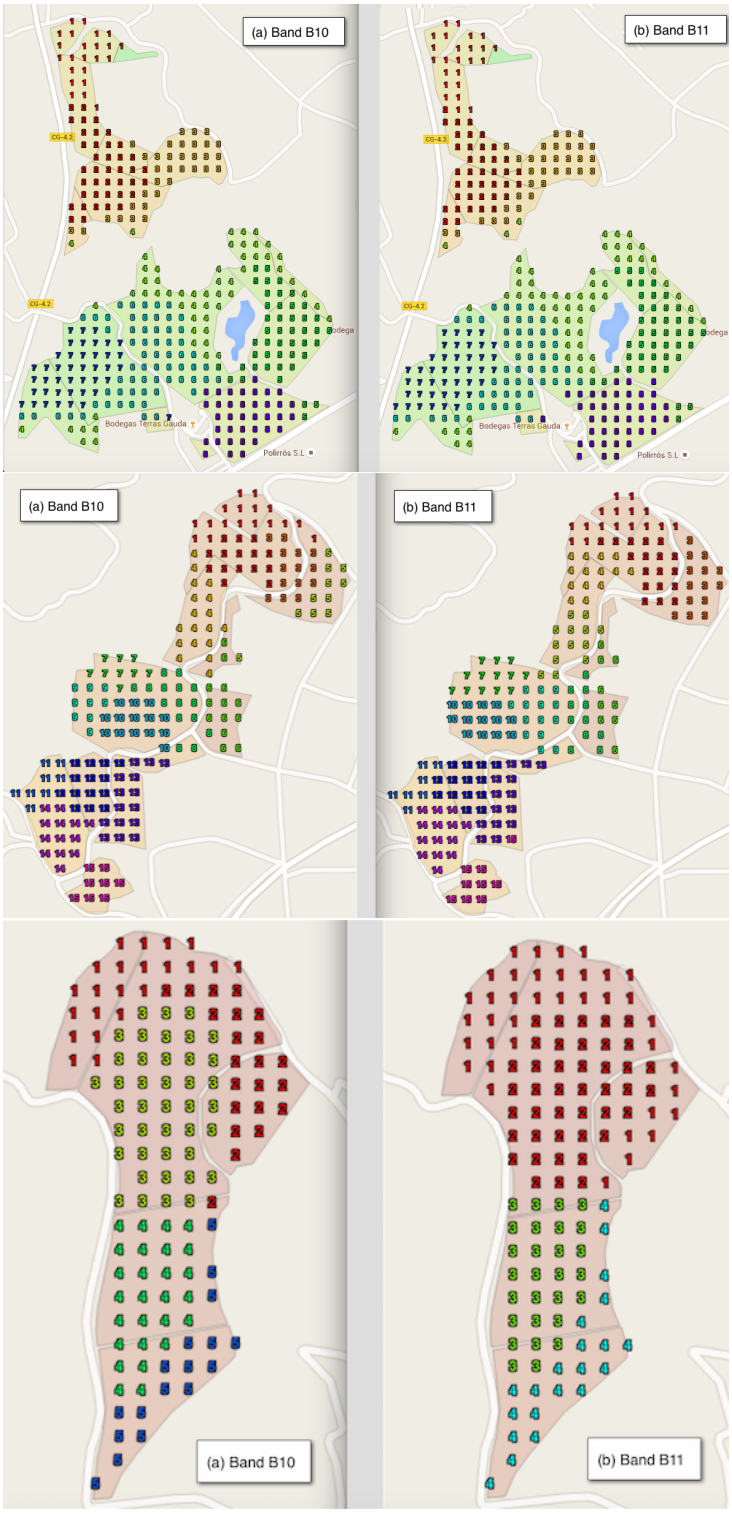



Figure 6.5: Clustering and MZs distribution. Each spatial data point is labelled according the number of the cluster to which it belongs. Plot 1 (top) considering: (a) band B10 with k=8, (b) band B11 with k=8; Plot 2 (middle left) considering: (a) band B10 with k=15, (b) band B11 with k=15; Plot 3 (middle right) considering: (a) band B10 with k=5, (b) band B11 with k=4.

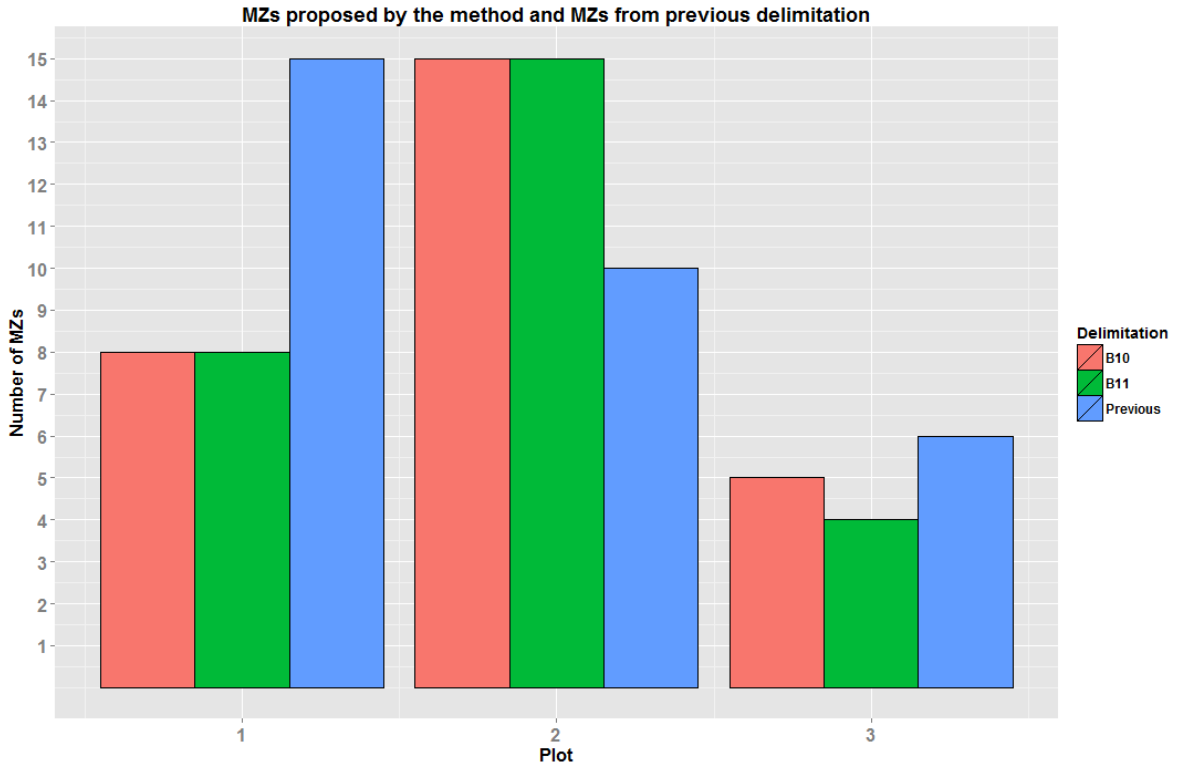



Figure 6.6: Number of MZs comparing the outcome of the clustering method with a previous delimitation based on a study for the deployment of an agro-meteorological network. $X$ axis represents the number of plot while $Y$ axis contains the number of MZs.

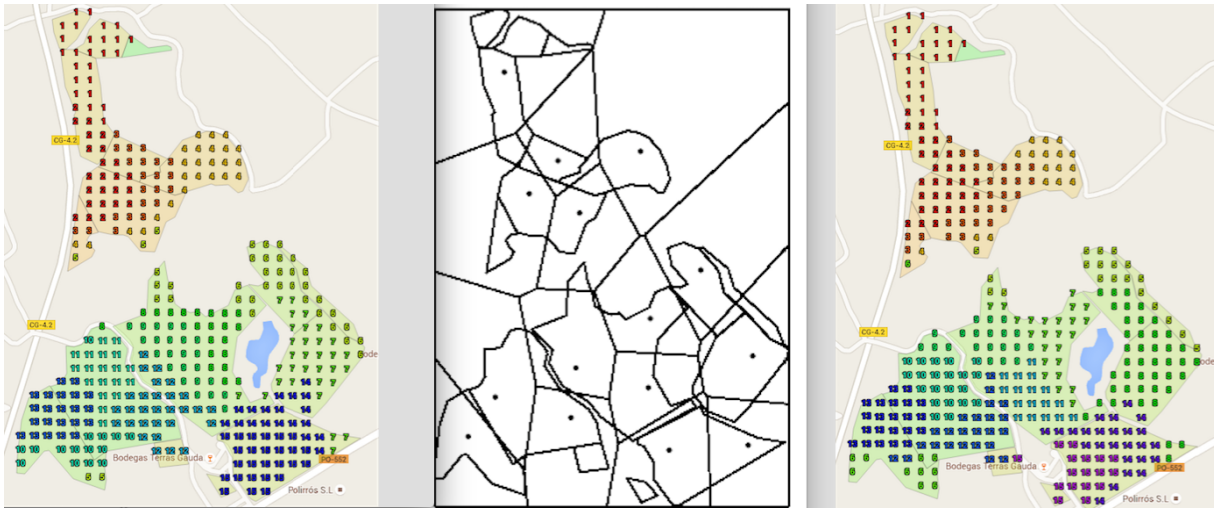


Figure 6.7: The clustering delimitation of Parcel 1 for 15 MZs and considering the band B10 (left) and the band 11 (right) has a spatial distribution similar to the one obtained with the Voronoi tessellation (middle)

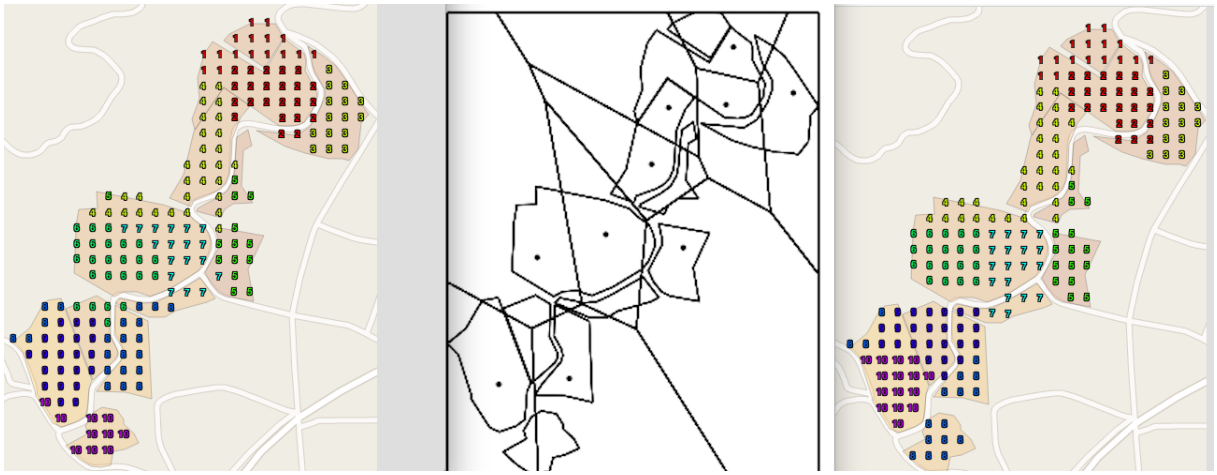


Figure 6.8: The clustering delimitation of Parcel 2 for 10 MZ considering the band B10 (left) and the band 11 (right) has a spatial distribution similar to the one obtained with the Voronoi tessellation (middle)

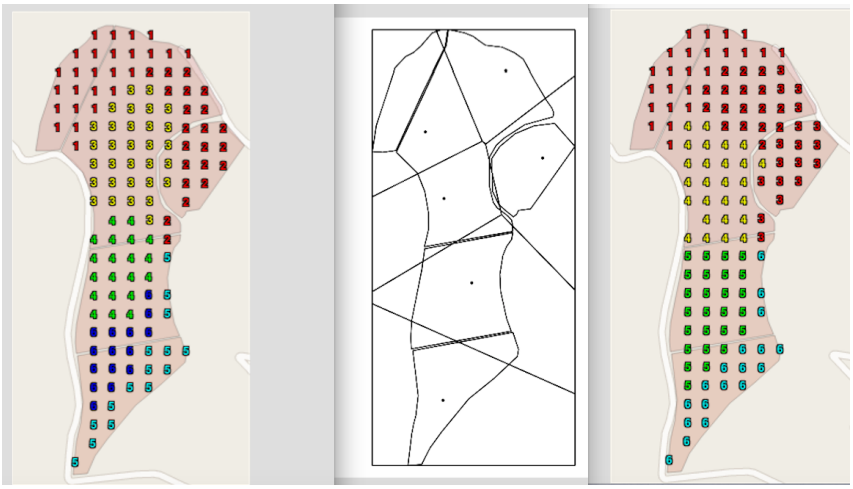


Figure 6.9: The clustering delimitation of Parcel 3 for 6 MZ considering the band B10 (left) and the band 11 (right) has a spatial distribution similar to the one obtained with the Voronoi tessellation (middle)

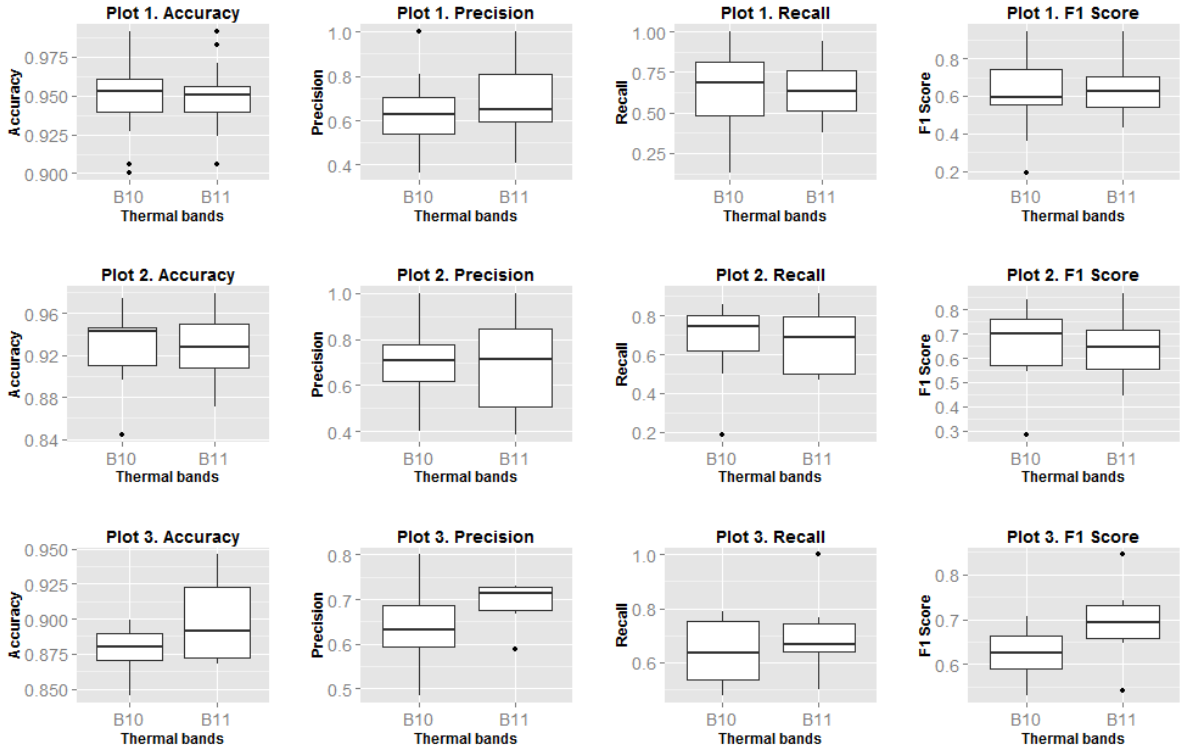



Figure 6.10: Accuracy, Precision, Recall and F1 of the MZs obtained using $M_C$ with thermal bands B10 and B11

# Chapter 7

# APPLYING SUPERVISED LEARNING ALGORITHMS TO GRAPEVINE YIELD FORECASTING

Early and very early grape yield prediction allows winegrowers to make decisions such as those concerning logistics and consider alternative grape providers in order to maintain the expected wine production. Yield prediction is also valuable as an indicator for the quality of the final product, managed by means of agricultural practices for controlling the vigour of the vines. However, the procedure for the estimation of grape production is often a time-consuming method that requires counting the number of inflorescences or the number of bunches per vine (Murisier et al., 1986) and also requires to estimate the average weight of the bunches. On the other hand, the results of this kind of methods strongly depends on the randomness and representativeness of the samples (Wulfsohn et al., 2012) regarding the in-field variability. Hence, more time-efficient methods are required to obtain early grape yield predictions.

This Chapter studies models of grape yield forecasting based on SML algorithms, considering heterogeneous data sources such as soil analysis, meteorological variables associated to phenological stages (Baggiolini, 1952; Eichhorn et al., 1977), satellite imagery, crop production and viticultural climatic indices (Tonietto and Carbonneau, 2004). The work is conducted by means of a case of study based on 10 years of data from the plots of Terras Gauda.

The experimental results show soil composition, in particular the content of potassium (K), as one of the most important variables, in addition to satellite data collected in the phenological stages: flowering begins and berries begin to soften.

The Chapter is organized as follows. Section 7.1 explains the data sources used for the prediction model of the study case and summarize basic statistics of the data. Section 7.2 explain the techniques used in this work for modelling yield grape prediction. Finally, Section 7.3 exposes

the results of the case of study and shows the model obtained and the regression equations associated to each node are shown in order to apply the yield forecasting.

## 7.1 Drawing computationally manageable data

One of the first steps in decision support processes and data analysis is processing the data to be used. In this case were acquired heterogeneous data sources for a period of ten years, from 2004 to 2015, inclusive:

- Soil analysis of the MZs (see Section 7.1.1)
- Phenological stages of the vines (see Section 7.1.2)
- Meteorological variables (see Section 7.1.3)
- Bioclimatic indices (see Section 7.1.4)
- Satellite imagery (see Section 7.1.5)
- Crop production (see Section 7.1.6)

With these data sources it is studied a model for grape yield forecast as described on Section 7.2. The following Sections review the aforementioned data sources.

### 7.1.1 Soil analysis

Data were drawn from soil analysis regarding the chemical composition of the soil for the MZs included on the plots of the case of study. Table 7.2 shows statistics regarding the data used in this case of study.

| | P (mg/kg) | K (mg/kg) | Mg (cmol/kg) | Ca (cmol/kg) | pH Water | Org. Matter (%) |
|---|---|---|---|---|---|---|
| Minimum | 0.00 | 0.07 | 0.22 | 1.70 | 4.70 | 0.00 |
| Maximum | 590.00 | 0.85 | 4.52 | 14.51 | 8.36 | 11.40 |
| 1. Quartile | 20.12 | 0.43 | 0.96 | 6.01 | 5.62 | 5.47 |
| 3. Quartile | 32.00 | 0.56 | 1.91 | 8.32 | 6.12 | 7.80 |
| Mean | 32.56 | 0.49 | 1.57 | 7.23 | 5.89 | 6.62 |
| Median | 26.00 | 0.49 | 1.50 | 7.23 | 5.89 | 6.62 |
| Stdev | 42.34 | 0.12 | 0.76 | 2.13 | 0.40 | 1.76 |

Table 7.1: Basic statistics for the variables related with the composition of the soil. Years 2004 to 2015

### 7.1.2 Phenological stages of the grapevine

The phenological stages describes the developmental stages of the grapevine (see Section3.3.5). This case of study considers the dates of the following ones according to the codification of Coombe (1995):

- Stage 19. Flowering begins.
- Stage 23. Full bloom.
- Stage 32. Beginning of bunch closure.

Table 7.2 shows the statistical data related with the phenological stages 19, 23 and 32; and harvesting begins. The values of these variables are expressed on the number of days past from the 1st of January.

| | Stage 19 (days) | Stage 23 (days) | Stage 32 (days) | Harvesting begins (days) |
|---|---|---|---|---|
| Minimum | 4.80 | 137.00 | 190.00 | 248.00 |
| Maximum | 175.00 | 179.00 | 210.00 | 287.00 |
| 1. Quartile | 134.00 | 139.00 | 199.50 | 259.50 |
| 3. Quartile | 147.00 | 154.00 | 202.00 | 266.00 |
| Mean | 143.76 | 150.48 | 200.22 | 263.64 |
| Median | 142.00 | 148.00 | 200.22 | 264.00 |
| Stdev | 13.23 | 11.07 | 4.78 | 8.81 |

Table 7.2: Basic statistics for the variables related with the phenological stages of the vine. Years 2004 to 2015

### 7.1.3 Meteorological variables

The public agro-meteorological station located at Terras Gauda was used for generate the meteorological dataset. In particular, the variables enumerated on Appendix B.1 were considered.

A total of twenty two meteorological variables were acquired for the harvesting begins and the dates of the phenological stages 19, 23 and 32; as well as for a seven-day time window centred on each one of these dates for each variable, as the average value of the time window for each variable. For instance, if stage 19 on the year 2014 was on May 21th, the period of time considered for the calculation of the average of the variables was from May 18th to May 24th, three days before and after the considered date.

The variables were named as follows. $A1 \dots A22$ and $E1 \dots E22$ for the meteorological variables associated to the dates of the stage 19 and for the seven-day window centred on those dates, respectively. The same information associated to the stage 23 is on variables $B1 \dots B22$ y $F1 \dots F23$. The variables for the stage 32 are $C1 \dots C22$ y $G1 \dots G22$ and for harvesting begins $D1 \dots D22$ y $H1 \dots H22$.

Tables 7.3, 7.4 and 7.5 show the statistical the meteorological variables associated to the dates of the considered phenological stages.

| Variable | Minimum | Maximum | 1. Quartile | 3. Quartile | Mean | Median | Stdev |
|---|---|---|---|---|---|---|---|
| A1 | 12.50 | 20.90 | 15.55 | 18.40 | 16.90 | 16.70 | 2.11 |
| A2 | 17.90 | 30.40 | 20.55 | 26.00 | 23.29 | 23.00 | 3.65 |
| A3 | 6.00 | 13.50 | 9.60 | 13.10 | 11.00 | 11.90 | 2.21 |
| A4 | 43.70 | 89.00 | 59.00 | 84.00 | 73.94 | 78.00 | 14.29 |
| A5 | 0.00 | 71.00 | 0.00 | 24.00 | 18.47 | 6.80 | 22.08 |
| A6 | 66.00 | 99.00 | 83.00 | 99.00 | 91.33 | 98.00 | 12.21 |
| A7 | 21.00 | 68.00 | 38.00 | 63.00 | 50.37 | 49.00 | 13.83 |
| A8 | 6.90 | 12.40 | 10.22 | 11.30 | 10.22 | 10.22 | 1.52 |
| A9 | 16.20 | 20.70 | 16.60 | 20.40 | 18.55 | 19.10 | 1.74 |
| A10 | 16.40 | 21.00 | 19.28 | 19.50 | 19.28 | 19.28 | 1.04 |
| A11 | 1.37 | 6.37 | 2.20 | 3.53 | 3.02 | 3.02 | 1.28 |
| A12 | 0.00 | 7.20 | 0.00 | 0.00 | 1.10 | 0.00 | 2.33 |
| A13 | 17.03 | 28.44 | 17.86 | 21.13 | 21.13 | 21.13 | 2.99 |
| A14 | 4.70 | 12.40 | 10.65 | 12.10 | 10.65 | 10.65 | 1.95 |
| A15 | 1160.00 | 2993.00 | 1722.00 | 2835.00 | 2148.69 | 2025.00 | 564.04 |
| A16 | 19.00 | 321.00 | 155.43 | 188.00 | 155.43 | 155.43 | 70.73 |
| A17 | 0.10 | 156.20 | 30.10 | 55.59 | 55.59 | 55.59 | 40.92 |
| A18 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| A19 | 31.90 | 84.10 | 73.40 | 83.80 | 73.40 | 73.40 | 13.57 |
| A20 | −1.00 | 315.00 | 0.00 | 315.00 | 167.62 | 180.00 | 130.22 |
| A21 | −6.90 | 1.60 | −6.90 | −4.89 | −4.89 | −4.89 | 2.16 |
| A22 | 3.60 | 6.90 | 5.93 | 6.90 | 5.93 | 5.93 | 0.85 |

Table 7.3: Basic statistics for meteorological data associated to the phenological stage 19 (Flowering begins) Years 2004 to 2015

### 7.1.4 Bioclimatic indices

This case of study includes the following viticultural climatic indices which are calculated for each year as described in Section 3.2.5.

- Average temperature for the active period of vegetation.
- Average of maximum and minimum temperatures for the active period of vegetation.
- Frosts (FD) calculated as the number of days of the year with average temperature below 0°
- Number of days with maximum temperature above 25° and above 30° (ND25, ND30).

| Variable | Minimum | Maximum | 1. Quartile | 3. Quartile | Mean | Median | Stdev |
|---|---|---|---|---|---|---|---|
| B1 | 11.70 | 23.90 | 14.60 | 19.10 | 17.17 | 17.30 | 3.38 |
| B2 | 16.30 | 30.90 | 18.05 | 27.45 | 22.87 | 22.30 | 5.02 |
| B3 | 7.20 | 18.60 | 11.30 | 13.15 | 12.28 | 12.60 | 2.76 |
| B4 | 46.00 | 95.00 | 76.00 | 83.00 | 77.01 | 79.00 | 12.87 |
| B5 | 0.00 | 77.00 | 0.00 | 12.20 | 19.24 | 6.70 | 26.76 |
| B6 | 80.00 | 99.00 | 96.00 | 99.00 | 94.02 | 97.00 | 7.26 |
| B7 | 19.00 | 78.00 | 47.50 | 61.00 | 54.28 | 55.00 | 14.72 |
| B8 | 8.10 | 14.50 | 12.40 | 14.20 | 12.40 | 12.40 | 1.74 |
| B9 | 15.90 | 22.50 | 17.80 | 21.30 | 19.38 | 19.40 | 1.93 |
| B10 | 12.60 | 20.90 | 18.19 | 18.70 | 18.19 | 18.19 | 1.92 |
| B11 | 1.12 | 6.37 | 2.48 | 4.37 | 3.41 | 3.13 | 1.56 |
| B12 | 0.00 | 11.20 | 0.00 | 0.20 | 1.86 | 0.00 | 4.02 |
| B13 | 19.66 | 40.61 | 29.80 | 30.31 | 29.80 | 29.80 | 5.84 |
| B14 | 4.50 | 12.40 | 8.68 | 9.10 | 8.68 | 8.68 | 1.75 |
| B15 | 836.00 | 3035.00 | 1521.00 | 2230.00 | 1901.59 | 2042.00 | 614.45 |
| B16 | 143.00 | 318.00 | 196.00 | 203.12 | 203.12 | 203.12 | 39.67 |
| B17 | 0.10 | 178.20 | 72.40 | 92.40 | 76.72 | 76.72 | 44.93 |
| B18 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| B19 | 30.70 | 83.70 | 58.75 | 63.00 | 58.75 | 58.75 | 11.85 |
| B20 | −1.00 | 315.00 | 0.00 | 202.50 | 120.03 | 180.00 | 121.01 |
| B21 | −7.30 | 7.40 | −4.10 | −0.60 | −0.60 | −0.60 | 4.04 |
| B22 | 2.80 | 7.30 | 4.90 | 4.91 | 4.91 | 4.91 | 0.97 |

Table 7.4: Basic statistics for meteorological data associated to the phenological stage 23 (Full bloom). Years 2004 to 2015

- Active Thermal Integral.
- Thermal Index of Winkle is the sum of effective daily mean temperatures, calculated from the monthly average temperatures multiplied by days of each month during the growing season from April to October.
- Heliothermic product.
- Heliothermic index (P).
- Huglin index of helio-thermal aptitude.
- Average temperature on April, May, June and July (Tapril, Tmay, Tjune, Tjuly ).
- Average rainfall (mm) on April, May, June and July (Llapril, Llmay, Lljunw, Lljuly).
- Total annual rainfall.
- Annual rainfall for the active period of vegetation.
- Maximum rainfall (Pmax).

Table 7.6 shows the statistics of these data for the ten years considered in this study.

| Variable | Minimum | Maximum | 1. Quartile | 3. Quartile | Mean | Median | Stdev |
|---|---|---|---|---|---|---|---|
| C1 | 15.40 | 22.40 | 19.00 | 19.90 | 19.27 | 19.27 | 1.59 |
| C2 | 20.20 | 31.90 | 25.30 | 27.55 | 25.71 | 25.71 | 2.91 |
| C3 | 10.10 | 16.80 | 12.60 | 13.40 | 13.38 | 13.38 | 1.69 |
| C4 | 70.70 | 95.00 | 77.00 | 81.55 | 81.55 | 81.00 | 6.98 |
| C5 | 3.70 | 61.00 | 9.15 | 37.00 | 24.45 | 24.45 | 17.06 |
| C6 | 98.00 | 99.00 | 98.77 | 99.00 | 98.77 | 99.00 | 0.36 |
| C7 | 40.00 | 84.00 | 49.00 | 55.55 | 55.55 | 55.00 | 12.06 |
| C8 | 14.20 | 15.90 | 14.97 | 14.97 | 14.97 | 14.97 | 0.37 |
| C9 | 19.10 | 25.30 | 22.36 | 23.40 | 22.36 | 22.36 | 1.53 |
| C10 | 20.70 | 22.00 | 21.33 | 21.33 | 21.33 | 21.33 | 0.27 |
| C11 | 1.51 | 5.29 | 2.45 | 2.87 | 2.72 | 2.72 | 0.91 |
| C12 | 0.00 | 14.00 | 0.00 | 1.57 | 1.57 | 0.00 | 3.80 |
| C13 | 17.75 | 24.70 | 21.90 | 21.90 | 21.90 | 21.90 | 1.52 |
| C14 | 10.90 | 12.20 | 11.70 | 11.70 | 11.70 | 11.70 | 0.29 |
| C15 | 691.00 | 2872.00 | 1935.00 | 2648.00 | 1953.98 | 1953.98 | 634.18 |
| C16 | 192.00 | 315.00 | 272.92 | 272.92 | 272.92 | 272.92 | 29.04 |
| C17 | 0.10 | 200.00 | 37.30 | 58.77 | 58.77 | 58.77 | 47.67 |
| C18 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| C19 | 73.00 | 83.60 | 79.90 | 79.90 | 79.90 | 79.90 | 2.48 |
| C20 | −1.00 | 180.00 | 0.00 | 180.00 | 99.55 | 99.55 | 77.53 |
| C21 | −6.40 | −5.60 | −6.03 | −6.03 | −6.03 | −6.03 | 0.17 |
| C22 | 5.60 | 6.40 | 6.03 | 6.03 | 6.03 | 6.03 | 0.17 |

Table 7.5: Basic statistics for meteorological data associated to the phenological stage 32 (Berries begin to soften) Years 2004 to 2015

### 7.1.5 Satellite imagery

Satellite imagery from MODIS (see Section 3.1) was download and processed for the plots of Terras Gauda considering the same seven-day temporal window, as explained on Section 7.1.3, centred on the dates of the phenological stages 19, 23 and 32. In particular, the MOD09GQ data products from 2004 to 2015 were downloaded and processed following the procedures described on Section A.1 in order to generate a dataset with the reflectivity values of band 1, band 2 and NDVI (see Section 3.3.6)

Tables 7.7, 7.8 and 7.9 show the basic statistics for the dataset on each phenological stage.

### 7.1.6 Crop production

Regarding the historical records of yield for the varieties of grapevines considered from 2004 to 2015: Albariño, Caiño, Loureiro and Treixadura; Table 7.10 shows the basic statical data and Figure7.1 shows the sum of kg/Ha for each year and variety.

| Variable | Minimum | Maximum | 1. Quartile | 3. Quartile | Mean | Median | Stdev |
|---|---|---|---|---|---|---|---|
| GSTavg | 15.69 | 17.47 | 15.91 | 17.00 | 16.47 | 16.43 | 0.60 |
| GSTmax | 20.74 | 23.52 | 21.38 | 22.75 | 22.01 | 21.91 | 0.82 |
| GSTmin | 10.86 | 12.74 | 11.21 | 11.99 | 11.59 | 11.59 | 0.53 |
| FD | 2.00 | 16.00 | 5.00 | 9.00 | 8.23 | 8.00 | 3.94 |
| ND25 | 9.00 | 29.00 | 13.00 | 21.00 | 17.39 | 15.00 | 6.00 |
| ND30 | 7.00 | 27.00 | 7.50 | 21.00 | 14.75 | 14.00 | 7.30 |
| ATI | 3358.64 | 3738.84 | 3404.11 | 3639.63 | 3525.75 | 3516.00 | 128.32 |
| WI | 1218.64 | 1598.84 | 1264.12 | 1499.63 | 1385.75 | 1376.00 | 128.32 |
| H | 0.00 | 1484.50 | 0.00 | 1426.10 | 596.13 | 0.00 | 697.40 |
| PH | 0.00 | 2.34 | 0.00 | 1.98 | 0.85 | 0.00 | 1.00 |
| IH | 1829.10 | 2322.85 | 1921.88 | 2198.39 | 2057.08 | 2041.26 | 157.46 |
| Tapril | 10.72 | 16.79 | 12.93 | 14.37 | 13.47 | 13.51 | 1.54 |
| Llapril | 44.40 | 252.40 | 79.20 | 140.80 | 120.19 | 120.40 | 56.42 |
| Tmay | 13.89 | 18.05 | 15.24 | 16.00 | 15.62 | 15.35 | 1.12 |
| Llmay | 28.80 | 194.80 | 67.60 | 122.00 | 95.76 | 100.20 | 45.50 |
| Tjune | 16.93 | 20.08 | 18.07 | 18.79 | 18.52 | 18.28 | 0.83 |
| Lljune | 9.80 | 134.40 | 22.30 | 68.00 | 53.37 | 50.00 | 38.63 |
| Tjuly | 17.94 | 21.54 | 18.87 | 20.44 | 19.78 | 20.10 | 1.05 |
| Lljuly | 5.20 | 99.60 | 13.90 | 53.50 | 34.89 | 31.20 | 28.43 |
| Taugust | 18.49 | 21.59 | 19.01 | 21.31 | 19.96 | 19.66 | 1.11 |
| Llaugust | 3.60 | 104.40 | 13.10 | 66.20 | 43.51 | 45.20 | 32.76 |
| P | 3461.07 | 8124.53 | 4270.30 | 7355.91 | 5542.52 | 4965.43 | 1521.59 |
| Pannual | 621.20 | 1972.00 | 1078.00 | 1775.80 | 1346.15 | 1313.20 | 383.41 |
| Pgs | 293.40 | 683.40 | 381.60 | 610.80 | 480.37 | 476.20 | 116.32 |
| Pmax | 37.40 | 117.80 | 59.60 | 91.40 | 73.29 | 82.60 | 22.34 |

Table 7.6: Basic statistics for the viticultural climatic indices. Years 2004 to 2015

| Statistic | reflb01.STG19 | reflb02.STG19 | NDVI.STG19 |
|---|---|---|---|
| Minimum | 487.58 | 2518.10 | 0.12 |
| Maximum | 6198.57 | 6589.14 | 0.72 |
| 1. Quartile | 1203.95 | 3207.48 | 0.29 |
| 3. Quartile | 2990.61 | 3973.01 | 0.59 |
| Mean | 2175.12 | 3779.65 | 0.45 |
| Median | 1777.21 | 3474.31 | 0.48 |
| Stdev | 1446.28 | 891.04 | 0.16 |

Table 7.7: Basic statistics for MODIS reflectivity and NDVI associated with the stage 19 (Flowering begins) Years 2004 to 2015

## 7.2 Modelling yield forecasting

The model tree technique is based on combining decision trees with linear regression functions at the leaves. Among these learners, M5 builds multivariate linear tree-based models, producing a classification based on piecewise linear functions (Quinlan et al., 1992). This predictive mechanism is added in this work to classical statistical procedures in order to characterize the

| Statistic | reflb01.STG23 | reflb02.STG23 | NDVI.STG23 |
|---|---|---|---|
| Minimum | 781.62 | 2821.31 | 0.03 |
| Maximum | 6825.93 | 7210.63 | 0.64 |
| 1. Quartile | 1780.41 | 3659.75 | 0.26 |
| 3. Quartile | 3910.86 | 5118.07 | 0.50 |
| Mean | 3074.38 | 4460.54 | 0.37 |
| Median | 2351.57 | 4104.43 | 0.40 |
| Stdev | 1734.81 | 1130.50 | 0.16 |

Table 7.8: Basic statistics for MODIS reflectivity and NDVI associated with the stage 23 (Full bloom) Years 2004 to 2015

| Statistic | reflb01.STG32 | reflb02.STG32 | NDVI.STG32 |
|---|---|---|---|
| Minimum | 464.07 | 2381.25 | 0.19 |
| Maximum | 5497.00 | 6422.43 | 0.73 |
| 1. Quartile | 2014.27 | 3726.46 | 0.43 |
| 3. Quartile | 2290.43 | 3995.52 | 0.50 |
| Mean | 2239.36 | 3949.06 | 0.46 |
| Median | 2239.36 | 3949.06 | 0.46 |
| Stdev | 1083.20 | 725.23 | 0.12 |

Table 7.9: Basic statistics for MODIS reflectivity and NDVI associated with the stage 32 (Beginning of bunch closure) Years 2004 to 2015

| Surface (Ha) | Variety | Kg/plot | Yield (Kg/Ha) |
|---|---|---|---|
| Min. : 0.1000 | Albariño : 228 | Min. : 252.4 | Min. : 486 |
| 1st Qu.: 0.3436 | Caiño blanco: 51 | 1st Qu.: 1961.4 | 1st Qu.: 3867 |
| Median : 0.9000 | Loureiro : 72 | Median : 5832.0 | Median : 6685 |
| Mean : 1.5773 | Treixadura : 48 | Mean : 9691.9 | Mean : 7697 |
| 3rd Qu.: 2.1300 | | 3rd Qu.: 14640.5 | 3rd Qu.: 9766 |
| Max. : 12.0000 | | Max. : 55159.0 | Max. :32260 |

Table 7.10: Basic statistics for the variables related with the yield, including the surface (hectares) of the plots and the grapevine variety. Years 2004 to 2015

variables involved and to identify the most important factors affecting production. There are several techniques to predict numeric values instead of just a label. Standard regression imposes a linear relation on data so it is not quite powerful. On the other hand, Neural Networks, SVR or lazy classifiers can be quite powerful but their interpretability is low. The solution proposed in this case of study represents a trade-off between interpretability and performance.

Regression trees, have been demonstrated as suitable methods to crop yield prediction. The most common algorithms to build regression trees are CART (Breiman et al., 2001), M5 (Quinlan et al., 1992) and M5-Prime (Wang and Witten, 1996). The strategy to construct the tree is similar in all of them (ElGibreen and Aksoy, 2015).

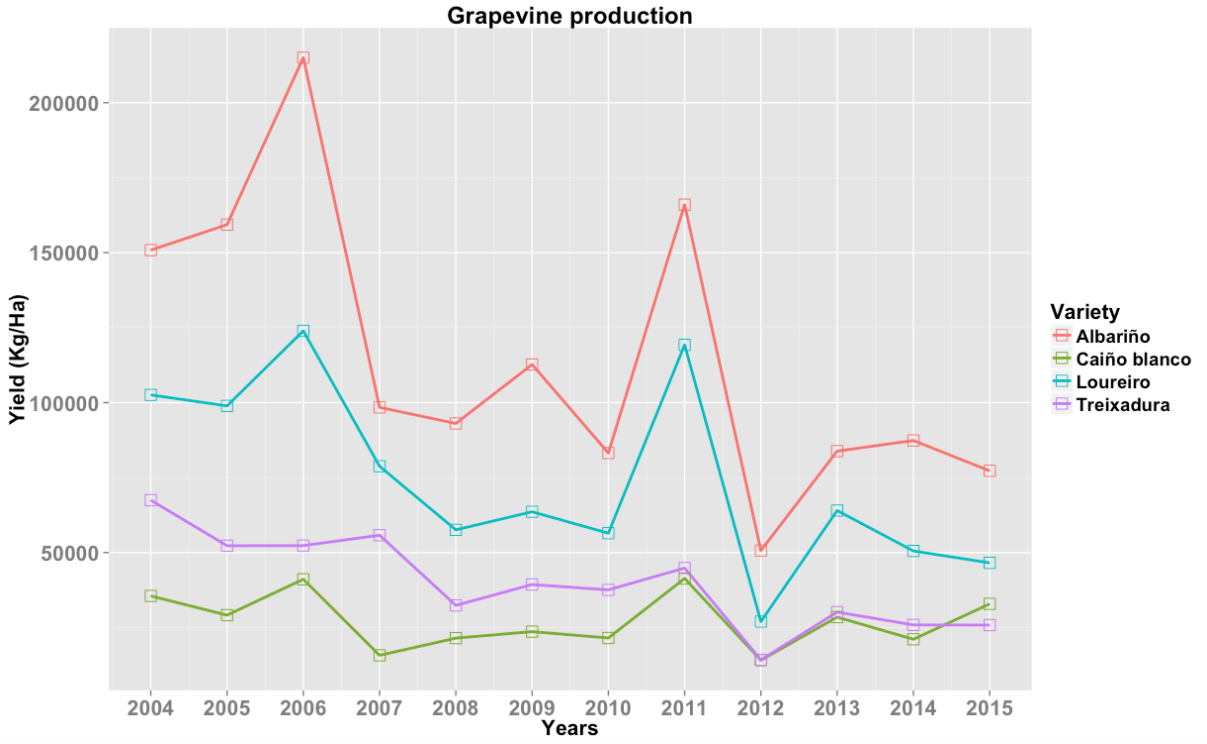



Figure 7.1: Grapevine production of Terras Gauda for the years 2004 to 2015

According to the results obtained in Gonzalez-Sanchez et al. (2014), M5-prime is the more suitable modelling tool for yield crop prediction with regard to accuracy metrics. What is more, it is more interpretable than other ML techniques, such as KNN. Thus, M5-Prime is the method selected in this work to predict grape production. This algorithm was already explained on Section 2.1.1

As production predictive tasks require the learned model to predict a numeric value associated with a variable rather than the class to which the example belongs, model regression trees are proposed. Hence, this work checks the effectiveness of ML techniques in order to determine variables affecting and classify vineyards according to production.

## 7.3 Results

The grape yield prediction for each parcel was obtained using the M5-Prime method. Other methods were also used as SVM machines or artificial NNs. However, the results obtained were far worse than the obtained with M5-Prime, regarding both efficiency and efficacy, and also decreasing interpretability. These preliminary results are consistent with previous works which state that M5-Prime is more suitable for yield prediction than other methods.

The experiments are conducted using the RWeka Package (https://cran.r-project.org/web/packages/RWeka), using M5-Prime function with the standard configuration, i.e., with pruning, smoothing and 4 being the minimum number of examples per node. It was used bootstrap resampling (see Section 2.1.6). In aggregate, the results reduce the effects of

random selection. The experiments performed here were repeated 100 times.

The accuracy of this method is studied in terms of the error in predicting production according to the following formula:

$$\frac{abs(\sum_i P_{predicted_i} - \sum_i P_{real_i})}{\sum P_{real_i}} \tag{7.1}$$

### 7.3.1 Feature selection

Taking into account that the number of variables considered in this case of study is 236 (see Section 7.1), it was studied which ones are mostly representatives for the variability of the production. On the other hand, it is also valuable to study which kind of variables affect yield production. Hence, the input variables were grouped for this study and feature selection process was done considering the following groups:

- Grape variety (G01.VARI)
- Soil composition (G02.SOIL)
- Viticultural climatic indices (G03.VITI)
- Meteorological variables for the date of phenological stage 19 (G04.METE.STG19)
- Meteorological variables for the date of phenological stage 23 (G05.METE.STG23)
- Meteorological variables for the date of phenological stage 32 (G06.METE.STG32)
- Meteorological variables for seven-day time window centred on stage 19 (G07.METE.STG19.7DAYS)
- Meteorological variables for seven-day time window centred on stage 23 (G08.METE.STG23.7DAYS)
- Meteorological variables for seven-day time window centred on stage 32 (G09.METE.STG32.7DAYS)
- Satellite variables for seven-day time window centred on stages 19, 23 and 32 (G10.SATE)

### 7.3.2 Experimental results for combinations of input variables

This Section shows the results of the experiments considering the following combinations of the aforementioned groups of variables for training the model:

- All variables
- Grape variety + Each one of the other groups.
- Grape variety + Any combination of 2 other groups.

Fig. 7.2 summarizes the results showing the errors obtained in yield forecasting using all the variables (grey line), considering only one of the aforementioned groups (pink line), and considering 2 groups (red line). It should be noted that Fig. 7.2 shows the minimum yearly error which means that it the error obtained with the best combination of variables. These combinations are shown on Table 7.11.

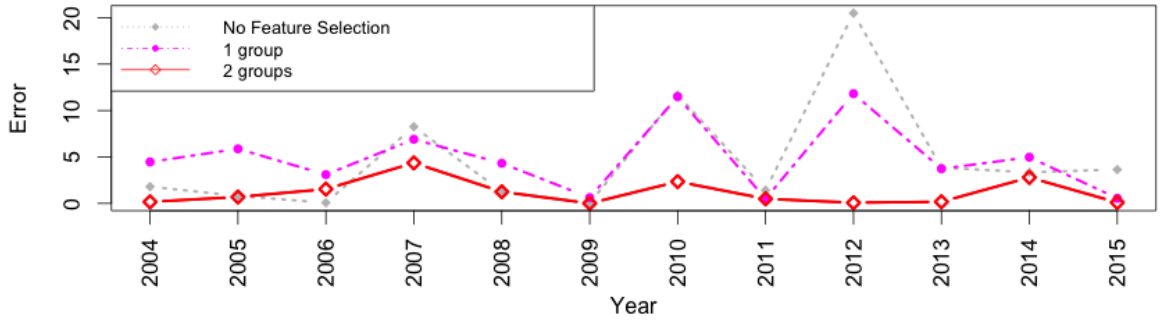



Figure 7.2: Minimum yearly error considering the best combination of variables

| Year | Best combination 1 attribute group | Best combination 2 attribute groups |
|---|---|---|
| 2004 | G05.METE.STG23 | G02.SOIL - G08.METE.STG23.7DAYS |
| 2005 | G10.SATE | G02.SOIL - G03.VITI |
| 2006 | G07.METE.STG19.7DAYS | G02.SOIL - G04.METE.STG19 |
| 2007 | G07.METE.STG19.7DAYS | G02.SOIL - G09.METE.STG32.7DAYS |
| 2008 | G04.METE.STG19 | G03.VITI - G05.METE.STG23 |
| 2009 | G08.METE.STG23.7DAYS | G02.SOIL - G05.METE.STG23 |
| 2010 | G09.METE.STG32.7DAYS | G02.SOIL - G09.METE.STG32.7DAYS |
| 2011 | G04.METE.STG19 | G04.METE.STG19 - G07.METE.STG19.7DAYS |
| 2012 | G10.SATE | G10.SATE - G06.METE.STG32 |
| 2013 | G08.METE.STG23.7DAYS | G07.METE.STG19.7DAYS - G08.METE.STG23.7DAYS |
| 2014 | G03.VITI | G10.SATE - G03.VITI |
| 2015 | G04.METE.STG19 | G04.METE.STG19 - G05.METE.STG23 |

Table 7.11: Combinations of groups of attributes that minimizes the error of the yield forecasting

The results show that the combination of two groups of attributes reduces the errors on the predictions. However, Fig. 7.3 indicates that the local minimum of each year is not extrapolable as it shown for the combination of two groups of attributes that drives to the minimum and maximum local error by year, and also in the combination which minimizes the total error. This "optimal" combination corresponds to soil composition (G02.SOIL) with satellite imagery (G10.SATE).


Figure 7.3: Comparative of local minimum, local maximum and averaged minimum for the best combinations of groups of attributes

### 7.3.3 Forecasting model considering grape variety, soil composition and satellite data

Fig. 7.4 shows the forecasting model considering the grape variety, the soil composition and the satellite attributes like independent variables. The textual mode of the M5 pruned model is the following:

```
K <= 0.445 :
|   Ca <= 5.535 :
|   |   P <= 16.5 : LM1 (30/20.586\%)
|   |   P >  16.5 :
|   |   |   pHWater <= 6.025 : LM2 (14/30.883\%)
|   |   |   pHWater >  6.025 : LM3 (4/12.094\%)
|   Ca >  5.535 :
|   |   refl_b01.STG19 <= 1880.357 :
|   |   |   refl_b02.STG19 <= 3193.357 : LM4 (23/27.49\%)
|   |   |   refl_b02.STG19 >  3193.357 : LM5 (25/43.656\%)
|   |   refl_b01.STG19 >  1880.357 : LM6 (22/42.92\%)
K >  0.445 :
```

```
|   NDVI.STG19 <= 0.425 :
|   |   AlbarinoVariety <= 0.5 :
|   |   |   refl_b02.STG19 <= 3847.029 : LM7 (11/80.375\%)
|   |   |   refl_b02.STG19 >  3847.029 :
|   |   |   |   O.M. <= 8.385 :
|   |   |   |   |   LoureiroVariety <= 0.5 : LM8 (17/32.729\%)
|   |   |   |   |   LoureiroVariety >  0.5 : LM9 (14/31.025\%)
|   |   |   |   O.M. >  8.385 : LM10 (8/26.362\%)
|   |   AlbarinoVariety >  0.5 :
|   |   |   refl_b02.STG32 <= 2929.583 : LM11 (12/10.365\%)
|   |   |   refl_b02.STG32 >  2929.583 :
|   |   |   |   refl_b01.STG32 <= 2191.929 :
|   |   |   |   |   O.M. <= 8.335 : LM12 (18/30.544\%)
|   |   |   |   |   O.M. >  8.335 : LM13 (7/8.645\%)
|   |   |   |   refl_b01.STG32 >  2191.929 : LM14 (13/17.623\%)
|   NDVI.STG19 >  0.425 :
|   |   refl_b01.STG19 <= 1138.071 : LM15 (41/62.007\%)
|   |   refl_b01.STG19 >  1138.071 :
|   |   |   AlbarinoVariety <= 0.5 :
|   |   |   |   O.M. <= 8.16 :
|   |   |   |   |   LoureiroVariety <= 0.5 : LM16 (34/51.002\%)
|   |   |   |   |   LoureiroVariety >  0.5 :
|   |   |   |   |   |   refl_b02.STG19 <= 3431.214 : LM17 (14/35.616\%)
|   |   |   |   |   |   refl_b02.STG19 >  3431.214 : LM18 (9/51.131\%)
|   |   |   |   O.M. >  8.16 : LM19 (15/62.266\%)
|   |   |   AlbarinoVariety >  0.5 : LM20 (68/78.835\%)
```

The tree shows that the most relevant variable for the discrimination of the groups is the content of potassium (K). The value of discretization of this variable is 0.45, very close to the mean. However, the content of Magnesium (Mg) is the only variable related with soil composition which does not discriminate any group. It is also noticed that the grapes varieties of Albariño and Loureiro show a different behaviour.

Regarding the satellite data, those associated to the phenological stages 19 (Flowering begins) and 32 (Beginning of bunch closure), are the most important satellite variables used in the partition.

The coefficient of correlation obtained with this combination of groups of attributes is 0.84. See Appendix B for the regression equations associated to each node in order to apply the yield forecasting.

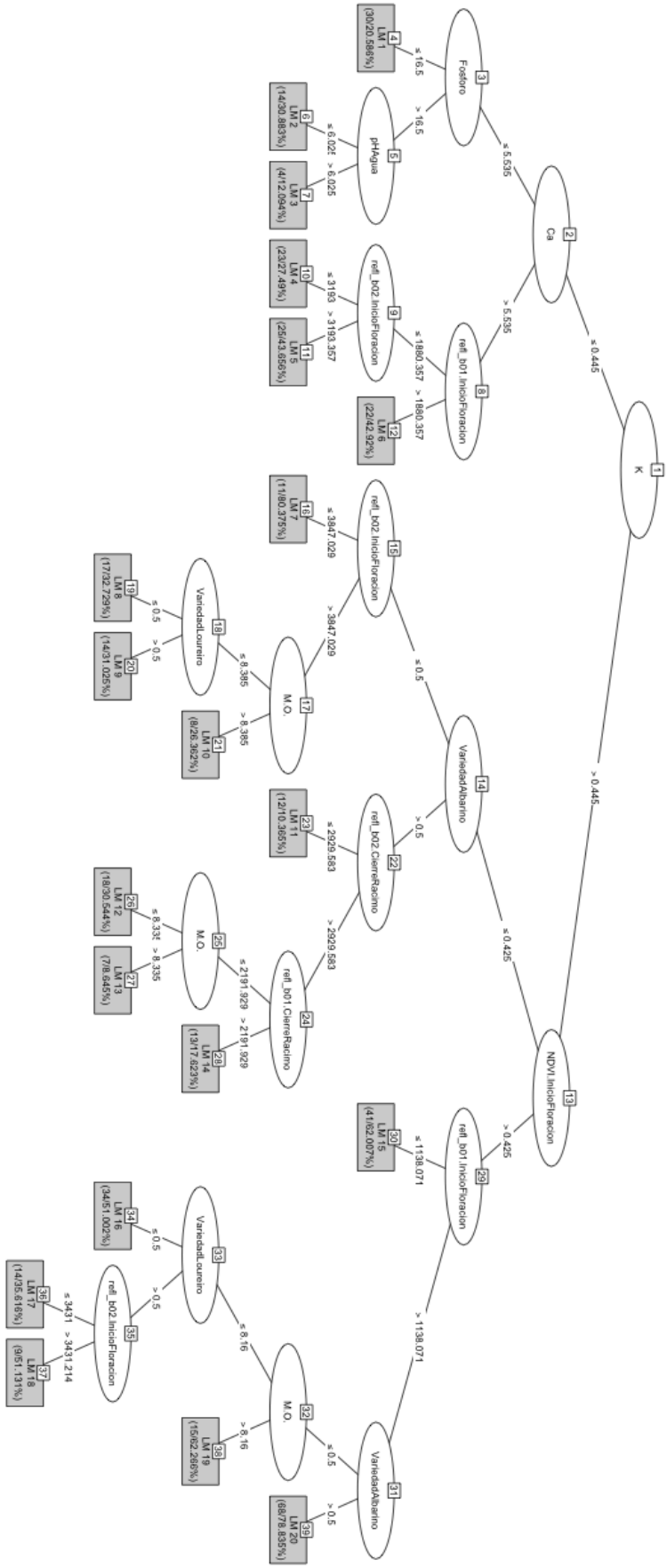



Figure 7.4: M5 pruned model tree considering grape variety, soil composition and satellite data

# Chapter 8

# CONCLUSIONS AND FUTURE WORK

This work studies and develops models for smart agro-services related with the automatic cultivable land delimitation, the automatic identification of homogeneous land zones and the forecasting of grapevine production.

The use of satellite imagery for automatic land delimitation requires high spatial resolution. However, in general, the higher the spatial resolution, the lower the temporal one. Considering this fact, it was studied how the temporal resolution affects the results of the automatic delimitation of land using satellite reflectivity and clustering algorithms (see Chapter 4). The experimental results obtained on the case of study show that satellite imagery with lower temporal resolution, as Landsat 8 has, does not affect land delimitation. Actually, the lower the resolution, the more compact the clusters of land. This work opens future lines of research as to study fusion techniques with satellite imagery of different temporal and spatial resolutions in order to improve the clustering of the land.

Regarding automatic cultivable land detection with supervised machine learning (see Section 5.1), it was studied the performance of five different SML methods in the task of arable land determination for a real case study. The results show evidence that selecting the most relevant attributes with CFS lead to classifiers that are both efficient and effective, with values of the area under the ROC curve that are above 0.95 in most cases. In particular, C5.0 seems a good choice since it performs constantly well both with and without feature selection. Future work is intended to extend this study in order to classify the type of crop based on satellite imagery and supervised machine learning.

Automatic cultivable land detection was also studied using unsupervised machine learning algorithms (see Section 5.3) and a methodology for mapping cultivable land was proposed. The experimental results show the great potential of this method for cultivable land monitoring from remote-sensed multispectral imagery. The methodology could be used for cultivable land monitoring from remote-sensed multispectral imagery, for instance regarding the census of agri-

cultural land or assisting in the process of checking the use of agricultural aids provided by governments. Future work is focused on developing feature selection techniques to improve the clustering as well as exploring different clustering approaches.

Concerning to the identification of management zones, in Section 6.1 is proposed a cost-efficient method for mapping MZs that combines clustering algorithms with publicly available satellite imagery. The method does not require exploring the parcels with any special equipment or taking samples of the soil for laboratory analysis. The results show that thermal infrared and spectral data remotely sensed by satellite instruments have great potential for land clustering. The findings also indicate that NDVI, the vegetation index usually considered for decision making with regard to the delimitation of MZs, produces clusters with a quality significantly lower than the ones obtained with thermal bands. Future work will involve the inclusion of topographical characteristics and biophysical features as data input for the clustering algorithm in order to study the improvement of the identification of MZs.

Regarding early grape yield prediction, Chapter 7 studies models of grape yield forecasting based on SML algorithms, considering heterogeneous data sources such as soil analysis, meteorological variables associated to phenological stages, satellite imagery, crop production and viticultural climatic indices. The experimental results show soil composition, in particular the content of potassium (K), as one of the most important variables, in addition to satellite data collected in the phenological stages: flowering begins and berries begin to soften.

Summing up, the potential of smart agro-services for providing solutions for planning and decision-making processes in the agricultural and environmental domains is high. Opening a scientific opportunity in this field to continue with the study and development of robust ML models relaying on heterogeneous data sources, that could help to improve the sustainable agriculture model in a eco-friendly context.

# Bibliography


Adams, R. M., Bryant, K. J., McCarl, B. A., Legler, D. M., O'Brien, J., Solow, A. and Weiher, R. (1995). Value of improved long-range weather information, *Contemporary Economic Policy* **13**(3): 10–19.

Aggelopooulou, K., Castrignanò, A., Gemtos, T. and Benedetto, D. D. (2013). Delineation of management zones in an apple orchard in Greece using a multivariate approach, *Computers and Electronics in Agriculture* **90**(C): 119–130.

Aizerman, A., Braverman, E. M. and Rozoner, L. (1964). Theoretical foundations of the potential function method in pattern recognition learning, *Automation and remote control* **25**: 821–837.

Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression, *The American Statistician* **46**(3): 175–185.

Amazon (2015). Public landsat datasets on amazon, `http://aws.amazon.com/es/public-data-sets/landsat/`.

American Society of Agronomy (ASA) (1989). Decision reached on sustainable agriculture, *Agronomy News* p. 15.

Amerine, M. A. and Winkler, A. J. (1944). *Composition and quality of musts and wines of California grapes*, University of California Berkeley.

Arango, R., Campos, A., Combarro, E., Canas, E. and Díaz, I. (2016). Mapping cultivable land from satellite imagery with clustering algorithms, *International Journal of Applied Earth Observation and Geoinformation* **49**: 99–106.

Arango, R., Díaz, I., Campos, A., Combarro, E. and Canas, E. (2015). On the influence of temporal resolution on automatic delimitation using clustering algorithms, *Appl. Math. Inf. Sci.* **9**(2L): 339–347.

Bae, J. K. and Kim, J. (2011). Combining models from neural networks and inductive learning algorithms, *Expert Systems with Applications* **38**(5): 4839–4850.

Baggiolini, M. (1952). *Stades repères de la vigne*, Station fédérale d'essais et de contrôle de semences.

Baggliolini, M. (1952). Les stades repèeres dans le développement annuel de la vigne et leur utilisation pratique, *Rev Romande Agric Vitic Arboric* **8**: 4–6.

Basak, D., Pal, S. and Patranabis, D. C. (2007). Support vector regression, *Neural Information Processing-Letters and Reviews* **11**(10): 203–224.

Basso, B., Cammarano, D. and De Vita, P. (2004). Remotely sensed vegetation indices: Theory and applications for crop management, *Rivista Italiana di Agrometeorologia* **1**: 36–53.

Bekkerman, R., El-Yaniv, R., Tishby, N. and Winter, Y. (2003). Distributional word clusters vs. words for text categorization, *Journal of Machine Learning Research* **3**(Mar): 1183–1208.

Bellman, R. E. et al. (1978). *An introduction to artificial intelligence: Can computers think?*, Boyd & Fraser Publishing Company.

Bhatti, A., Mulla, D. and Frazier, B. (1991). Estimation of soil properties and wheat yields on complex eroded hills using geostatistics and thematic mapper images, *Remote Sensing of Environment* **37**(3): 181–191.

Bingfang, W., Jihua, M., FeiFei, Z., Xin, D., Miao, Z. and Xueyang, C. (2010). Applying remote sensing in precision farming-a case study in Yucheng, *World Automation Congress*, pp. 1–6.

Bivand, R., Keitt, T. and Rowlingson, B. (2013). rgdal: Bindings for the geospatial data abstraction library, *R package version 0.8-10* .

Blackmore, B., Wheeler, P., Morris, J., Robert, P., Rust, R., Larson, W. et al. (1995). The role of precision farming in sustainable agriculture: a european perspective., *Site-specific management for agricultural systems: proceedings of Second International Conference, Minneapolis, MN, USA, March 27-30, 1994.*, American Society of Agronomy, pp. 777–793.

Blackmore, S., Godwin, R. J. and Fountas, S. (2003). The analysis of spatial and temporal trends in yield map data over six years, *Biosystems engineering* **84**(4): 455–466.

Bongiovanni, R. and Lowenberg-DeBoer, J. (2000). Nitrogen management in corn using site-specific crop response estimates from a spatial regression model, *Proceedings of the Fifth International Conference on Precision Agriculture*.

Bongiovanni, R. and Lowenberg-DeBoer, J. (2004). Precision agriculture and sustainability, *Precision agriculture* **5**(4): 359–387.

Boso, S., Santiago, J. and Martínez, M. C. (2008). The influence of 110-Ritcher and SO4 rootstocks on the performance of scions of Vitis vinifera L. cv. Albariño clones, *Spanish journal of agricultural research* **6**(1): 96–104.

Bourbakis, N. G. (1992). *Artificial Intelligence Methods and Applications*, Vol. 1, World Scientific.

Bradford, J. P., Kunz, C., Kohavi, R., Brunk, C. and Brodley, C. E. (1998). Pruning decision trees with misclassification costs, *European Conference on Machine Learning*, Springer, pp. 131–136.

Branas, J., Bernon, G., Levadoux, L. et al. (1946). A treatise on the elements of viticulture., *Elements de viticulture generale* .

Breiman, L. (1996). Bagging predictors, *Machine learning* **24**(2): 123–140.

Breiman, L., Friedman, J., Stone, C. J. and Olshen, R. A. (1984). *Classification and regression trees*, CRC press.

Breiman, L. et al. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author), *Statistical Science* **16**(3): 199–231.

Caliński, T. and Harabasz, J. (1974). A dendrite method for cluster analysis, *Communications in Statistics-theory and Methods* **3**(1): 1–27.

Cambardella, C. and Karlen, D. (1999). Spatial analysis of soil fertility parameters, *Precision Agriculture* **1**(1): 5–14.

Cardie, C. (1993). Using decision trees to improve case-based learning, *Proceedings of the tenth international conference on machine learning*, pp. 25–32.

Cassman, K. G. (1999). Ecological intensification of cereal production systems: Yield potential, soil quality, and precision agriculture, *Proceedings of the National Academy of Sciences* **96**(11): 5952–5959.

Ceccato, P., Gobron, N., Flasse, S., Pinty, B. and Tarantola, S. (2002). Designing a spectral index to estimate vegetation water content from remote sensing data: Part 1: Theoretical approach, *Remote Sensing of Environment* **82**(2): 188–197.

Chang, Y.-W., Hsieh, C.-J., Chang, K.-W., Ringgaard, M. and Lin, C.-J. (2010). Training and testing low-degree polynomial data mappings via linear SVM, *The Journal of Machine Learning Research* **11**: 1471–1490.

Chang, Y.-W. and Lin, C.-J. (2008). Feature ranking using linear SVM, *WCCI Causation and Prediction Challenge*, pp. 53–64.

Charvat, K. and Pavel, G. (2012). Using linear programming for tactical planning in agriculture in the frame of the COIN project, *in* T. Mildorf and K. C. jr. (eds), *ICT for Agriculture, Rural Development and Environment*, Czech Centre for Science and Society, pp. 300–309.

Chawla, N. V. (2003). C4.5 and imbalanced data sets: investigating the effect of sampling method, probabilistic estimate, and decision tree structure, *Proceedings of the ICML*, Vol. 3.

Chen, J. M. and Black, T. (1992). Defining leaf area index for non-flat leaves, *Plant, Cell & Environment* **15**(4): 421–429.

Chiang, T.-H., Lo, H.-Y. and Lin, S.-D. (2012). A ranking-based KNN approach for multi-label classification, *ACML* **25**: 81–96.

Chou, C.-H., Su, M.-C. and Lai, E. (2004). A new cluster validity measure and its application to image compression, *Pattern Analysis and Applications* **7**(2): 205–220.

Chou, J.-S. (2012). Comparison of multilabel classification models to forecast project dispute resolutions, *Expert Systems with Applications* **39**(11): 10202–10211.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K. and Kuksa, P. (2011). Natural language processing (almost) from scratch, *Journal of Machine Learning Research* **12**(Aug): 2493–2537.

Combarro, E. F., Montanes, E., Diaz, I., Ranilla, J. and Mones, R. (2005). Introducing a family of linear measures for feature selection in text categorization, *IEEE Transactions on knowledge and data engineering* **17**(9): 1223–1232.

Coombe, B. (1995). Growth stages of the grapevine: Adoption of a system for identifying grapevine growth stages, *Australian Journal of Grape and Wine Research* **1**(2): 104–110.

Cooter, E., Bash, J., Benson, V. and Ran, L. (2012). Linking agricultural crop management and air quality models for regional to national-scale nitrogen assessments, *Biogeosciences* **9**(10): 4023–4035.

Cormack, R. M. (1971). A review of classification, *Journal of the Royal Statistical Society* **Series A**(134): 321–367.

Cortes, C. and Vapnik, V. (1995). Support-vector networks, *Machine Learning* **20**(3): 273–297.

Corwin, D. and Lesch, S. (2005). Apparent soil electrical conductivity measurements in agriculture, *Computers and electronics in agriculture* **46**(1): 11–43.

Cutter, M. A., Lobb, D. R. and Cockshott, R. A. (2000). Compact high resolution imaging spectrometer (CHRIS), *Acta Astronautica* **46**(2): 263–268.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function, *Mathematics of control, signals and systems* **2**(4): 303–314.

DAAC, L. (1990). MODIS products table, http://lpdaac.usgs.gov/products/modis_products_table.

Day, W. H. and Edelsbrunner, H. (1984). Efficient algorithms for agglomerative hierarchical clustering methods, *Journal of classification* **1**(1): 7–24.

Dayhoff, J. and DeLeo, J. (2001). Artificial neural networks: Opening the black box, *Cancer* **91**(8): 1615.

Delegido, J., Fernandez, G., Gandia, S. and Moreno, J. (2008). Retrieval of chlorophyll content and LAI of crops using hyperspectral techniques: Application to PROBA/CHRIS data, *International Journal of Remote Sensing* **29**(24): 7107–7127.

Delgado, J., Follett, R., Buchleiter, G., Stuebe, A., Sparks, R., Dillon, M., Thompson, A. and Thompson, K. (2001). Use of geospatial information for N management and conservation of underground water quality, *The Third International Conference on Geospatial Information in Agriculture and Forestry*, pp. 5–7.

Díaz, I., Montañés, E., Ranilla, J. and Espuña-Pons, M. (2011). A framework for diagnosis of urinary incontinence disease based on scoring measures and automatic classifiers, *Computers in biology and medicine* **41**(1): 11–17.

Díaz, I., Ranilla, J., Montañes, E., Fernández, J. and Combarro, E. F. (2004). Improving performance of text categorization by combining filtering and support vector machines, *Journal of the American society for information science and technology* **55**(7): 579–592.

Ding, C. and He, X. (2002). Cluster merging and splitting in hierarchical clustering algorithms, *Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on*, IEEE, pp. 139–146.

Domingos, P. (2012). A few useful things to know about machine learning, *Communications of the ACM* **55**(10): 78–87.

Dreiseitl, S. and Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: a methodology review, *Journal of biomedical informatics* **35**(5): 352–359.

Duca, R. and Del Frate, F. (2008). Hyperspectral and multiangle CHRIS–PROBA images for the generation of land cover maps, *IEEE Transactions on Geoscience and Remote Sensing* **46**(10): 2857–2866.

Duda, R. O., Hart, P. E. et al. (1973). *Pattern classification and scene analysis*, Vol. 3, Wiley New York.

Duro, D. C., Franklin, S. E. and Dubé, M. G. (2012). A comparison of pixel-based and object-based image analysis with selected machine learning algorithms for the classification of agricultural landscapes using SPOT-5 HRG imagery, *Remote Sensing of Environment* **118**: 259–272.

Dwyer, M. J. and Schmidt, G. (2006). The MODIS reprojection tool, *Earth science satellite remote sensing*, Springer, pp. 162–177.

*EarthOnline* (2000). `https://earth.esa.int/web/guest/data-access`.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife, *Annals of Statistics* **7**: 1–26.

Efron, B. and Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation, *The American Statistician* **37**(1): 36–48.

Eichhorn, K., Lorenz, D. et al. (1977). Phenological development stages of the grape vine, *Nachrichtenblatt des Deutschen Pflanzenschutzdienstes* **29**(8): 119–120.

El-Nasr, M. S., Drachen, A. and Canossa, A. (2013). *Game analytics: Maximizing the value of player data*, Springer Science & Business Media.

ElGibreen, H. and Aksoy, M. S. (2015). Classifying continuous classes with reinforcement learning rules, *Intelligent Information and Database Systems*, Springer, pp. 116–127.

ESA (2014). Sentinel missions, `http://www.esa.int/Our_Activities/Observing_the_Earth/Copernicus/Overview4/`. Accessed: 2016-05-14.

FAO, U. et al. (2009). How to feed the world in 2050, *Rome: High-Level Expert Forum*.

Farid, D. M., Rahman, M. Z. and Rahman, C. M. (2011). Article: Adaptive intrusion detection based on boosting and naive bayesian classifier, *International Journal of Computer Applications* **24**(3): 12–19.

Fawcett, C. B. (1930). The extent of the cultivable land, *Geographical Journal* **76**(6): 504–509.

Fawcett, T. (2006). An introduction to ROC analysis, *Pattern Recogn. Lett.* **27**(8): 861–874.

Fensholt, R. and Sandholt, I. (2003). Derivation of a shortwave infrared water stress index from MODIS near-and shortwave infrared data in a semiarid environment, *Remote Sensing of Environment* **87**(1): 111–121.

Fernández-Quintanilla, C., Dorado, J., San Martín, C., Conesa-Muñoz, J. and Ribeiro, A. (2011). A five-step approach for planning a robotic site-specific weed management program for winter wheat, *Robotics and Associated High-Technologies and Equipment For Agriculture, 2011* .

Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population, *Biometrika* **10**(4): 507–521.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems, *Annals of eugenics* **7**(2): 179–188.

Frank, E., Wang, Y., Inglis, S., Holmes, G. and Witten, I. H. (1998). Using model trees for classification, *Machine Learning* **32**(1): 63–76.

Franzen, D. W., Hopkins, D. H., Sweeney, M. D., Ulmer, M. K. and Halvorson, A. D. (2002). Evaluation of soil survey scale for zone development of site-specific nitrogen management, *Agronomy Journal* **94**(2): 381–389.

Freund, Y. and Schapire, R. E. (1995). A desicion-theoretic generalization of on-line learning and an application to boosting, *European conference on computational learning theory*, Springer, pp. 23–37.

Friedl, M. A. and Brodley, C. E. (1997). Decision tree classification of land cover from remotely sensed data, *Remote sensing of environment* **61**(3): 399–409.

Friedman, N., Geiger, D. and Goldszmidt, M. (1997). Bayesian network classifiers, *Machine learning* **29**(2-3): 131–163.

*FSelector R Package* (n.d.). `https://cran.r-%project.org/web/packages`. Accessed: 2015-07-21.

Fu, Q., Wang, Z. and Jiang, Q. (2010). Delineating soil nutrient management zones based on fuzzy clustering optimized by PSO, *Mathematical and Computer Modelling* **51**(11–12): 1299–1305. Mathematical and Computer Modelling in Agriculture.

Gislason, P. O., Benediktsson, J. A. and Sveinsson, J. R. (2006). Random forests for land cover classification, *Pattern Recognition Letters* **27**(4): 294–300.

Godwin, R. and Miller, P. (2003). A review of the technologies for mapping within-field variability, *Biosystems engineering* **84**(4): 393–407.

Goldstein, I. and Papert, S. (1977). Artificial intelligence, language, and the study of knowledge*,, *Cognitive Science* **1**(1): 84–123.

Gonzalez-Sanchez, A., Frausto-Solis, J. and Ojeda-Bustamante, W. (2014). Predictive ability of machine learning methods for massive crop yield prediction, *Spanish Journal of Agricultural Research* **12**(2): 313–328.

Griepentrog, H.-W. and Kyhn, M. (2000). Strategies for site specific fertilization in a highly productive agricultural region, *Proceedings of the 5th International Conference on Precision Agriculture*, Citeseer.

Group, H. et al. (2000). Hierarchical data format version 5, `http://www.hdfgroup.org/HDF5`.

Gualtieri, J. A. and Cromp, R. F. (1999). Support vector machines for hyperspectral remote sensing classification, *The 27th AIPR Workshop: Advances in Computer-Assisted Recognition*, International Society for Optics and Photonics, pp. 221–232.

Guerrero, J., Pajares, G., Montalvo, M., Romeo, J. and Guijarro, M. (2012). Support vector machines for crop/weeds identification in maize fields, *Expert Systems with Applications* **39**(12): 11149–11155.

Guo, H., Wang, L., Chen, F. and Liang, D. (2014). Scientific big data and digital earth, *Chinese Science Bulletin* **59**(35): 5066–5073.

Gurrutxaga, I., Muguerza, J., Arbelaitz, O., Pérez, J. M. and Martín, J. I. (2011). Towards a standard methodology to evaluate internal cluster validity indices, *Pattern Recognition Letters* **32**(3): 505–515.

Gustafson, D. E. and Kessel, W. C. (1978). Fuzzy clustering with a fuzzy covariance matrix, *Decision and Control including the 17th Symposium on Adaptive Processes, 1978 IEEE Conference on*, Vol. 17, IEEE, pp. 761–766.

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection, *Journal of Machine Learning Research* **3**(Mar): 1157–1182.

Guzella, T. S. and Caminhas, W. M. (2009). A review of machine learning approaches to spam filtering, *Expert Systems with Applications* **36**(7): 10206–10222.

Haba, M., Mulet, A. and Berna, A. (1997). Stability in wine differentiation of two close viticultural zones, *American Journal of Enology and Viticulture* **48**(3): 285–290.

Hadsell, R., Sermanet, P., Ben, J., Erkan, A., Scoffier, M., Kavukcuoglu, K., Muller, U. and LeCun, Y. (2009). Learning long-range vision for autonomous off-road driving, *Journal of Field Robotics* **26**(2): 120–144.

Halkidi, M., Batistakis, Y. and Vazirgiannis, M. (2001). On clustering validation techniques, *Journal of Intelligent Information Systems* **17**(2): 107–145.

Hall, M. A. and Smith, L. A. (1997). Feature subset selection: a correlation based filter approach.

HAN, N., WU, J., Tahmassebi, A. R. S., wei XU, H. and WANG, K. (2011). NDVI-Based lacunarity texture for improving identification of torreya using object-oriented method, *Agricultural Sciences in China* **10**(9): 1431–1444.

Hartigan, J. A. (1975). *Clustering Algorithms*, 99th edn, John Wiley & Sons, Inc., New York, NY, USA.

Hatfield, J. L. (2000). *Precision agriculture and environmental quality: Challenges for research and education*, USDA National Resources Conservation Service.

Heisel, T., Christensen, S. and Walter, A. (1996). Weed managing model for patch spraying in cereal, *Proceedings of the 3rd International Conference on Precision Agriculture*, American Society of Agronomy, Crop Science Society of America, Soil Science Society of America, pp. 999–1007.

Hennig, C. (2010). fpc: Flexible procedures for clustering, *R package version* **2**: 0–3.

Hergert, G., Ferguson, R., Gotway, C. and Peterson, T. (1996). The impact of variable rate N application on N use efficiency of furrow irrigated corn, *Proceedings of 3rd International Conference on Precision Agriculture*, American Society of Agronomy, Crop Science Society of America, Soil Science Society of America, pp. 389–397.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N. et al. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups, *IEEE Signal Processing Magazine* **29**(6): 82–97.

Hu, H., Wen, Y., Chua, T.-S. and Li, X. (2014). Toward scalable systems for big data analytics: A technology tutorial, *IEEE Access* **2**: 652–687.

Huang, C., Davis, L. and Townshend, J. (2002). An assessment of support vector machines for land cover classification, *International Journal of remote sensing* **23**(4): 725–749.

Huete, A., Liu, H., Batchily, K. and Van Leeuwen, W. (1997). A comparison of vegetation indices over a global set of tm images for eos-modis, *Remote sensing of environment* **59**(3): 440–451.

Huglin, P. (1978). Nouveau mode d'évaluation des possibilités héliothermiques d'un milieu viticole, *Comptes rendus des seances* .

Hunt, E. B., Marin, J. and Stone, P. J. (1966). Experiments in induction, *Induction of decision trees* .

Hunt Jr, E. R. and Rock, B. N. (1989). Detection of changes in leaf water content using near-and middle-infrared reflectances, *Remote sensing of environment* **30**(1): 43–54.

Hutchinson, A. (1994). *Algorithmic learning*, Oxford University Press, Inc.

Jackson, R. D. and Huete, A. R. (1991). Interpreting vegetation indices, *Preventive veterinary medicine* **11**(3): 185–200.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means, *Pattern recognition letters* **31**(8): 651–666.

Jain, A. K. and Dubes, R. C. (1988). *Algorithms for clustering data*, Prentice-Hall, Inc.

Jardine, N. and van Rijsbergen, C. J. (1971). The use of hierarchic clustering in information retrieval, *Information storage and retrieval* **7**(5): 217–240.

Jiang, S., Pang, G., Wu, M. and Kuang, L. (2012). An improved k-nearest-neighbor algorithm for text categorization, *Expert Systems with Applications* **39**(1): 1503–1509.

Johnson, C. K., Mortensen, D. A., Wienhold, B. J., Shanahan, J. F. and Doran, J. W. (2003). Site-specific management zones based on soil electrical conductivity in a semiarid cropping system, *Agronomy Journal* **95**(2): 303–315.

Jones, R. J., Hiederer, R., Rusco, E. and Montanarella, L. (2005). Estimating organic carbon in the soils of europe for policy support, *European Journal of Soil Science* **56**(5): 655–671.

Jordan, C. F. (1969). Derivation of leaf-area index from quality of light on the forest floor, *Ecology* **50**(4): 663–666.

Kang, D.-K. and Kim, M.-J. (2011). Propositionalized attribute taxonomies from data for data-driven construction of concise classifiers, *Expert Systems with Applications* **38**(10): 12739–12746.

Karlik, B. and Olgac, A. V. (2011). Performance analysis of various activation functions in generalized MLP architectures of neural networks, *International Journal of Artificial Intelligence and Expert Systems* **1**(4): 111–122.

Kavzoglu, T. and Mather, P. (2003). The use of backpropagating artificial neural networks in land cover classification, *International Journal of Remote Sensing* **24**(23): 4907–4938.

Kholsa, R., Shaver, T., Reich, R. and Gangloff, W. (2001). Evaluating management zones for variable rate nitrogen management in corn, *The Third International Conference on Geospatial Information in Agriculture and Forestry*, pp. 5–7.

King, B. (1967). Step-wise clustering procedures, *Journal of the American Statistical Association* **62**(317): 86–101.

Kitchen, N., Snyder, C., Franzen, D. and Wiebold, W. (2002). Educational needs of precision agriculture, *Precision Agriculture* **3**(4): 341–351.

Klein, I., Gessner, U. and Kuenzer, C. (2012). Regional land cover mapping and change detection in central asia using modis time-series, *Applied Geography* **35**(1–2): 219–234.

Kneubuehler, M., Koetz, B., Huber, S., Schopfer, J., Itten, K. and Richter, R. (2006). Monitoring vegetation growth using multitemporal CHRIS/PROBA data, *IEEE International Conference on Geoscience and Remote Sensing Symposium.*, IEEE, pp. 2677–2680.

Knyazikhin, Y., Martonchik, J., Diner, D., Myneni, R., Verstraete, M., Pinty, B. and Gobron, N. (1998). Estimation of vegetation canopy leaf area index and fraction of absorbed photosynthetically active radiation from atmosphere-corrected misr data, *Journal of Geophysical Research: Atmospheres* **103**(D24): 32239–32256.

Koc, L., Mazzuchi, T. A. and Sarkani, S. (2012). A network intrusion detection system based on a hidden naïve bayes multiclass classifier, *Expert Systems with Applications* **39**(18): 13492–13500.

Kogan, F. (1990). Remote sensing of weather impacts on vegetation in non-homogeneous areas, *International Journal of Remote Sensing* **11**(8): 1405–1419.

Kogan, F. (2002). World droughts in the new millennium from AVHRR–based vegetation health indices, *Eos, Transactions American Geophysical Union* **83**(48): 557–563.

Kogan, F., Gitelson, A., Zakarin, E., Spivak, L. and Lebed, L. (2003). AVHRR-based spectral vegetation index for quantitative assessment of vegetation state and productivity: Calibration and validation, *Photogrammetric Engineering & Remote Sensing* **69**(8): 899–906.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection, *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'95, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 1137–1143.

Kohavi, R. and Provost, F. (1998). Glossary of terms, *Machine Learning* **30**(2–3): 271–274.

Kohavi, R. et al. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection, *International joint conference on Artificial Intelligence*, pp. 1137–1145.

Kotsiantis, S. (2007). Supervised machine learning: A review of classification techniques, *Informatica* **31**: 249–268.

Kriegler, F., Malila, W., Nalepka, R. and Richardson, W. (1969). Preprocessing transformations and their effects on multispectral recognition, *Remote Sensing of Environment, VI*, Vol. 1, p. 97.

Krizhevsky, A., Sutskever, I. and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks, *Advances in neural information processing systems*, pp. 1097–1105.

Kuhn, M. (2008). Building predictive models in R using the caret package, *Journal of Statistical Software* **28**(5): 1–26.

Kuhn, M. and Johnson, K. (2013). *Applied Predictive Modeling*, Springer, New York, Heidelberg, Dordrecht, London.

Kumar, J., Mills, R. T., Hoffman, F. M. and Hargrove, W. W. (2011). Parallel K-means clustering for quantitative ecoregion delineation using large data sets, *Procedia Computer Science* **4**: 1602–1611.

Lakshminarayan, K., Harp, S. A., Goldman, R. P., Samad, T. et al. (1996). Imputation of missing data using machine learning techniques, *KDD*, pp. 140–145.

Lance, G. and Williams, W. (1966). A generalized sorting strategy for computer classifications, *Nature* **212**: 218.

Landgraf, G. and Fusco, L. (1997). Earthnet online: The ESA earth observation multi-mission user information services start the operational phase, *Future Trends in Remote Sensing-Proceedings of the 17th EARSel Symposium. AA Balkema, Rotterdam Brookfield*, Vol. 5156, p. 40.

*Landsat Programme* (1972). `http://landsat.gsfc.nasa.gov/?p=3231`.

Larson, W. and Pierce, F. (1991). Conservation and enhancement of soil quality, *Evaluation for sustainable land management in the developing world: Proceedings of the International Workshop on Evaluation for Sustainable Land Management in the Developing World*, International Board for Soil Research and Management.

Lau, B. C., Ma, E. W. and Chow, T. W. (2014). Probabilistic fault detector for wireless sensor network, *Expert Systems with Applications* **41**(8): 3703–3711.

Lawrence, I. and Lin, K. (1989). A concordance correlation coefficient to evaluate reproducibility, *Biometrics* pp. 255–268.

Lee Rodgers, J. and Nicewander, W. A. (1988). Thirteen ways to look at the correlation coefficient, *The American Statistician* **42**(1): 59–66.

Li, X. (2009). K-Means and K-Medoids., *in* L. Liu and M. T. Özsu (eds), *Encyclopedia of Database Systems*, Springer US, pp. 1588–1589.

Lim, T.-S., Loh, W.-Y. and Shih, Y.-S. (2000). A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms, *Machine learning* **40**(3): 203–228.

Lin, S.-l. and Liu, Z. (2007). Parameter selection in SVM with RBF kernel function, *Journal-Zhejiang University of Technology* **35**(2): 163.

Liu, M. and Samal, A. (2002). A fuzzy clustering approach to delineate agroecozones, *Ecological modelling* **149**(3): 215–228.

Lorena, A. C., De Carvalho, A. C. and Gama, J. M. (2008). A review on the combination of binary classifiers in multiclass problems, *Artificial Intelligence Review* **30**(1-4): 19–37.

Lorenzo, M., Taboada, J., Lorenzo, J. and Ramos, A. (2013). Influence of climate on grape production and wine quality in the Rías Baixas, north-western spain, *Regional Environmental Change* **13**(4): 887–896.

Loureiro, M. D., Martínez, M. C., Boursiquot, J.-M. and This, P. (1998). Molecular marker analysis of Vitis vinifera "Albariño" and some similar grapevine cultivars, *Journal of the American Society for Horticultural Science* **123**(5): 842–848.

Lowenberg-DeBoer, J. and Swinton, S. (1997). *Economics of site-specific management in agronomic crops*, American Society of Agronomy, Crop Science Society of America, Soil Science Society of America.

Ludena, R. D. A. and Ahrary, A. (2013). A big data approach for a new ICT agriculture application development, *Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), 2013 International Conference on*, IEEE, pp. 140–143.

Ludwig, B., Nitschke, R., Terhoeven-Urselmans, T., Michel, K. and Flessa, H. (2008). Use of mid-infrared spectroscopy in the diffuse-reflectance mode for the prediction of the composition of organic matter in soil and litter, *Journal of Plant Nutrition and Soil Science* **171**(3): 384–391.

Luger, G. F. and Stubblefield, W. A. (1990). *Artificial intelligence and the design of expert systems*, Benjamin-Cummings Publishing Co., Inc.

Luoa, Z., Yaolin, L., Jiana, W. and Jingb, W. (2008). Quantitative mapping of soil organic material using field spectrometer and hyperspectral remote sensing, *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Science* **37**: 901–906.

Lynch, C. (2008). Big data: How do your data grow?, *Nature* **455**(7209): 28–29.

MacQueen, J. et al. (1967). Some methods for classification and analysis of multivariate observations, *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, Vol. 1, Oakland, CA, USA., pp. 281–297.

Mahdavi, M., Chehreghani, M. H., Abolhassani, H. and Forsati, R. (2008). Novel meta-heuristic algorithms for clustering web documents, *Applied Mathematics and Computation* **201**(1): 441–451.

Marconcini, M., Camps-Valls, G. and Bruzzone, L. (2009). A composite semisupervised SVM for classification of hyperspectral images, *Geoscience and Remote Sensing Letters, IEEE* **6**(2): 234–238.

Marr, D. (1977). Artificial intelligence – a personal view, *Artificial Intelligence* **9**(1): 37–48.

Marshall, J., Langille, R. and Palmer, W. M. K. (1947). Measurement of rainfall by radar, *Journal of Meteorology* **4**(6): 186–192.

Maulik, U. and Bandyopadhyay, S. (2002). Performance evaluation of some clustering algorithms and validity indices, *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **24**(12): 1650–1654.

Mayer-Schönberger, V. and Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*, Houghton Mifflin Harcourt.

McBratney, A., Whelan, B., Ancev, T. and Bouma, J. (2005). Future directions of precision agriculture, *Precision Agriculture* **6**(1): 7–23.

Melton, R. B., DeVaney, D. M. and French, J. C. (1995). The role of metadata in managing large environmental science datasets, *Proceedings of SDM-95, A Planning Workshop*, Pacific Northwest Laboratory, pp. 3–5.

Meyer-Aurich, A., Matthes, U., Osinski, E., Association, A. A. E. et al. (2001). *Integrating Sustainability in Agriculture: Trade-offs and Economic Consequences Demonstrated with a Farm Model in Bavaria*, Citeseer.

Meyer, G. E., Camargo Neto, J., Jones, D. D. and Hindman, T. W. (2004). Intensified fuzzy clusters for classifying plant, soil, and residue regions of interest from color images, *Computers and electronics in agriculture* **42**(3): 161–180.

Midgarden, D., Fleisher, R., Weisz, R. and Smilowitz, Z. (1997). Impact of site-specific IPM on the development of esfenvalerate resistance in Colorado potato beetle (Coleoptera: Chrysomelidae) and on densities of natural enemies, *Journal Economic Entomology* **90**: 855–867.

Milligan, G. W. and Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set, *Psychometrika* **50**(2): 159–179.

Milligan, G. W. and Cooper, M. C. (1987). Methodology review: Clustering methods, *Applied psychological measurement* **11**(4): 329–354.

Min, J. H. and Lee, Y.-C. (2005). Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters, *Expert systems with applications* **28**(4): 603–614.

Mirkin, B. (2012). *Clustering: a data recovery approach*, CRC Press.

Mistikoglu, G., Gerek, I. H., Erdis, E., Usmen, P. M., Cakan, H. and Kazan, E. E. (2015). Decision tree analysis of construction fall accidents involving roofers, *Expert Systems with Applications* **42**(4): 2256–2263.

Mitchell, T. M. et al. (1997). *Machine learning. WCB*, McGraw-Hill Boston, MA:.

Mladenic, D. and Grobelnik, M. (1999). Feature selection for unbalanced class distribution and naive bayes, *ICML*, Vol. 99, pp. 258–267.

Montanes, E., Diaz, I., Ranilla, J., Combarro, E. F. and Fernandez, J. (2005). Scoring and selecting terms for text categorization, *IEEE Intelligent Systems* **20**(3): 40–47.

Moody, J., Hanson, S., Krogh, A. and Hertz, J. A. (1992). A simple weight decay can improve generalization, *Advances in neural information processing systems* **4**: 950–957.

Moral, F., Terrón, J. and Rebollo, F. (2011). Site-specific management zones based on the Rasch model and geostatistical techniques, *Computers and Electronics in Agriculture* **75**(2): 223–230.

Morgan, J. N. and Sonquist, J. A. (1963). Problems in the analysis of survey data, and a proposal, *Journal of the American statistical association* **58**(302): 415–434.

Murisier, F., Jeangros, B. and Aerny, J. (1986). Maîtrise du rendement et maturitédu raisin, *Revue suisse de viticulture, arboriculture, horticulture* **18**(3): 149–156.

Murtagh, F. and Legendre, P. (2014). Ward's hierarchical agglomerative clustering method: Which algorithms implement Ward's criterion?, *Journal of Classification* **31**(3): 274–295.

Murthy, S. K. (1998). Automatic construction of decision trees from data: A multi-disciplinary survey, *Data mining and knowledge discovery* **2**(4): 345–389.

Myneni, R., Hoffman, S., Knyazikhin, Y., Privette, J., Glassy, J., Tian, Y., Wang, Y., Song, X., Zhang, Y., Smith, G. et al. (2002). Global products of vegetation leaf area and fraction absorbed par from year one of modis data, *Remote sensing of environment* **83**(1): 214–231.

NASA and USGS (1990). NASA land processes distributed active archive center (LP DAAC). ASTER L1B. USGS/Earth resources observation and science (EROS) center, Sioux Falls, South Dakota. 2001.

Neto, M., Baptista, F., Navas, L. and Ruiz, G. (2012). A business intelligence approach to support a greenhouse tomato crop grey mould disease early warning system, *in* T. Mildorf and K. C. jr. (eds), *ICT for Agriculture, Rural Development and Environment*, Czech Centre for Science and Society, pp. 175–184.

Nilsson, N. J. (1965). *Learning machines: foundations of trainable pattern-classifying systems*, McGraw-Hill.

*NOAA Products* (2005). `http://www.ospo.noaa.gov/Products/land/index.html`.

Oceanic, N. and (NOAA), A. A. (2005). Level 1b format, `http://www2.ncdc.noaa.gov/docs/podug/html/c2/sec2-0.htm`.

Ormeño Villajos, S., Arozarena Villar, A., Martínez Peña, M., Palomo Arroyo, M., Villa Alcázar, G., Peces Morera, J. and Pérez García, L. (2008). Los satélites de media y baja resolución espacial como fuente de datos para la obtención de indicadores ambientales, *IX Congreso Nacional de Medio Ambiente, Madrid*.

Ortega, R. A. and Santibáñez, O. A. (2007). Determination of management zones in corn (Zea mays L.) based on soil fertility, *Computers and Electronics in agriculture* **58**(1): 49–59.

Ottinger, M., Kuenzer, C., Liu, G., Wang, S. and Dech, S. (2013). Monitoring land cover dynamics in the Yellow River Delta from 1995 to 2010 based on Landsat 5 TM, *Applied Geography* **44**: 53–68.

Pal, M. and Mather, P. (2005). Support vector machines for classification in remote sensing, *International Journal of Remote Sensing* **26**(5): 1007–1011.

Pal, N. R. and Bezdek, J. C. (1995). On cluster validity for the fuzzy c-means model, *Fuzzy Systems, IEEE Transactions on* **3**(3): 370–379.

Paliwal, M. and Kumar, U. A. (2009). Neural networks and statistical techniques: A review of applications, *Expert Systems with Applications* **36**(1): 2–17.

Panagos, P., Ballabio, C., Yigini, Y. and Dunbar, M. B. (2013). Estimating the soil organic carbon content for european NUTS2 regions based on LUCAS data collection, *Science of The Total Environment* **442**: 235–246.

Pearson, K. (1920). Notes on the history of correlation, *Biometrika* **13**(1): 25–45.

Peralta, N. R. and Costa, J. L. (2013). Delineation of management zones with soil apparent electrical conductivity to improve nutrient management, *Computers and Electronics in Agriculture* **99**: 218–226.

Perez-Ariza, C. B., Nicholson, A. E. and Flores, M. J. (2012). Prediction of coffee rust disease using bayesian networks, *Proceedings of the Sixth European Workshop on Probabilistic Graphical Models*, pp. 259–266.

Phipps, J. (1971). Dendrogram topology, *Systematic Biology* **20**(3): 306–308.

Pinty, B. and Verstraete, M. (1992). GEMI: a non-linear index to monitor global vegetation from satellites, *Vegetatio* **101**(1): 15–20.

Powers, D. M. W. (2007). Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation, *Technical Report SIE-07-001*, School of Informatics and Engineering, Flinders University, Adelaide, Australia.

Qing Liu, H. and Huete, A. (1995). A feedback based modification of the ndvi to minimize canopy background and atmospheric noise, *Geoscience and Remote Sensing, IEEE Transactions on* **33**(2): 457–465.

Quinlan, J. R. (1987). Simplifying decision trees, *International journal of man-machine studies* **27**(3): 221–234.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*, Vol. 1, Morgan Kaufmann.

Quinlan, J. R. (1996a). Bagging, boosting, and C4.5, *AAAI/IAAI, Vol. 1*, pp. 725–730.

Quinlan, J. R. (1996b). Improved use of continuous attributes in C4.5, *Journal of artificial intelligence research* **4**: 77–90.

Quinlan, J. R. et al. (1992). Learning with continuous classes, *5th Australian joint conference on artificial intelligence*, Vol. 92, Singapore, pp. 343–348.

Quinlan, R. J. (1994). C4.5: Programs for machine learning, *Machine Learning* **16**(3): 235–240.

Raghupathi, W. and Raghupathi, V. (2014). Big data analytics in healthcare: promise and potential, *Health Information Science and Systems* **2**(1): 1.

Ripley, B. D. and Hjort, N. L. (1995). *Pattern Recognition and Neural Networks*, 1st edn, Cambridge University Press, New York, NY, USA.

Robert, P. (1993). Characterization of soil conditions at the field level for soil specific management, *Geoderma* **60**(1): 57–72.

Rouse Jr, J., Haas, R., Schell, J. and Deering, D. (1974). Monitoring vegetation systems in the great plains with ERTS, *NASA special publication* **351**: 309.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, *Journal of Computational and Applied Mathematics* **20**(1): 53–65.

Rubio, M., Riaño, D., Cheng, Y. and Ustin, S. (2006). Estimation of canopy water content from modis using artificial neural networks trained with radiative transfer models, *6th EMS/6th ECAC* .

Ruggieri, S. (2002). Efficient C4. 5 [classification algorithm], *IEEE transactions on knowledge and data engineering* **14**(2): 438–444.

Russell, S., Norvig, P. and Intelligence, A. (1995). A modern approach, *Artificial Intelligence. Prentice-Hall, Egnlewood Cliffs* **25**: 27.

Samadi, M., Jabbari, E. and Azamathulla, H. M. (2014). Assessment of M5´ model tree and classification and regression trees for prediction of scour depth below free overfall spillways, *Neural Computing and applications* **24**(2): 357–366.

Schuster, E., Kumar, S., Sarma, S. E., Willers, J. and Milliken, G. (2011). Infrastructure for data-driven agriculture: identifying management zones for cotton using statistical modeling and machine learning techniques, *Emerging Technologies for a Smarter World (CEWIT), 2011 8th International Conference Expo on*, pp. 1–6.

Schwartz, D. B., Samalam, V. K., Solla, S. A. and Denker, J. S. (1990). Exhaustive learning, *Neural Computation* **2**(3): 374–385.

Sebastiani, F. (2002). Machine learning in automated text categorisation, *ACM Computing Survey* **34**(1).

Seiffert, U., Bollenbeck, F., Mock, H. and Matros, A. (2010). Clustering of crop phenotypes by means of hyperspectral signatures using artificial neural networks, *Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS), 2010 2nd Workshop on*, pp. 1–4.

Seppelt, R. and Voinov, A. (2002). Optimization methodology for land use patterns using spatially explicit landscape models, *Ecological Modelling* **151**(2): 125–142.

Shao, J. (1993). Linear model selection by cross-validation, *Journal of the American statistical Association* **88**(422): 486–494.

Simbahan, G. C. and Dobermann, A. (2006). An algorithm for spatially constrained classification of categorical and continuous soil properties, *Geoderma* **136**(3): 504–523.

*Sistema de Información Geográfica de Parcelas Agrícolas* (n.d.). `http://sigpac.magrama.es/fega/h5visor/`. Accessed: 2015-01-20.

Sneath, P. H., Sokal, R. R. et al. (1973). *Numerical taxonomy. The principles and practice of numerical classification.*, Freeman, London, UK.

Sokolova, M. and Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks, *Information Processing & Management* **45**(4): 427–437.

Stafford, J. and Miller, P. (1996). Spatially variable treatment of weed patches, *Precision Agriculture* (precisionagricu3): 465–474.

Sugar, C. A. and James, G. M. (2003). Finding the number of clusters in a dataset: An information-theoretic approach, *Journal of the American Statistical Association* **98**(463): 750–763.

Sun, G. (1991). Prediction of vegetable yields by grey model GM (1, 1), *Journal of Grey System* **2**(2): 187–197.

Suzuki, R. and Shimodaira, H. (2006). Pvclust: an R package for assessing the uncertainty in hierarchical clustering, *Bioinformatics* **22**(12): 1540–1542.

Suzuki, R. and Shimodaira, H. (2013). Hierarchical clustering with P-values via multiscale bootstrap resampling, *R package* .

Swinton, S. (1997). Precision farming as green and competitive, *Draft for publication* p. 9.

Taigman, Y., Yang, M., Ranzato, M. and Wolf, L. (2014). Deepface: Closing the gap to human-level performance in face verification, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1701–1708.

Thrikawala, S., Weersink, A., Fox, G. and Kachanoski, G. (1999). Economic feasibility of variable-rate technology for nitrogen on corn, *American Journal of Agricultural Economics* **81**(4): 914–927.

Timmermann, C., Gerhards, R., Krohmann, P., Sokefeld, M. and Kuhbauch, W. (2001). The economical and ecological impact of the site-specific weed control, *Proceedings of the 3rd European conference on precision agriculture*, pp. 563–568.

Tonietto, J. and Carbonneau, A. (2004). A multicriteria climatic classification system for grape-growing regions worldwide, *Agricultural and Forest Meteorology* **124**(1): 81–97.

Tripathy, A., Adinarayana, J., Sudharsan, D., Merchant, S., Desai, U., Vijayalakshmi, K., Reddy, D. R., Sreenivas, G., Ninomiya, S., Hirafuji, M. et al. (2011). Data mining and wireless sensor network for agriculture pest/disease predictions, *Information and Communication Technologies (WICT), 2011 World Congress on*, IEEE, pp. 1229–1234.

Trombetti, M., Riaño, D., Rubio, M., Cheng, Y. and Ustin, S. (2008). Multi-temporal vegetation canopy water content retrieval and interpretation using artificial neural networks for the continental USA, *Remote Sensing of Environment* **112**(1): 203–215.

USGS (1972). LANDSAT project, `http://landsat.usgs.gov/`. Accessed: 2015-02-30.

Van Alphen, B. (2002). A case study on precision nitrogen management in dutch arable farming, *Nutrient Cycling in Agroecosystems* **62**(2): 151–161.

Villajos, S. O., Villar, A. A., Pena, M. M., Arroyo, M. P., Alcázar, G. V., Morera, J. P. and Garcıa, L. P. (2008). Los satélites de media y baja resolución espacial como fuente de datos para la obtención de indicadores ambientales, *IX Congreso Nacional de Medio Ambiente, Madrid (In Spanish)*.

VITO (1998). VITO-SPOT-VEGETATION, `http://www.spot-vegetation.com/`.

Voronoï, G. (1908). Nouvelles applications des paramètres continus à la théorie des formes quadratiques. Deuxième mémoire. Recherches sur les parallélloèdres primitifs., *Journal für die reine und angewandte Mathematik* **134**: 198–287.

Wang, H. and Ma, Z. (2011). Prediction of wheat stripe rust based on support vector machine, *Natural Computation (ICNC), 2011 Seventh International Conference on*, Vol. 1, IEEE, pp. 378–382.

Wang, K., Wang, B. and Peng, L. (2009). Cvap: validation for cluster analyses, *Data Science Journal* **8**: 88–93.

Wang, Y. and Witten, I. H. (1996). *Induction of model trees for predicting continuous classes*, Department of Computer Science, University of Waikato.

Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function, *Journal of the American statistical association* **58**(301): 236–244.

Wasserman, P. D. (1989). *Neural computing*, Van Nostrand Reinhold, New York.

Westfall, P. H. and Young, S. S. (1993). *Resampling-based multiple testing: Examples and methods for p-value adjustment*, Vol. 279, John Wiley & Sons.

Whitley, K., Davenport, J., Manley, S., Robert, P., Rust, R., Larson, W. et al. (2000). Differences in nitrate leaching under variable and conventional nitrogen fertilizer management in irrigated potato systems., *Proceedings of the 5th International Conference on Precision Agriculture, Bloomington, Minnesota, USA, 16-19 July, 2000.*, American Society of Agronomy, pp. 1–9.

Wiebold, B., Sudduth, K., Davis, G., Shannon, K. and Kitchen, N. (1998). Determining barriers to adoption and research needs of precision agriculture, *Precise News, VI* .

Wilson, D. R. and Martinez, T. R. (2000). An integrated instance-based learning algorithm, *Computational Intelligence* **16**(1): 1–28.

Winkler, A. J. (1962). *General viticulture*, Univ of California Press.

Wulfsohn, D., Zamora, F. A., Téllez, C. P., Lagos, I. Z. and García-Fiñana, M. (2012). Multilevel systematic sampling to estimate total fruit number for yield forecasts, *Precision agriculture* **13**(2): 256–275.

Xie, H., Yang, X., Drury, C., Yang, J. and Zhang, X. (2011). Predicting soil organic carbon and total nitrogen using mid-and near-infrared spectra for brookston clay loam soil in southwestern ontario, canada, *Canadian Journal of Soil Science* **91**(1): 53–63.

Yang, T. and Pedersen, J. P. (1997a). Feature selection in statistical learning of text categorization, *14th Int. Conf. on Machine Learning*, pp. 412–420.

Yang, Y. and Pedersen, J. O. (1997b). A comparative study on feature selection in text categorization, *ICML*, Vol. 97, pp. 412–420.

Ye, X., Sakai, K., Manago, M., Asada, S.-i. and Sasao, A. (2007). Prediction of citrus yield from airborne hyperspectral imagery, *Precision Agriculture* **8**(3): 111–125.

Yidan, H. Y. B. (1992). Grey-Markov forecasting model and its application, *Systems Engineering-theory & Practice* **4**: 012.

Yu, H., Liu, D., Chen, G., Wan, B., Wang, S. and Yang, B. (2010). A neural network ensemble method for precision fertilization modeling, *Mathematical and Computer Modelling* **51**(11): 1375–1382.

Zadeh, L. A. (1965). Fuzzy sets, *Information and control* **8**(3): 338–353.

Zadrozny, B. (2004). Learning and evaluating classifiers under sample selection bias, *Proceedings of the twenty-first international conference on Machine learning*, ACM, p. 114.

Zhang, B., Li, S., Wu, C., Gao, L., Zhang, W. and Peng, M. (2013). A neighbourhood-constrained K-means approach to classify very high spatial resolution hyperspectral imagery, *Remote Sensing Letters* **4**(2): 161–170.

Zhang, H., Perng, C.-S., Cai, Q. et al. (2002). An improved algorithm for feature selection using fractal dimension, *Proceedings of the Second International Workshop on Databases, Documents, and Information Fusion*, Citeseer.

Zhang, Q. and Han, S. (2002). An information table for yield data analysis and management, *Biosystems engineering* **83**: 299–306.

Zhao, Y. and Karypis, G. (2002). Evaluation of hierarchical clustering algorithms for document datasets, *Proceedings of the eleventh international conference on Information and knowledge management*, ACM, pp. 515–524.

Zhou, Y., Yang, X. and Wang, L. (2007). Study on Grey-Markov method and its application in agricultural production forecast, *IEEE International Conference on Grey Systems and Intelligent Services*, pp. 553–557.

Zhu, H. and Basir, O. (2005). An adaptive fuzzy evidential nearest neighbor formulation for classifying remote sensing images, *Geoscience and Remote Sensing, IEEE Transactions on* **43**(8): 1874–1889.

Zhu, Y., Toth, Z., Wobus, R., Richardson, D. and Mylne, K. (2002). The economic value of ensemble-based weather forecasts, *Bulletin of the American Meteorological Society* **83**(1): 73.

# Appendix A

# Downloading and processing satellite imagery

## A.1 Procesing MODIS data products

The MODIS instrument is available on the Terra and Aqua satellites covering the surface of the Earth every 1 or 2 days. As mentioned in Section 3.1.1, MODIS produces several data products related with the agriculture.

### A.1.1 Downloading MODIS data products

MODIS data products are freely available for downloading from several sites like the LP DAAC or (http://e4ftl01.cr.usgs.gov/MOLT). It also can be found at the following FTP sites:

- ftp://nrt1.modaps.eosdis.nasa.gov
- ftp://nrt2.modaps.eosdis.nasa.gov.

The MODIS data products follow a naming convention which includes the name of the product, the dates of acquisition and production and the tile identifier of the region of the Earth. This tile identifier consist on a horizontal and vertical coordinates which correspond to the tile of the grid that divides the Earth in adjacent non-overlapping tiles that are approximately 10 degrees square (at the equator). It is called the MODIS Sinusoidal Tiling System.

For example, the file name *MOD09GQ.A2015063.h18v03.005.2015069183737.hdf* corresponds to the *MOD09GQ* data product for the region *18H*, *3V* and the day March the 4th of 2015:

- MOD09GQ - Product Short Name
- A2015063 - Julian Date of Acquisition (A-YYYYDDD)
- h18v03 - Tile Identifier (horizontalXXverticalYY)

- 005 - Collection Version
- 2015069183737 - Julian Date of Production (YYYYDDDHHMMSS)
- hdf - Data Format (HDF-EOS)

### A.1.2 Processing MODIS data products

The data products MOD09GQ and MOD09GA are stored in HDF files which contain the raw values of the reflectivity bands (see Table A.1) for a specific tile and period of time.

The objective of the process, is to extract the values of each band stored on the HDF, for each pixel of the image, generating a dataset with the results. Thus, the process will transforms the data of each band into a GeoTIFF file. The process can also consider a specific region of the tail and extract only the values of that region.

Once the data products associated to the region of interest and the period of time considered are downloaded, the raw values are extracted.

### A.1.3 Calculation of vegetation and moisture indices from MODIS data products

Using MODIS MYD09GQ data product it can be obtained the following vegetation indices with a spatial resolution of 500m. See Table A.1 for the description of the bands and the correspondences with the indices defined on Section 3.3.6

- NDVI
- NDVI.SCALED [0..255]

| MODIS | Description | Data product |
|---|---|---|
| B1 | Red | MYD09GQ, MYD09GA |
| B2 | Near-Infrared (NIR) | MYD09GQ, MYD09GA |
| B3 | Blue | MYD09GA |
| B4 | Green | MYD09GA |
| B6 | Short Wavelength Infrared (SWIR) 1 | MYD09GA |
| B7 | SWIR 2 | MYD09GA |

Table A.1: Description of MODIS bands for data products MYD09GQ and MYD09GA

From MODIS MYD09GA it can be obtained NDVI, NDVI.scaled and the following vegetation and moisture indices with a spatial resolution of 250m.:

- EVI
- NDI7
- SIWSI

- SWIRR
- MSI
- GVMI

## A.2 Processing Landsat 8 data products

OLI and TIRS data products are delivered as compressed files with as many GeoTIFF files as bands are included in the data product. Each GeoTIFF file contains an scene of a specific region of the Earth and each pixel has a value. For instance, for band 2 pixels contain the visible blue colour. Table A.2 shows the bands included in each Landsat 8 data product.

| Bands | Description | Wavelength (micrometers) | Resolution (meters) |
|---|---|---|---|
| Band 1 | Coastal aerosol | 0.43 - 0.45 | 30 |
| Band 2 | Blue | 0.45 - 0.51 | 30 |
| Band 3 | Green | 0.53 - 0.59 | 30 |
| Band 4 | Red | 0.64 - 0.67 | 30 |
| Band 5 | Near Infrared (NIR) | 0.85 - 0.88 | 30 |
| Band 6 | SWIR 1 | 1.57 - 1.65 | 30 |
| Band 7 | SWIR 2 | 2.11 - 2.29 | 30 |
| Band 8 | Panchromatic | 0.50 - 0.68 | 15 |
| Band 9 | Cirrus | 1.36 - 1.38 | 30 |
| Band 10 | Thermal Infrared (TIRS) 1 | 10.60 - 11.19 | 100 * (30) |
| Band 11 | Thermal Infrared (TIRS) 2 | 11.50 - 12.51 | 100 * (30) |

Table A.2: Description of OLI and TIRS bands

The ZIP files also include a GeoTIFF for the Quality Assurance band and a metadata file associated to the data product. This metadata file contains information to calculate:

- **Radiance Top of Atmosphere (TOA)**:

$$L_\lambda = M_L * Q_{cal} + A_L, \tag{A.1}$$

where:

  - $L_\lambda$ is the TOA spectral radiance measured on Watts / ($m^2$ * srad * $\mu m$)
  - $M_L$ is the multiplicative scale factor for an specific band. This factor is provided in the variable $RADIANCE\_MULT\_BAND\_x$ included on the metadata file, where $x$ is the number of the band.

- $A_L$ is the additive scale factor corresponding to the variable $RADIANCE\_ADD\_BAND\_x$ included on the metadata file, where $x$ is the number of the band.
- $Q_{cal}$ is the quantified standard product calibrated with the values of the DN pixel. This value is for each one of the bands of the image.

- **Brightness temperature**:

$$T = \frac{K_2}{ln(\frac{K_1}{L_\lambda} + 1)}, \tag{A.2}$$

where:

- T is the at-satellite brightness temperature measured in degrees kelvin (K)
  $L_\lambda$ is the TOA spectral radiance
- $K_1$ is the conversion constant specific for the thermal bands. Corresponds to the value of the variable $K1\_CONSTANT\_BAND\_x$ included on the metadata file, where $x$ is the number of the band (10 or 11)
- $K_2$ is the conversion constant specific for the thermal bands. Corresponds to the value of the variable $K2\_CONSTANT\_BAND\_x$ included on the metadata file, where $x$ is the number of the band (10 or 11)

- **Reflectance TOA with angular correction**:

$$\rho_\lambda = \frac{M_\rho Q_{cal} + A_\rho}{sin(\theta_{SE})}, \tag{A.3}$$

where:

- $\rho_\lambda$ is the TOA planetary reflectance.
- $\theta_{SE}$ is the local sun elevation angle in degrees provided in the metadata $SUN\_ELEVATION$.
- $M_\rho$ is the multiplicative scale factor for an specific band. It is provided in the variable $REFLECTANCE\_MULT\_BAND\_x$ included on the metadata file, where $x$ is the number of the band.
- $A_\rho$ is the additive scale factor corresponding to the variable $REFLECTANCE\_ADD\_BAND\_x$ included on the metadata file, where $x$ is the number of the band.
- $Q_{cal}$ is the quantified standard product calibrated with the values of the DN pixel. This value is for each one of the bands of the image.

Once the Landsat 8 data products are downloaded for a specific date interval and geographical area, for instance via the web portal EarthExplorer (http://earthexplorer.usgs.gov/) the raw values of each pixel for each band of the image can be extracted.

### A.2.1 Calculation of vegetation and moisture indices from Operational Land Imager data products

From the OLI data product it is possible to calculate the following vegetation and moisture indices according to Table A.2 and the indices defined on Section 3.3.6

- NDVI
- NDVI.SCALED [0..255]
- EVI
- NDI7
- SIWSI
- SWIRR
- MSI
- GVMI

### A.2.2 LST8 package

LST8 is an R package developed by the author and published on GitHub (https://github.com/rbarango/LST8) under a open-source license.

The package relies on the LST8 function that processes Level 1 GeoTIFF Data Product of Landsat 8 satellite retrieving OLI & TIRS raw values and calculating: reflectance TOA, radiance TOA and brightness temperature.

The function considers as input (1) the folder where the GEOTIF data products are stored taking into account that each data product must be decompress in one folder and (2) a geometry object corresponding to the geographical zone to retrieve the data from (for instance a SpatialPolygonsDataFrame)

The output of this function is a CSV file including the spatial data points of the extent with the raw data of the OLI (Band B8 panchromatic is not included) and TIRS bands and their corresponding reflectance and radiance TOA values for the spectral bands and, for thermal bands, their brightness temperature. The name of the CSV file is LST8_startDate_endDate.

This function has dependences on *rgdal*, *raster* and *sp* R packages. It should be noticed that the function does not download the data products. The data products should be downloaded before running the function, for instance from (http://earthexplorer.usgs.gov/)

The structure of the CSV file generated by the function is the following:

- x: Coordinate x of the data point in the CRS of the data product
- y: Coordinate y of the data point in the CRS of the data product crs_proj: CRS projection string of the data product

- date: Year + Day number of the year in the format YYYYddd
- 30_B1: OLI band 1
- 30_B2: OLI band 2
- 30_B3: OLI band 3
- 30_B4: OLI band 4
- 30_B5: OLI band 5
- 30_B6: OLI band 6
- 30_B7: OLI band 7
- 30_B9: OLI band 9
- 30_B10: TIRS band 10
- 30_B11: TIRS band 11
- 30_QA: Quality indicator of the measures
- 30_B1_RADIANCE: RADIANCE (TOA) for band 1
- 30_B1_REFLECTANCE: REFLECTANCE (TOA) for band 1
- 30_B2_RADIANCE: RADIANCE (TOA) for band 2
- 30_B2_REFLECTANCE: REFLECTANCE (TOA) for band 2
- 30_B3_RADIANCE: RADIANCE (TOA) for band 3
- 30_B3_REFLECTANCE: REFLECTANCE (TOA) for band 3
- 30_B4_RADIANCE: RADIANCE (TOA) for band 4
- 30_B4_REFLECTANCE: REFLECTANCE (TOA) for band 4
- 30_B5_RADIANCE: RADIANCE (TOA) for band 5
- 30_B5_REFLECTANCE: REFLECTANCE (TOA) for band 5
- 30_B6_RADIANCE: RADIANCE (TOA) for band 6
- 30_B6_REFLECTANCE: REFLECTANCE (TOA) for band 6
- 30_B7_RADIANCE: RADIANCE (TOA) for band 7
- 30_B7_REFLECTANCE: REFLECTANCE (TOA) for band 7
- 30_B9_RADIANCE: RADIANCE (TOA) for band 9

- 30_B9_REFLECTANCE: REFLECTANCE (TOA) for band 9
- 30_B10_TEMPERATURE: BRIGHTNESS TEMPERATURE (TOA) for band 11
- 30_B11_TEMPERATURE: BRIGHTNESS TEMPERATURE (TOA) for band 12

### A.2.3 Example of use of LST8 package

The following code corresponds to an example of script in R testing the use of package:

```
# Loading package dependencies
require(sp)
require(rgdal)
require(raster)
require(devtools)

# Installing LST8 package from GitHub
install_github("rbarango/LST8")
require(LST8)

# Loading polygons for retrieving REFLECTANCE, RADIANCE
# and BRIGHTNESS TEMPERATURE
shapesTG = load("shapesTG.RData")

# Executing the function
# As input the location of the GeoTIFF files from Landsat 8
# and the target geometry
getLST8(dir="../Documents/LST8/", geometry=shapesTG)
```

## A.3 R scripts related with MODIS data extraction and calculation of vegetation indices

The functions for processing and extracting the MODIS data products MOD09GQ and MYD09GA are published on Github under an open-source license:

- getMOD09GQ (https://github.com/rbarango/MODIS/getMOD09GQ.R)
- getMYD09GA (https://github.com/rbarango/MODIS/getMYD09GA.R)

Also it was published a function for the calculation of vegetation and moisture indices from MYD09GA: getVI_MYD09GA (https://github.com/rbarango/MODIS/getVI_MYD09GA.R)

### A.3.1 Function getMOD09GQ

The function getMOD09GQ extracts daily Tile Surface Reflectance Bands 1-2 product, spatial resolution: 250 m, for a particular extent and date interval (https://lpdaac.usgs.gov/products/modis_products_table/mod09gq)

It requires access to GDAL commands, for instance via the following command providing the right path to GDAL in your system:

```
Sys.setenv(PATH="$PATH:/Library/Frameworks/GDAL.framework/Programs")
```

The function requires the following parameters:

- **targetPoints**. JSON with the coordinates of the bounding box for the target zone:
  - ULx X coordinate of the upper-left corner
  - ULy Y coordinate of the upper-left corner
  - URx X coordinate of the upper-right corner
  - URy Y coordinate of the upper-right corner
  - LRx X coordinate of the lower-right corner
  - LRy Y coordinate of the lower-right corner
  - LLx X coordinate of the lower-left corner
  - LLy Y coordinate of the lower-left corner
- **CRSString** Coordinate reference system. CRS string of the targetPoints. Default = "+proj=utm +zone=29"
- **startDate** A string with the start date. Default=Sys.Date()-2. Format: yyyy/mm/dd
- **endDate** A string with the end date. Default = Sys.Date()-2. Format: yyyy/mm/dd

The function returns a JSON with the reflectivity values: x, y, date, refl_b01, refl_b02, QC_250m

#### Example of use

```
getMOD09GQ({\"ULx\":516750,\"ULy\":4644000,\"URx\":517750,\"URy\":4644000,
\"LRx\":517750,\"LRy\":4642750,\"LLx\":516750,\"LLy\":4642750}",
startDate = "2016/01/01", endDate = "2016/01/03")
```

### A.3.2 Function getMYD09GA

The function getMYD09GA extracts daily Tile Surface Reflectance Bands 1-7, spatial resolution: 500/1000m, for a particular extent and date interval (https://lpdaac.usgs.gov/products/modis_products_table/myd09ga)

The function requires access to GDAL commands, for instance via the following command providing the right path to GDAL in your system:

```
Sys.setenv(PATH="$PATH:/Library/Frameworks/GDAL.framework/Programs")
```

The function requires the same parameters described for the getMYD09GA function (see Section **??** and returns a JSON with the following reflectivity values:

```
x, y, date, refl_b01, refl_b02, refl_b03, refl_b04, refl_b05, refl_b06, refl_b07, QC_250m
```

**Example of use of getMYD09GA**

```
getMYD09GA({\"ULx\":516750,\"ULy\":4644000,\"URx\":517750,\"URy\":4644000,
\"LRx\":517750,\"LRy\":4642750,\"LLx\":516750,\"LLy\":4642750}",
startDate = "2016/01/01", endDate = "2016/01/03")
```

### A.3.3 Function getVI_MYD09GA

Function to calculate vegetation and moisture indices from the reflectivity values of bands 1 to 7 of the MOD09GA dataproduct. It requires the following parameters:

- Band1: a list of reflectivity values from MOD09GA band 1
- Band2: a list of reflectivity values from MOD09GA band 2
- Band3: a list of reflectivity values from MOD09GA band 3
- Band4: a list of reflectivity values from MOD09GA band 4
- Band5: a list of reflectivity values from MOD09GA band 5
- Band6: a list of reflectivity values from MOD09GA band 6
- Band7: a list of reflectivity values from MOD09GA band 7

The function returns a data frame with the following indices:

- NDVI
- NDVI.SCALED [0..255]
- EVI
- NDWI
- NDI7
- SIWSI
- SWIRR
- SRWI
- MSI_1
- MSI_2
- GVMI

### A.3.4 NDVI calculation from MOD09GQ

The following function calculates the NDVI and NDVI scaled from the reflectivity bands 1 and 2 of MOD09GQ dataproduct, returning a data frame with the NDVI and NDVI.SCALED [0..255] values.

The parameters of the function are the following:

- **band1**: a list of reflectivity values from MOD09GQ band 1
- **band2**: a list of reflectivity values from MOD09GQ band 2

The R script implementation is the following:

```
getNDVI_MOD09GQ  <- function(band1, band2){

  ndvi  <-  (band2 - band1) / (band2 + band1)
  ndvi.scaled  <- (ndvi+1)*127

  df  <- data.frame(ndvi=ndvi, ndvi.scaled=ndvi.scaled)
  names(df)  <- c("NDVI","NDVI.SCALED")

  return(df)
}
```

# Appendix B

# Meteorological features and regression equations for yield forecasting

## B.1 List of meteorological features considered for yield prediction

The following type of observations are collected by Meteogalicia with the agro-meteorological station As Eiras. The station is located at O Rosal, Pontevedra, Spain (41.94° latitude, −8.79° longitude) at an elevation of 52m.

1. Average temperature of the air (°C)
2. Max. temperature of the air(°C)
3. Min. temperature of the air (°C)
4. Average relative humidity (%)
5. Max. relative humidity (%)
6. Min. relative humidity (%)
7. Foliar humidity (h)
8. Rain temperature (°C)
9. Soil temperature (°C)
10. Air temperature (°C)
11. Wind speed (km/h)

12. Rainfall ($l/m^2$)
13. Wind gusts (km/h)
14. Sunshine hours (h)
15. Daily global irradiation ($10kJ/(m^2.day)$)
16. Wind gusts direction (degrees)
17. Soil humidity ($m^3/m^3$)
18. Hours of cold temperature (h)
19. Insolation (%)
20. Wind direction (degrees)
21. Hours of light (h)
22. Evapotranspiration of reference ($L/m^2$)

## B.2 Regression equations for yield forecasting

```
LM num: 1
.outcome =
259.5716 * AlbarinoVariety
+ 1890.137 * LoureiroVariety
+ 466.9527 * TreixaduraVariety
- 1.849 * P
+ 2503.9048 * K
+ 625.1787 * Mg
+ 44.9976 * Ca
+ 1112.4852 * pHWater
- 76.5733 * M.O.
- 0.5054 * refl_b01.STG19
+ 0.4181 * refl_b02.STG19
+ 2720.4307 * NDVI.STG19
+ 0.1766 * refl_b01.STG23
+ 0.247 * refl_b02.STG23
+ 4389.7688 * NDVI.STG23
- 0.1327 * refl_b02.STG32
- 1714.8408 * NDVI.STG32
- 7976.5896

LM num: 2
```

```
.outcome =
259.5716 * AlbarinoVariety
+ 1890.137 * LoureiroVariety
+ 466.9527 * TreixaduraVariety
- 1.849 * P
+ 5408.749 * K
+ 794.1707 * Mg
+ 44.9976 * Ca
+ 494.2108 * pHWater
- 266.7401 * M.O.
- 0.3514 * refl_b01.STG19
+ 0.4181 * refl_b02.STG19
+ 2720.4307 * NDVI.STG19
+ 0.1766 * refl_b01.STG23
+ 0.247 * refl_b02.STG23
+ 4389.7688 * NDVI.STG23
- 0.1327 * refl_b02.STG32
- 1714.8408 * NDVI.STG32
- 3500.8108

LM num: 3
.outcome =
657.167 * AlbarinoVariety
+ 1890.137 * LoureiroVariety
+ 466.9527 * TreixaduraVariety
- 1.849 * P
+ 6937.6144 * K
+ 794.1707 * Mg
+ 44.9976 * Ca
+ 494.2108 * pHWater
- 366.8279 * M.O.
- 0.3514 * refl_b01.STG19
+ 0.4181 * refl_b02.STG19
+ 2720.4307 * NDVI.STG19
+ 0.1766 * refl_b01.STG23
+ 0.247 * refl_b02.STG23
+ 4389.7688 * NDVI.STG23
- 0.1327 * refl_b02.STG32
- 1714.8408 * NDVI.STG32
- 3151.7216
```

```
LM num: 4
.outcome =
192.3884 * AlbarinoVariety
+ 5549.4594 * LoureiroVariety
+ 466.9527 * TreixaduraVariety
- 8.7231 * P
+ 3944.5745 * K
+ 265.6116 * Mg
- 30.6798 * Ca
- 295.0018 * M.O.
- 0.1337 * refl_b01.STG19
+ 1.8806 * refl_b02.STG19
+ 4379.1378 * NDVI.STG19
+ 0.1766 * refl_b01.STG23
+ 0.1462 * refl_b02.STG23
+ 4425.6046 * NDVI.STG23
+ 0.1496 * refl_b01.STG32
- 0.1327 * refl_b02.STG32
- 1714.8408 * NDVI.STG32
- 4324.9393

LM num: 5
.outcome =
192.3884 * AlbarinoVariety
+ 7116.4475 * LoureiroVariety
+ 466.9527 * TreixaduraVariety
- 24.0734 * P
+ 3852.6808 * K
+ 265.6116 * Mg
- 26.8959 * Ca
- 288.0211 * M.O.
- 0.1337 * refl_b01.STG19
+ 1.8217 * refl_b02.STG19
+ 4379.1378 * NDVI.STG19
+ 0.1766 * refl_b01.STG23
+ 0.1462 * refl_b02.STG23
+ 4425.6046 * NDVI.STG23
+ 0.1496 * refl_b01.STG32
- 0.1327 * refl_b02.STG32
- 1714.8408 * NDVI.STG32
- 2776.7653
```

```
LM num: 6
.outcome =
192.3884 * AlbarinoVariety
+ 3422.726 * LoureiroVariety
+ 466.9527 * TreixaduraVariety
- 8.1269 * P
+ 2106.7004 * K
+ 368.6906 * Mg
+ 44.9976 * Ca
- 224.6963 * M.O.
- 0.1337 * refl_b01.STG19
+ 0.9035 * refl_b02.STG19
+ 5797.4729 * NDVI.STG19
+ 0.1766 * refl_b01.STG23
+ 0.1462 * refl_b02.STG23
+ 4958.8916 * NDVI.STG23
+ 0.2547 * refl_b01.STG32
- 0.1327 * refl_b02.STG32
- 1714.8408 * NDVI.STG32
- 3143.971

LM num: 7
.outcome =
-271.7469 * AlbarinoVariety
+ 2494.0767 * LoureiroVariety
+ 3124.8175 * TreixaduraVariety
+ 13.2208 * P
+ 1296.5386 * K
+ 140.0604 * Ca
+ 2192.0811 * pHWater
+ 320.005 * M.O.
- 0.7967 * refl_b01.STG19
+ 2.8169 * refl_b02.STG19
+ 2791.3134 * NDVI.STG19
- 0.3435 * refl_b01.STG23
- 0.0639 * refl_b02.STG23
+ 3113.284 * NDVI.STG23
+ 0.7701 * refl_b01.STG32
- 1.1499 * refl_b02.STG32
- 635.7881 * NDVI.STG32
```

```
- 18922.3577

LM num: 8
.outcome =
-271.7469 * AlbarinoVariety
+ 2980.4495 * LoureiroVariety
+ 2959.5654 * TreixaduraVariety
- 2.5103 * P
+ 1296.5386 * K
+ 140.0604 * Ca
+ 2335.9251 * pHWater
+ 257.8772 * M.O.
- 0.7967 * refl_b01.STG19
+ 2.2222 * refl_b02.STG19
+ 2791.3134 * NDVI.STG19
+ 0.0469 * refl_b01.STG23
- 0.0639 * refl_b02.STG23
+ 4305.464 * NDVI.STG23
+ 0.7701 * refl_b01.STG32
- 1.1499 * refl_b02.STG32
- 635.7881 * NDVI.STG32
- 17141.9845

LM num: 9
.outcome =
-271.7469 * AlbarinoVariety
+ 3032.9943 * LoureiroVariety
+ 2959.5654 * TreixaduraVariety
- 4.1376 * P
+ 1296.5386 * K
+ 140.0604 * Ca
+ 2335.9251 * pHWater
+ 257.8772 * M.O.
- 0.7967 * refl_b01.STG19
+ 2.2222 * refl_b02.STG19
+ 2791.3134 * NDVI.STG19
+ 0.059 * refl_b01.STG23
- 0.0639 * refl_b02.STG23
+ 4428.793 * NDVI.STG23
+ 0.7701 * refl_b01.STG32
- 1.1499 * refl_b02.STG32
```

```
- 635.7881 * NDVI.STG32
- 16607.8255

LM num: 10
.outcome =
-271.7469 * AlbarinoVariety
+ 2983.7935 * LoureiroVariety
+ 4925.0619 * TreixaduraVariety
+ 13.2208 * P
+ 1296.5386 * K
+ 341.9429 * Mg
+ 140.0604 * Ca
+ 3211.0296 * pHWater
+ 361.6778 * M.O.
- 0.7967 * refl_b01.STG19
+ 2.2222 * refl_b02.STG19
+ 2791.3134 * NDVI.STG19
- 0.0702 * refl_b01.STG23
- 0.0639 * refl_b02.STG23
+ 3113.284 * NDVI.STG23
+ 0.7701 * refl_b01.STG32
- 1.1499 * refl_b02.STG32
- 635.7881 * NDVI.STG32
- 22252.3621

LM num: 11
.outcome =
-271.7469 * AlbarinoVariety
+ 1466.4612 * LoureiroVariety
+ 1744.6079 * TreixaduraVariety
+ 58.9354 * P
+ 1296.5386 * K
+ 140.0604 * Ca
+ 2239.9206 * pHWater
- 0.7967 * refl_b01.STG19
+ 1.67 * refl_b02.STG19
- 1573.624 * NDVI.STG19
- 0.0496 * refl_b01.STG23
- 0.0639 * refl_b02.STG23
+ 3113.284 * NDVI.STG23
+ 0.7701 * refl_b01.STG32
```

```
- 1.1499 * refl_b02.STG32
+ 1441.294 * NDVI.STG32
- 14318.6034

LM num: 12
.outcome =
-271.7469 * AlbarinoVariety
+ 1466.4612 * LoureiroVariety
+ 1744.6079 * TreixaduraVariety
+ 56.079 * P
+ 1296.5386 * K
+ 140.0604 * Ca
+ 3051.8794 * pHWater
- 70.8351 * M.O.
- 0.7967 * refl_b01.STG19
+ 1.67 * refl_b02.STG19
+ 567.666 * NDVI.STG19
+ 0.0648 * refl_b01.STG23
- 0.0639 * refl_b02.STG23
+ 3113.284 * NDVI.STG23
+ 1.3043 * refl_b01.STG32
- 2.2705 * refl_b02.STG32
+ 422.3481 * NDVI.STG32
- 14775.9228

LM num: 13
.outcome =
-271.7469 * AlbarinoVariety
+ 1466.4612 * LoureiroVariety
+ 1744.6079 * TreixaduraVariety
+ 56.079 * P
+ 1296.5386 * K
+ 140.0604 * Ca
+ 2742.9288 * pHWater
- 106.2526 * M.O.
- 0.7967 * refl_b01.STG19
+ 1.67 * refl_b02.STG19
+ 567.666 * NDVI.STG19
+ 0.0648 * refl_b01.STG23
- 0.0639 * refl_b02.STG23
+ 3113.284 * NDVI.STG23
```

```
+ 1.6109 * refl_b01.STG32
- 2.4447 * refl_b02.STG32
+ 422.3481 * NDVI.STG32
- 12814.7864

LM num: 14
.outcome =
-271.7469 * AlbarinoVariety
+ 1466.4612 * LoureiroVariety
+ 1744.6079 * TreixaduraVariety
+ 84.2158 * P
+ 1296.5386 * K
+ 140.0604 * Ca
+ 2596.5497 * pHWater
- 0.0793 * refl_b01.STG19
+ 1.67 * refl_b02.STG19
+ 567.666 * NDVI.STG19
+ 0.0648 * refl_b01.STG23
- 0.0639 * refl_b02.STG23
+ 3113.284 * NDVI.STG23
+ 1.1088 * refl_b01.STG32
- 2.2532 * refl_b02.STG32
+ 422.3481 * NDVI.STG32
- 16423.198

LM num: 15
.outcome =
-2669.979 * AlbarinoVariety
+ 1808.6665 * LoureiroVariety
+ 592.5172 * TreixaduraVariety
- 2.1252 * P
+ 2905.1114 * K
+ 57.27 * Ca
- 507.2806 * pHWater
+ 1.1145 * refl_b01.STG19
+ 1.0964 * refl_b02.STG19
+ 1884.8598 * NDVI.STG19
+ 0.425 * refl_b01.STG23
+ 0.1987 * refl_b02.STG23
+ 3790.9026 * NDVI.STG23
+ 0.5894 * refl_b01.STG32
```

```
- 0.8095 * refl_b02.STG32
+ 664.4926 * NDVI.STG32
+ 1196.1389

LM num: 16
.outcome =
-408.5472 * AlbarinoVariety
+ 3640.5276 * LoureiroVariety
+ 592.5172 * TreixaduraVariety
- 6.7392 * P
+ 6617.3858 * K
+ 57.27 * Ca
- 183.2756 * pHWater
+ 173.1614 * M.O.
+ 1.5593 * refl_b01.STG19
+ 2.0087 * refl_b02.STG19
+ 4463.9743 * NDVI.STG19
+ 0.9811 * refl_b01.STG23
+ 0.0309 * refl_b02.STG23
+ 2199.3766 * NDVI.STG23
+ 3.2996 * refl_b01.STG32
- 3.8656 * refl_b02.STG32
+ 7879.468 * NDVI.STG32
- 5468.0753

LM num: 17
.outcome =
-408.5472 * AlbarinoVariety
+ 3912.3216 * LoureiroVariety
+ 592.5172 * TreixaduraVariety
- 14.533 * P
- 185.3004 * K
+ 57.27 * Ca
- 183.2756 * pHWater
+ 173.1614 * M.O.
+ 7.1825 * refl_b01.STG19
+ 1.1773 * refl_b02.STG19
+ 5983.576 * NDVI.STG19
+ 1.2246 * refl_b01.STG23
+ 0.0309 * refl_b02.STG23
+ 2199.3766 * NDVI.STG23
```

```
+ 1.9967 * refl_b01.STG32
- 1.7653 * refl_b02.STG32
- 358.5315 * NDVI.STG32
- 8801.3738

LM num: 18
.outcome =
-408.5472 * AlbarinoVariety
+ 3912.3216 * LoureiroVariety
+ 592.5172 * TreixaduraVariety
- 15.8784 * P
+ 1842.3785 * K
+ 57.27 * Ca
- 183.2756 * pHWater
+ 173.1614 * M.O.
+ 4.6891 * refl_b01.STG19
+ 1.0041 * refl_b02.STG19
+ 5983.576 * NDVI.STG19
+ 1.0149 * refl_b01.STG23
+ 0.0309 * refl_b02.STG23
+ 2199.3766 * NDVI.STG23
+ 1.9967 * refl_b01.STG32
- 1.7653 * refl_b02.STG32
+ 3537.1431 * NDVI.STG32
- 6011.6979

LM num: 19
.outcome =
-408.5472 * AlbarinoVariety
+ 3921.0097 * LoureiroVariety
+ 592.5172 * TreixaduraVariety
- 2.1252 * P
+ 57639.183 * K
- 222.0549 * Ca
- 183.2756 * pHWater
+ 415.5873 * M.O.
+ 1.0659 * refl_b01.STG19
+ 2.0087 * refl_b02.STG19
- 785.5589 * NDVI.STG19
+ 0.8686 * refl_b01.STG23
+ 0.0309 * refl_b02.STG23
```

```
+ 2199.3766 * NDVI.STG23
+ 0.6802 * refl_b01.STG32
- 0.9773 * refl_b02.STG32
+ 6067.6546 * NDVI.STG32
- 30707.6441

LM num: 20
.outcome =
-420.8601 * AlbarinoVariety
+ 1868.0503 * LoureiroVariety
+ 592.5172 * TreixaduraVariety
- 2.1252 * P
+ 3638.6707 * K
+ 57.27 * Ca
- 183.2756 * pHWater
- 4.7948 * refl_b01.STG19
+ 9.1932 * refl_b02.STG19
- 26501.6593 * NDVI.STG19
+ 0.515 * refl_b01.STG23
+ 1.0114 * refl_b02.STG23
+ 2199.3766 * NDVI.STG23
+ 1.2213 * refl_b01.STG32
- 1.0041 * refl_b02.STG32
+ 6406.2639 * NDVI.STG32
- 10552.5871
```