Statistical analysis of semi-natural areas and unsupervised climate clustering for the Alentejo region, Portugal

Filipa Inês Tavares Vilhena

Thesis to obtain the Master of Science Degree in

Computer Science and Engineering

Supervisors: Prof. Ana Gualdina Almeida Matos
Prof. Ana Patricia Subtil da Graça Freitas Garcia

Examination Committee

Chairperson: Prof. Diogo Manuel Ribeiro Ferreira
Supervisor: Prof. Ana Gualdina Almeida Matos
Member of the Committee: Prof. Patrícia Maria Nunes Tiago

June 2022

This work was created using LaTeX typesetting language
in the Overleaf environment (www.overleaf.com).

Acknowledgments

I would first like to thank my mother, for always working and ambitioning a proper education for me. Not only for financially supporting my studies, but also for the sometimes undervalued, but always crucial, logistical and emotional support.

In the same spirit, I would also to thank my partner, for his support and patience throughout this journey. It was not an easy one to share, but I'm so thankful that he did.

I would also like to thank my dissertation supervisors, Prof. Ana Matos and Prof. Ana Subtil. Not only for their knowledge and insight, but for staying by my side through difficult moments, always with friendship and understanding. Also, for their availability for numerous long meetings and for my never ending questions. This MSc dissertation was supported by Instituto de Telecomunicações.

A special thank you to the NBI team, particularly Prof. Hugo Rebelo and Mr. Francisco Marques. Their continuous guidance, both with software support and with a deep understanding of ecology and the Alentejo area, were vital for this dissertation.

Last, but not least, I would like to acknowledge my close friends for their companionship through these months, and for keeping me company while I worked until the early hours of the morning. Without them I wouldn't have been able to complete this.

To everyone I mentioned and to everyone I may have forgotten to mention—Thank you.

Abstract

Vineyards require specific ecological contexts for all stages of their life cycle, such as appropriate temperatures, sufficient water availability and some pest control, to maximize wine yield and quality. In Mediterranean Portugal, the region Alentejo is a known wine producer, with very high temperatures in the Summer and a worsening drought situation. Alentejo also contains multitudes of different regions; most of them derived from human settlement and land exploration, and some areas that are still semi-natural. Research published in 2021 showed a correlation between the landscape surrounding vineyards and pest outbreaks, in across 400 Spanish vineyards, suggesting that a landscape of semi-natural habitats may lower pests, decrease insecticide use and improve biodiversity. This thesis gathered 19 bioclimatic variables from the CHELSA dataset, as input for the ISO Cluster unsupervised algorithm, a variation of k-means, creating 10 clusters with specific climate characteristics. Those clusters were then integrated with a statistical description of semi-natural areas across the region, through land cover data from DGT — COS 2018. This project aims to provide an initial, descriptive statistical study of Alentejo, to, in future work, be possible to analyse semi-natural habitats and vineyard productivity.

Keywords

Portugal; climate description; land cover; GIS; vineyard.

Resumo

As vinhas requerem contextos ecológicos específicos para todas as fases do seu ciclo de vida, tais como temperaturas apropriadas, disponibilidade suficiente de água e algum controlo de pragas, para maximizar o rendimento e a qualidade do vinho. Em Portugal, um país mediterrâneo, a região do Alentejo é um produtor de vinho conhecido, com temperaturas muito elevadas no Verão e uma situação de seca em agravamento. O Alentejo conta com diferentes regiões; a maioria derivada da manipulação humana e da exploração da terra, e algumas áreas ainda semi-naturais. Pesquisas publicadas em 2021 mostram uma correlação entre a paisagem circundante das vinhas e os surtos de pragas, em 400 vinhas espanholas, sugerindo que habitats semi-naturais em redor da vinha podem reduzir as pragas, diminuir a utilização de insecticidas e melhorar a biodiversidade. Esta tese recolheu 19 variáveis bioclimáticas do conjunto de dados CHELSA e utilizou o algoritmo não supervisionado ISO Cluster, uma variação do algoritmo k-means, para criar novas regiões com características climáticas específicas. Estes clusters foram então integrados com uma descrição estatística de áreas semi-naturais em toda a região, através de dados de cobertura do solo da DGT, COS 2018. Este projecto visa fornecer um estudo estatístico inicial e descritivo do Alentejo, para, em trabalhos futuros, ser possível relacionar-se habitats semi-naturais com a produtividade das vinhas.

Palavras Chave

Portugal; descrição climática; carta de ocupação do solo; vinhas; GIS.

Contents

List of Figures

List of Tables

Acronyms

ANOVA Analysis of Variance

CHELSA Climatologies at high resolution for the earth’s land surface areas

COS *Carta de Ocupação dos Solos*

CRS Coordinate Reference System

CVRA Alentejo Regional Winegrowing Commission

DGT Directorate-General for the Territory

DOC *Denominação de Origem Controlada*

GIS Geographic Information System

NBI Natural Business Intelligence

NUTS Nomenclature of Territorial Units for Statistics

PC principal component

PCA Principal Component Analysis

POI Point of Interest

1

Introduction

Contents

1.1 Context and Motivation

Viticulture, the cultivation of grapes for the main purpose of producing wine, is a practice with a long history and still substantial throughout Europe. Particularly in the Mediterranean, wine is embedded in several countries' cultural identity, and grape cultivation is very significant. Two Portuguese regions, according to Nomenclature of Territorial Units for Statistics (NUTS) 2 from the European Union, dedicate more than 8% of their agricultural area to vineyards [1], from cooperatives and big companies to small wineries.

This thesis is a conjoined effort made by members of Instituto Superior Técnico and the company Natural Business Intelligence (NBI), regarding wine growing and its climatic and geographical context. NBI is a Portuguese environmental consultancy group, founded in the beginning of 2020. The group focuses on aiding companies improve their ecological impact, implementing different farming strategies to improve the surrounding environment, or to minimize prejudicial impact. NBI is currently partnered with Alentejo Regional Winegrowing Commission (CVRA), an organisation dedicated to certifying and protecting wines from a specific region in Portugal, Alentejo.

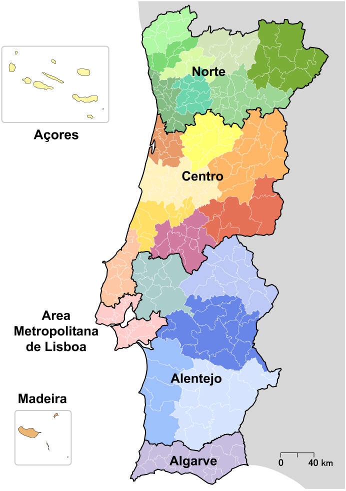


Figure 1.1: .
Portuguese NUTS2 region classification in 2013, with subregions.

Alentejo, the blue zone [2] in Figure 1.1, is the region on which this project will focus. Making up to one third of mainland Portugal's area, the climate is temperate, with dry, warm Summers and mild Winters. It is currently posing some challenges for agriculture, as it is the driest region in Portugal, and has some of the highest Summer temperatures. As the climate changes, the rising temperatures and less water availability may turn some areas inhospitable [3]. It becomes relevant to study new areas on which agricultural areas may be installed according to their ecological needs, for climate change adaptation [3, 4].

It is one of many wine-producing regions in Portugal, with its own wine classification label, *“DOC – Alentejo”*. *Denominação de Origem Controlada* (DOC), or Label of Controlled Origin, refers to wines that grow on a specific geographical location, follow a rigorous production method and have a higher standard of quality. Some producers on the Alentejo region, having met these requirements, produce DOC wines on the subregions pictured on fig. 1.2.

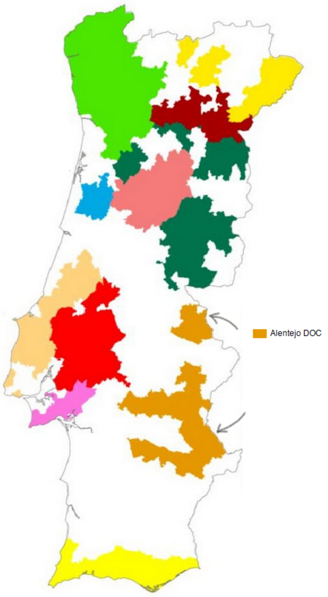


Figure 1.2: DOC wine growing regions in Portugal.

In the last years, an organic agricultural practice has been demonstrated to have positive effects, regarding changes in soil health, diversified biodiversity and higher product quality. In 2021, a study presented that the landscape surrounding vineyards may be also relevant: habitats classified as semi-natural were proven to decrease pests and improve biodiversity [5]. Therefore, it may be relevant to study surrounding land usage when managing a vineyard, which is the purpose of this project. This thesis will be the first step of a larger study, that intents to correlate these semi-natural areas with vine productivity, suggest a reclassification of the Alentejo region, and lead to a more resilient use of the land.

1.2 Objectives

This dissertation’s main objective is to assist decisions by future wine producers, such as the location of new vineyards or which grape varieties to plant. New regions in the Alentejo area will be proposed to aid the decision-making, created according to climatic data, and described by the presence of semi-natural habitats, such as nearby bodies of water or nearby native forests.

In future work, this project will integrate a larger research produced by NBI, that aims to associate climate and land cover data to vineyard productivity. By studying the presence of specific habitats and its impact on grape yield, wine producers may be advised to install vineyards near some habitats, and steer away from others.

Some sub-objectives of this thesis are:

land cover and climate statistical analyses The thesis should provide a statistical description of the area according to climatic variables and according to land usage.

Regionalization regarding climate New areas will be defined with consistent climatic properties, to assure climate stability in certain regions and to showcase differences in Alentejo

Description of newly calculated regions These new areas will be described according to their land cover, so producers know what would surround a possible new vineyard, and how that might impact the environment.

1.3 Contributions

This dissertation will contribute with proposed new areas that are climatically stable, so that climate might become a fixed parameter and other variables might be analysed, like land cover or other dataset at will.

This dissertation will also produce statistical analysis of semi-natural habitats on Alentejo, that may contribute to an improved land planning, according to each region's ecological needs and traits.

1.4 Thesis structure

This document is organised as follows:

Chapter 2 presents some literature on climate impact on vineyards, and how surrounding habitats may impact production.

Chapter 3 explains the datasets employed, and some technical information regarding location data and the used GIS software.

Chapter 4 contains the developed work, and is split into five sections: Data Preparation; Statistical Background, where the necessary mathematical concepts are described; Analysis of Climate Data, where the new regions are defined; land cover Data, where the *Carta de Ocupação dos Solos* (COS) is analysed, and Criteria Compilation, where we calculate the final statistics.

Chapter 5 presents the results of the previous chapter, an evaluation and a discussion.

Chapter 6 indicates some limitations of the project, mentions possible future work, and concludes the dissertation.

2

Literature review and Background

This chapter will share some light on the context and importance of surrounding areas for a viticulture practice.

The ecological context of a plant rules its development. Variables related to the climate may affect the plant's health [6], therefore making their analysis of vital importance. The highest and lowest temperatures the vine experiences, as well as occurrence of frost or lack of precipitation, may prejudice its development [7, 8], which will in turn affect wine quality [9] and yield [10]. In a Mediterranean climate like Portugal, it has not been common to have extreme weather events like prolonged frost, snow or consistently very high temperatures; so far, Portugal has been recognized as one of the best countries for wine growing, along with others in the Mediterranean area, but climate change should be accounted for in future wine growing implementations.

In Alentejo, a region known for its high Summer temperatures, low precipitation and low water availability, wine producers have already established the best cultivars for specific zones, regarding their preferred temperature and precipitation values, altitude and soil; however the surroundings of vineyards are often overlooked.

In order to maximize yield, an absence of pests and diseases should also be sought for. One of the first studies of its kind, an article by Paredes et al [5] correlated habitats surrounding Spanish vineyards with pest outbreaks. Organic farming is an approach proved to be effective as a more natural pest control, by introducing animals that are natural enemies of common pests, or by planting specific plants that repels nefarious microbes and insects [11]. With the changing climate, pest outbreaks and fungal diseases may increase their impact on viticulture, and pesticides start to become less effective as pests grow more resistant [12].

The aforementioned article modelled landscape effects on pest infestations and insecticide applications, and concluded that vineyards that were surrounded with simplified landscapes, such as more vineyard area, had a significantly higher pest outbreak and need of insecticide application. On the other hand, vineyards surrounded by semi-natural habitats would have four times less outbreaks of the studied species, the European grapevine moth.

We proposed to study Alentejo's semi-natural habitats. Since there is not a developed research in what constitutes a semi-natural area, we had to analyse, one by one, land cover classes of Alentejo. As described in the following chapter, the COS dataset provided us with 11 classifications, which the NBI team divided to 13 to clarify and distinguish some habitats. Variables that might prove beneficial to the environment of the vineyard were chosen to be Water, Riparian Area, Native Forest and Shrubland. Agricultural habitats, due to some practices by producers, negatively impact the soil with chemicals and contaminated animal waste. Urban areas and low vegetation areas do not provide enhanced biodiversity or any ecological benefit, and pastures, also due to wrongful management, strip the soil of vegetation and any other resource. Pine forests and eucalyptus, for their high water needs and for being an invasive species in Alentejo, were also considered non beneficial for vineyards. Native forests are mindful and appropriate of the Portuguese region, and water bodies provide a vital component for vine growth.

3

Data and Software

Contents

Location is a type of data that is increasingly useful to record. It is needed to create maps of cities, which in turn allows the creation of GPS, to register habitats or to plot mountains and lakes, ultimately to better know and explore the Earth we live in. Such type of data, due to its inherent relation to the planet and being so closely attached to the visual component of a map, has specific characteristics and needs. In this chapter we will explore how this data is described and how to do it universally, how it can be managed through software and which datasets were used in this project.

3.1 Tools and Software

Data that contains geospatial information falls under the scope of Geographic Information System (GIS). GIS is a system that manipulates this kind of data ('spatial data'), and arose from the technological advances of the 1960s and 1970s, when land description and planning were first being computationally calculated. Nowadays, GIS applications are more powerful: able to import spatial data, perform analyses, create new information and project it in maps, with two or three dimensions, with descriptive information.

This dissertation relies on GIS to analyse the datasets described in the following section. Through institutional access from Instituto Superior Técnico we were able to develop this project using the GIS application ArcGIS. Created by the company Esri, ArcGIS is a feature-rich tool and one of the most used in this field. It allows, among other functionalities, to visualize areas in 3-D, to perform statistical analyses on geographical data, or to edit and create new features, all through explicit menus and intuitive mouse commands. It also allows users, however, to program in Python for additional, personalized operations.

It is able to import some filetypes useful for geographical data, such as .csv, .tif or vector formats. In this project, the Climatologies at high resolution for the earth's land surface areas (CHELSA) dataset provides files of the type GEOtiff, while the COS dataset provides data in an Esri-only file type, .shapefile.

3.1.1 GIS datafiles – Vector and Raster

Vector data is an important type of data and the baseline for spatial representation. Vectors identify individual geographical features in three feature classes: points, lines and polygons [13].

Points represent discrete data elements. Having an area of zero, they are useful for marking Points of Interest (POIs) of irrelevant area or shape. Each point is internally represented by its coordinates. Lines, a one dimensional feature characterized by two separate points, are represented by their vertices. Polygons are a closed shape defined by connected points; they are used to show boundaries and convey area and shape information. Polygon features will be, from the three vector data types, the one most relevant in this project. Figure 3.1 exemplifies common uses for vector data types: points represent location, lines represent roads and polygon represent areas, such as green spaces or buildings.

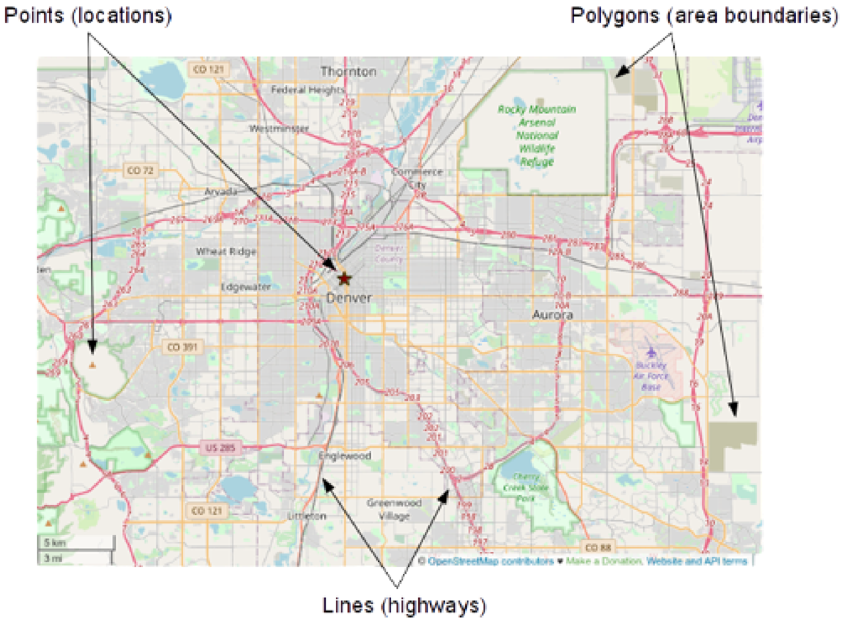


Figure 3.1: Examples of different vector data types in use. Base map from OpenStreetMap.

Besides being represented by their coordinates, features also possess attributes, extra information associated with a shape. Considering a simple database of points representing museums, each point may have the additional attributes “Name”, “Admission price” or “Closing hours”.

Alternatively, **raster** is a geospatial data that defines surfaces. Also known as grid data, rasters are a matrix of cells organized into rows and columns, and with a specific cell size (raster resolution). Within the scope of the project, raster data is the original data type of the CHELSA dataset and the main type of GIS data file used.

While vector data describes individual shapes, with no information about the space between them, rasters provide a way to describe geographical continuity. A classical use of raster data is for elevation, that has a different value for each cell of the grid. Points could be used to pinpoint mountain tops and the lowest point of valleys, but a considerable amount of information would be lost in-between.

The matrix’s cell size determines the level of detail of the raster and ultimately how accurate a surface is represented. A raster with a cell size of 10 kilometres may be enough to describe the shape of a city in the scope of a country, but not precise enough to describe elevation on a mountain range.

Raster resolution, although directly proportional to accuracy, comes with higher file size and need for processing power. Figure fig. 3.3 exemplifies how resolution can alter our perception of the subject. The most accurate representation from the polygon is the raster with smaller cell size, which brings an elevated number of rows and columns. The 2 meter cell raster loses some shape information, although

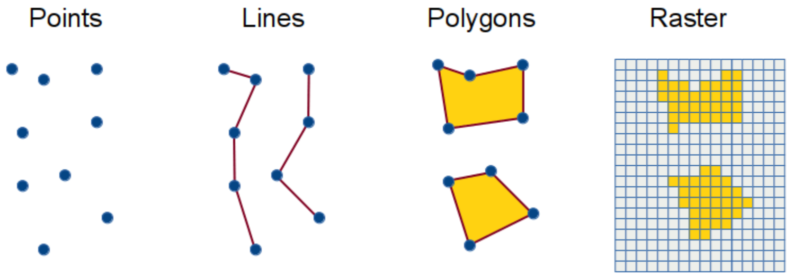


Figure 3.2: Rasters represent a continuous surface, rather than single shapes. [14]

maintaining an acceptable area, and the 4 meter cell raster exaggerates its proportions and loses all of its original shape. The less detailed raster, however, would be the one less computationally expensive and faster to load and operate on GIS software. Therefore, spatial resolution has to be carefully determined, balancing accuracy and computational speed.

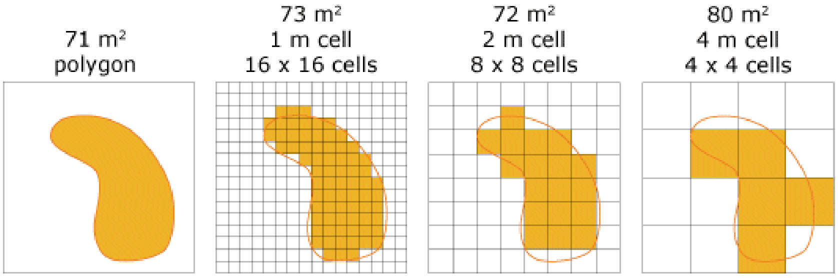


Figure 3.3: Raster resolution. Smaller cell size comes with higher accuracy and slower processing.

In the next section we will expound the factors behind the ideal raster resolution for this project, and how it was achieved using upscaling.

3.2 Dataset description

3.2.1 Land cover data

The land cover dataset COS used was provided by Directorate-General for the Territory (DGT) [15] and collected by the NBI team. For the months of June, July, August, September and October 2018 the data was produced by DGT through orthophotography, meaning it's derived from aerial photographs. Different than a standard picture, orthophotography accounts for Earth's topography and camera distortions, so an accurate picture with measurable distances can be obtained.Its spacial resolution is 25 meters, a highly detailed digital record, which allows for precise detection of land cover and each region's area.

For the remaining months of the year, the dataset collected information from other data sources available, like LUCAS (land cover and Coverage Area frame Survey) [16], CORINE Land Cover [17], and others. This is an accurate method, and logistically easier than through personal observation. Satellites photograph delimited areas, and through image recognition algorithms they are able to be automatically be classified; examples may be water bodies, shrubland, pastures, between other habitats.

More information about data collection methods and specifications may be found on DGT's technical document for the dataset [18].

A supplemental dataset was utilized for water bodies, obtained from the European Commission's Joint Research Centre within the Copernicus Programme [19]. This data was also collected through satellite imagery, derived from the Landsat dataset [20].

The original dataset by DGT provided all bodies of water of Alentejo, despite some of them only being active on some months of the year, possibly due to lack of precipitation and high temperatures. This dataset provides an additional attribute to their data, "Occurrence", that indicates, in percentage, how much time of the year a specific water body exists. The Guadiana river, and the water harnessed by the Alqueva dam have occurrence values superior to 95, while small ponds that disappear during the warmest months have occurrences of 20 or 25%. This dataset allowed for a much accurate assessment of water impact in Alentejo.

The COS classified the area in Alentejo into 9 major classes: (a) Artificial territory; (b) Agriculture; (c) Pastures; (d) Humid areas (e) Forests (f) Superficial water bodies (g) Agroforestry surfaces (h) Shrubland (i) Low Vegetation areas .

This classification is then divided into multiple subclassifications. The NBI team grouped many subclassifications and divided some classes, creating their own classification system with 13 categories:

1. ”Artificial territory” was renamed to ”Urban area”.
2. ”Agriculture” was divided into ”Agriculture”, ”Vineyards and orchards” and ”Olive groves”.
3. ”Pastures” remained the same classification.
4. ”Humid areas” was reclassified as ”Riparian area” and ”Water”.
5. ”Forests” was divided into ”Forests”, ”Pine forests” and ”Eucalyptus”.
6. ”Superficial water bodies”, grouped into ”Water”.
7. ”Agroforestry surfaces” were grouped into ”Forests”.
8. ”Shrubland” remained the same.
9. ”Low Vegetation areas” remained the same.

A comprehensive list of the original classifications, together with subclassifications and meanings, is annexed on the official technical guide [18].

3.2.2 Climate Data

The climate set of data was collected from CHELSA (Climatologies at high resolution for the earth's land surface areas) [21], by the NBI team. CHELSA is a dataset that consists of "downscaled model output temperature and precipitation estimates" for a climatological period from 1979-2013 [22], using data collected with a 6-hour period from the ERA-Interim dataset [23].

This dataset has a 30 arcsec resolution. While meter is a measure of distance commonly used on planar surfaces, arcsecond is an angular measurement, corresponding to $\frac{1}{3600}$ of a degree. Because of the spherical shape of the planet, arcsecond is a more precise way of describing the resolution [24]. When "flattened" and converted to meters, 30 arcsec are approximately 1000 meters.

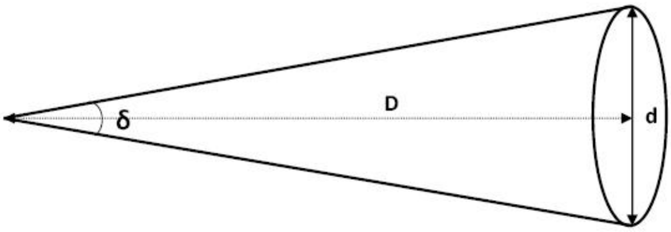


Figure 3.4: Angular diameter formula, of how an angular variable is converted to planar distance.

The baseline for climate data are individual registries from local weather stations and observatories, that create a local report and analysis of weather. These, stations, however, have a limited geographical scope. If the temperature in one city is registered, and it is also registered at a weather station at 100 km, there is no measurement for the area in between. Downscaled models of temperature or precipitation utilize climate records, which provides accurate readings for a finite number of locations, and interpolate it for a larger area using mathematical models.

ERA-Interim calculated that interpolation using specific models, which CHELSA reformulated using their own formulas, correcting possible biases.

For instance, for the mean daily air temperatures, CHELSA applied linear regression for each grid cell of ERA-Interim monthly means of daily temperature, and calculated lapse rates Γ_d according to the recorded pressure levels, in hPa. Using the lapse rates, temperature at sea level t_0 was also interpolated, and, for each cell, the elevation was gathered from a Global Terrain Elevation Data dataset. The final formula was

$$t = \Gamma_d * elev + t_0$$

For land surface temperatures, lst, the main component of this project's bioclimatic variables, CHELSA based the formula on the SRAD model, by Wilson and Gallant [25]:

$$lst = t - \Gamma_d(z_h - z_c) + C(S\frac{1}{S})(1\frac{0.1}{8})$$

in which z values correspond to elevation, t corresponds to the temperature grid, C is a constant that was chosen to be 1.0, and S is the short-wave radiation ratio.

3.2.2.A Bioclimatic variables description

This model was used to calculate 19 bioclimatic variables, known as being biologically meaningful and used for ecological and climatic models [26]. A compiled table of these variables, with respective titles and units, was adapted from the official dataset source [22] and is presented in fig. 3.5.

Bio01

- Definition: Annual Mean Temperature (℃/10)
- Calculation: average temperature for each month of the 1979-2013 period. Each month of this year set was averaged ($\frac{January\_1979+January\_1980+...}{34}$), and lastly the mean of these final twelve months was calculated as well.
- Usage: A simplistic temperature variable, for a general sense of average temperatures.

Bio02

- Definition: Mean diurnal air temperature range (℃/10)
- Calculation: average diurnal temperature range ($T_{max} - T_{min}$) for each month of the year set.
- Usage: it is useful to see how temperatures range between the various months. A large number means the highest temperatures and the lowest are consistently very different.

Bio03

- Definition: Isothermality (℃/10)
- Calculation: $Bio02/Bio08 * 100$
- Usage: Isothermality quantifies how temperatures oscillate during the day, relative to the seasonal oscillations.

Bio04

- Definition: Temperature Seasonality (℃/10)
- Calculation: standard deviation of the mean monthly temperatures
- Usage: variable to measure range of temperatures, in this case how they deviate from the mean.

Bio05

- Definition: Mean daily maximum air temperature of the warmest month (℃/10)

- Calculation: maximum temperature across all monthly means of the year set.
- Usage: useful to see the maximum mean temperature a specific pixel of the map reached in the last years, which is important to prevent a potential sunburn

Bio06

- Definition: mean daily maximum air temperature of the coldest month (℃/10)
- Calculation: minimum temperature of any monthly mean temperature values
- Usage: may be valuable to detect too low temperatures for the vine

Bio07

- Definition: Temperature Annual Range (℃/10)
- Calculation: $Bio05 - Bio06$
- Usage: the highest temperature a pixel reached minus its lowest temperature. Another useful variable for temperature range; a high number indicates high variability between the warmest and coldest seasons.

Bio08

- Definition: Mean air temperature of the wettest quarter (℃/10)
- Calculation: together with the precipitation values, the quarter (set of three months) with the highest cumulative precipitation was calculated for each year, and that quarter's temperatures were collected. These temperatures for different years were then averaged for the whole year set.
- Usage: useful for insight about temperatures when precipitation is abundant.

Bio09

- Definition: Mean air temperature of the driest quarter (℃/10)
- Calculation: similar to $Bio08$, but the lowest cumulative precipitation was sought
- Usage: similar to $Bio08$, it may be a useful variable to relate temperature and precipitation.

Bio10

- Definition: Mean air temperature of the warmest quarter (℃/10)
- Calculation: for each year, the three-month set with the highest mean temperature was calculated, and these values were subsequently averaged.
- Usage: general notion of the temperature of the warmest season of the year.

Bio11

- Definition: Mean air temperature of the coldest quarter (℃/10)
- Calculation: similar to $Bio10$, but the lowest mean temperature was sought

- Usage: similar to $Bio10$, provides a general knowledge of temperatures on the coldest season.

Bio12

- Definition: Annual Precipitation Amount (kg/m^2)
- Calculation: cumulative precipitation values of each year were summed, and then averaged according to the 34-year time period.
- Usage: may be an important measure when accounting for total water availability for a specific area, for the whole year

Bio13

- Definition: Precipitation Amount of Wettest Month (kg/m^2)
- Calculation: maximum value between precipitation amounts for each year, averaged
- Usage: might be useful to account for extreme climate conditions, such as floods or extreme lack of precipitation. It might not have been the same month throughout the year set.

Bio14

- Definition: Precipitation Amount of Driest Month (kg/m^2)
- Calculation: similar to $Bio13$, but with the lowest precipitation value, averaged
- Usage: similar to $Bio13$. A particularly high value in this variable means the water availability is high throughout the year.

Bio15

- Definition: Precipitation Seasonality (kg/m^2)
- Calculation: standard deviation of precipitation amounts for each month, averaged for the time period
- Usage: Coefficient of Variation of precipitation. A high value indicates greater water variability, which may be important for some species that require a constant water amount.

Bio16

- Definition: Mean monthly precipitation amount of the wettest quarter (kg/m^2)
- Calculation: similar to ($Bio08$), with calculated precipitation amounts.
- Usage: more than knowing the wettest month, this variable refers to the wettest three-month period, guaranteeing insight about precipitation values for a larger period of time.

Bio17

- Definition: Mean monthly precipitation amount of the driest quarter (kg/m^2)
- Calculation: similar to ($Bio16$)
- Usage: similar to ($Bio16$), it refers to precipitation values for the driest period of the year, which

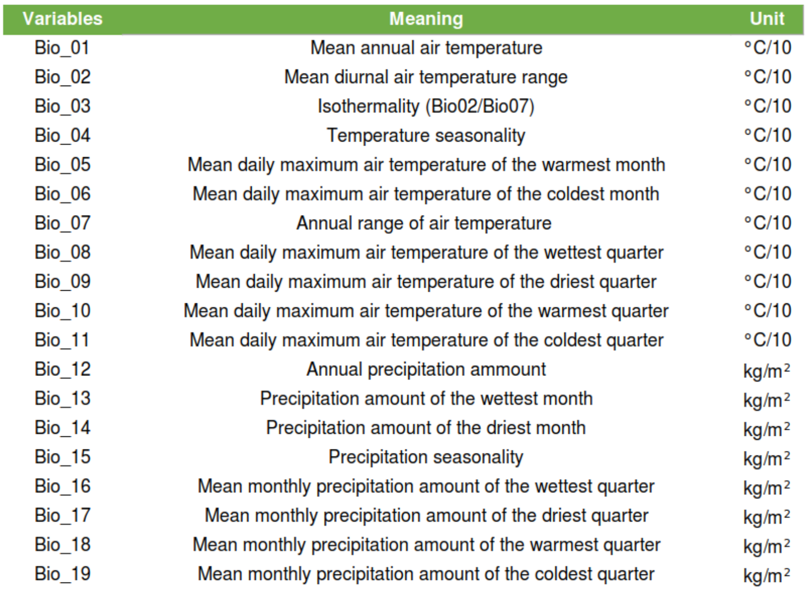
may be considerably more than the driest month.

Bio18

- Definition: Mean monthly precipitation amount of the warmest quarter (kg/m^2)
- Calculation: for each year, the three-month period with the highest mean temperature was found, its precipitation calculated, and lastly averaged.
- Usage: useful to relate precipitation to temperatures, especially in the warmest quarter, when water needs are more prevalent.

Bio19

- Definition: Mean monthly precipitation amount of the coldest quarter (kg/m^2)
- Calculation: similar to $(Bio18)$
- Usage: might provide useful information of how precipitation relates to low temperatures, in a period where vines might be in dormancy.



Variables	Meaning	Unit
Bio_01	Mean annual air temperature	°C/10
Bio_02	Mean diurnal air temperature range	°C/10
Bio_03	Isothermality (Bio02/Bio07)	°C/10
Bio_04	Temperature seasonality	°C/10
Bio_05	Mean daily maximum air temperature of the warmest month	°C/10
Bio_06	Mean daily maximum air temperature of the coldest month	°C/10
Bio_07	Annual range of air temperature	°C/10
Bio_08	Mean daily maximum air temperature of the wettest quarter	°C/10
Bio_09	Mean daily maximum air temperature of the driest quarter	°C/10
Bio_10	Mean daily maximum air temperature of the warmest quarter	°C/10
Bio_11	Mean daily maximum air temperature of the coldest quarter	°C/10
Bio_12	Annual precipitation ammount	kg/m^2
Bio_13	Precipitation amount of the wettest month	kg/m^2
Bio_14	Precipitation amount of the driest month	kg/m^2
Bio_15	Precipitation seasonality	kg/m^2
Bio_16	Mean monthly precipitation amount of the wettest quarter	kg/m^2
Bio_17	Mean monthly precipitation amount of the driest quarter	kg/m^2
Bio_18	Mean monthly precipitation amount of the warmest quarter	kg/m^2
Bio_19	Mean monthly precipitation amount of the coldest quarter	kg/m^2

Figure 3.5: Climate variables, respective meaning and units.

4

Experimental Process

Contents

In this chapter, we will go through the experimental process, from the initial data cleaning to the creation of Principal Component Analyses (PCAs) and region demarcation. We will also explain the statistical tools applied.

4.1 Data preparation

4.1.1 Land cover data

DGT originally provided the COS dataset fragmented into sections of Portugal, for easier handling and downloading. Those sections, which include, but are not limited to, North Alentejo, South Alentejo and Coastal Alentejo, were previously compiled into a single shapefile (file of vector features) by the NBI team, and clipped to remove unnecessary map area. After the compilation and clipping, the COS files received no further preparation.

The specific dataset used for the water habitat was gathered from the online repository, where it was split by many small areas of the globe. We selected the .tif file that included Portugal, this file was then clipped to maintain the area at study, and filtered by the attribute "Occurrence", with $occurrence >= 50\%$, to only keep polygons of water that are actively present for at least half of the year.

4.1.2 Climate data

The bioclimatic variables were considerably processed by NBI and already ready to be utilized from the beginning of this project. The dataset, as mentioned, offers their data separated by twelve parts, corresponding to each month of the 1979-2013 time period. Since the scope of this project was not a climate variation throughout the months, the NBI team averaged the twelve months in advance, resulting in the mean of values of each variable from 1979 to 2013. The geographical range of this dataset is global, so each file also had to be clipped to exclusively maintain the Alentejo area.

4.2 Mathematical background

4.2.1 PCA – Statistical analysis

The main statistical tool in this project is the Principal Component Analysis (PCA), a technique widely utilized to lower data dimensionality and improve computational efficiency. PCA, most appropriately used for a dataset of three or more variables, summarizes data into new and relevant features, leaving behind redundant information or outliers that could potentially interfere with final results. In this particular context, PCA was not only a cleaning aid, but was also used to portray data visually and reduce it to the essential information.

An important concept in this tool is the concept of variance, mathematically symbolised by σ^2. In statistics, variance is a measure of dispersion of a variable, meaning how far the values of a variable are to its mean.

Given multiple variables, the one with higher variance will explain data variability better than others. A variable with low variance may not be relevant: in a dataset where samples refer to wine, and are described by their properties, like alcohol content, price or acidity, a certain property with the same value consistently for every wine sample (low variance) would not aid the characterization. With this technique, that variable would have a very low attributed importance.

To apply a PCA, features are expected to go through a normalisation process. Normalisation is necessary for PCA due to that consideration of variance. Considering two example features, if the variance of the first variable is significantly higher than the other, this would wrongly lead to PCA considering the first variable a better descriptor of the dataset. With it, all variables have a much more similar variance value, and are considered of equal importance for this tool, which is the intended. Normalisation consists of rescaling all values to a specific range, in this project being from 0 to 1.

The formula utilized to perform that transformation, having $\mathbf{x} = (x_1, ..., x_n)$ as a set of n observations, is:

$$x_{i,normalized} = \frac{x_i - \min(\mathbf{x})}{\max(\mathbf{x}) - \min(\mathbf{x})}, \tag{4.1}$$

with $i = 1, ..., n$.

The denominator, $max(x) - min(x)$, is also named the range of the dataset.

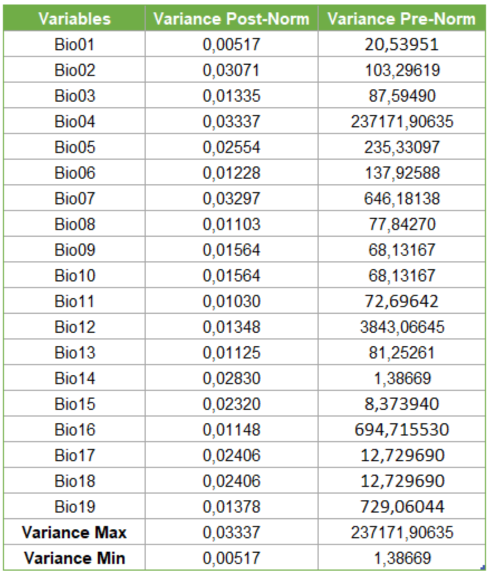
Having determined the minimum and maximum values, each value is subtracted by the minimum and divided by the range, resulting in a rescaled dataset, but with identical variation between values.

Figure 4.1 displays the variance of the bioclimatic variables before normalisation, and after, and the maximum and minimum values for range comparison.

The output of this technique is principal components (PCs), which are linear transformations of the original data. The first principal component is a vector that represents the maximum variance direction, the second principal component describes the second most important variance direction, and so on. In Figure 4.2 below, a graph describes a simple two-dimensional dataset. The direction in which the variability is most significant is displayed in red, and it is the first PC returned by the technique.

Referring to the wine example, the first PC could be $0.5 * acidity + 2 * alcohol\_content$, meaning that this linear combination would be a more summarized way with which we could describe a wine sample, together with the rest of the PCs.

The calculus of each PC, how that linear combination is calculated, comes from the eigendecomposition of the original matrix. The premise behind eigendecomposition, the creation of eigenvalues and



Variables	Variance Post-Norm	Variance Pre-Norm
Bio01	0,00517	20,53951
Bio02	0,03071	103,29619
Bio03	0,01335	87,59490
Bio04	0,03337	237171,90635
Bio05	0,02554	235,33097
Bio06	0,01228	137,92588
Bio07	0,03297	646,18138
Bio08	0,01103	77,84270
Bio09	0,01564	68,13167
Bio10	0,01564	68,13167
Bio11	0,01030	72,69642
Bio12	0,01348	3843,06645
Bio13	0,01125	81,25261
Bio14	0,02830	1,38669
Bio15	0,02320	8,373940
Bio16	0,01148	694,715530
Bio17	0,02406	12,729690
Bio18	0,02406	12,729690
Bio19	0,01378	729,06044
Variance Max	0,03337	237171,90635
Variance Min	0,00517	1,38669

Figure 4.1: Variance table pre- and post-normalization.

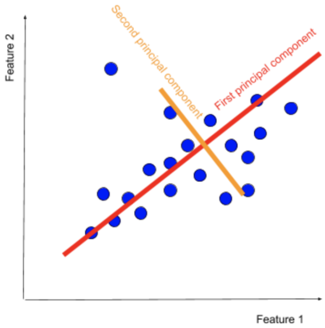


Figure 4.2: Example of PCA applied to a two-dimensional dataset.

eigenvectors, is that any matrix, multiplied by a vector υ, outputs the same as a certain value λ multiplied by that same vector. Each of these values λ are eigenvalues, and the vectors υ are eigenvectors (demonstrations present on [27]).

This relationship between eigenvalues, λ, eigenvectors, υ, and the original matrix M can be described as:

$$M * \upsilon = \lambda * \upsilon \tag{4.2}$$

which is equivalent to

$$M * \upsilon - \lambda * \upsilon = 0 \equiv (M - \lambda) * \upsilon = 0 \tag{4.3}$$

Internally, this technique will first calculate the eigenvalues of M, by solving the first part of the equation,

$(M - \lambda)$. In case of a 2x2 M, it was demonstrated to be equivalent to:

$$M - \lambda * \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix} \equiv M - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix} \equiv \begin{bmatrix} (M_1 - \lambda) & M_2 \\ M_3 & (M_4 - \lambda) \end{bmatrix}$$

When replaced on eq. (4.3), we have

$$\begin{bmatrix} (M_1 - \lambda) & M_2 \\ M_3 & (M_4 - \lambda) \end{bmatrix} * v = 0$$

When the procedure calculates the determinant of $\begin{bmatrix} (M_1 - \lambda) & M_2 \\ M_3 & (M_4 - \lambda) \end{bmatrix}$ and equates it to 0, it will obtain the eigenvalues of the matrix, which are as many as the number of rows and columns.

From the eigenvalues it quickly obtains the eigenvectors, through the formula 4.2, and we have reached our new variables.

4.2.2 Clustering – Machine Learning

Clustering is a widely used method for grouping information. Given a set of observations, clustering algorithms intend to assemble such observations into groups, according to user-defined variables. They can be applied to many fields, such as (a) the medical field, in which patients might be grouped by risk of disease; (b) product recommendations, by clustering customers' preferences and defining profiles; or (c) to group documents or music by topic or genre.

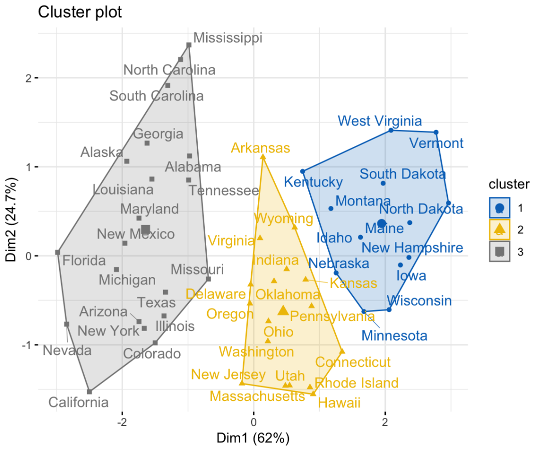


Figure 4.3: Example of clustering algorithm on USA states [28].

When projected on a two-dimensional graph, it becomes easier to visualize clusters. The figure below, fig. 4.3, exemplifies a clustering algorithm applied to states of the United States of America, according to two Principal Components.

Between supervised and unsupervised machine learning, clustering is less commonly used for supervised. A supervised learning algorithm refers to a set of observations in which samples are already labelled, and the model uses a small training sample to learn and be able to predict future observations. Given the previous example of wine data, if the dataset had a label variable for classifying wine as “high quality” or “low quality”, a supervised model would train with those examples. If the model was to be given a new wine sample, with new values, it would be able to classify it based on the previous learning.

On unsupervised learning, namely clustering, there are no labels for a model to learn and train with. Therefore, an unsupervised algorithm must cluster observations purely based on other factors, such as the distance between them.

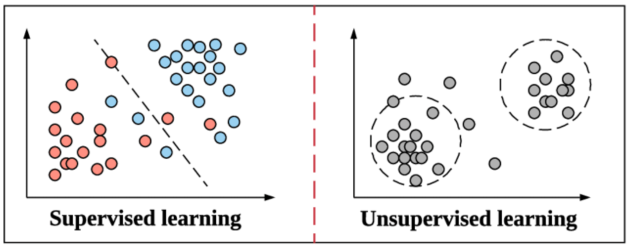


Figure 4.4: Example of supervised and unsupervised machine learning algorithms [29].

This project will use the ISO Cluster algorithm, a variation of K-Means, one of the most used techniques for unsupervised clustering. ISO Cluster is an algorithm that initializes K number of cluster centres (means), and assigns each sample to the nearest cluster. After the assigning step, the centroids are recalculated, samples are reassigned them to the nearest one, and this is done iteratively until convergence.

$$\underset{\mathbf{S}}{\arg\min} \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2$$

Figure 4.5: ISO Cluster, and K-Means, main formula of minimizing variance.

ISO Cluster attributes a sample to a specific cluster by calculating the variance within the cluster in each step. For $i = 1$ until k, and for each observation x of the dataset, it will try to minimise the absolute value between x and the cluster S_i mean, μ. The cluster in which this value is minimised is the cluster assigned to x.

The algorithm proceeds in the following manner:

- *k* number of clusters is chosen, and the algorithm initializes *k* aleatoric seeds to be the centroids of the clusters.

- For each observation, ISO Cluster will calculate the Euclidean distance between the sample and all centroids, on any number of dimensions;

- The centroid which produces the smallest distance value will be chosen, and that observation will belong to that cluster;
- This step is repeated for all observations of the dataset. In the end of this step, the centroids of these new clusters will be generated;
- For all observations, the algorithm will calculate if adding one to a specific cluster will increase or decrease the variance, and will choose the option that minimizes variance;
- When clusters converge and an iteration brings no new changes, the algorithm ends.

4.3 Analysis of climate data

This project was organized into three distinct steps, which are represented below in Figure 4.6:

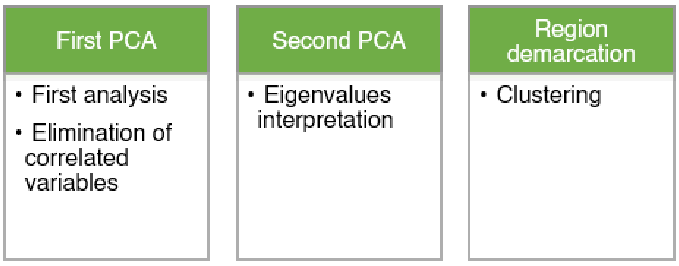

First PCA
• First analysis
• Elimination of correlated variables
Second PCA
• Eigenvalues interpretation
Region demarcation
• Clustering

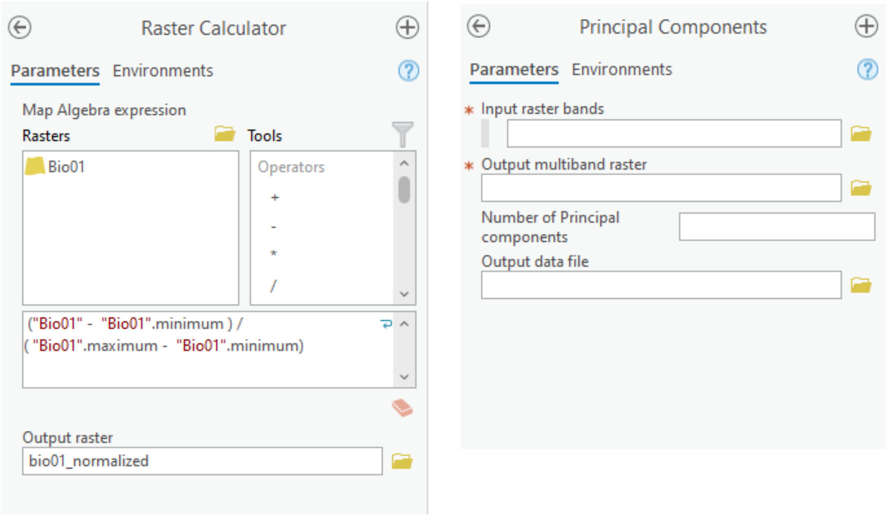
Figure 4.6: Steps of climate study until region demarcation.

All 19 bioclimatic variables were analysed for their correlation, the ones which provided redundant data were removed, and it was performed a final PCA. This second PCA showcased the most relevant bioclimatic features, and created new variables, with which we have performed a clustering algorithm to demarcate new regions in Alentejo.

4.3.1 Principal Component Analyses

The first PCA was created through the ArcGIS Pro menu “Principal Components”, pictured in Figure 4.7. The input raster bands in this initial PCA were the nineteen bioclimatic variables, normalised as explained above, and it created the maximum number of PCs. This allows each PC to be as accurate as possible.

With the Raster Calculator tool, the equation eq. (4.1) was applied to each bioclimatic variable. Selecting each raster from the list of variables currently on the Map, the *.minimum* and *.maximum* functions are available as ArcGIS may calculate statistics automatically. The output is a new raster with the same value distribution, but on a different range.



("Bio01" - "Bio01".minimum) / ("Bio01".maximum - "Bio01".minimum)

Output raster: bio01_normalized

Figure 4.7: Normalization procedure and menu of the Principal Components tool.

On the Principal Components menu, the input raster bands are the ones previously calculated. Since all climate variables have the same spatial reference, extent and Coordinate Reference System (CRS), no changes were necessary in the Environments panel, where these parameters, cell size and others can be altered.

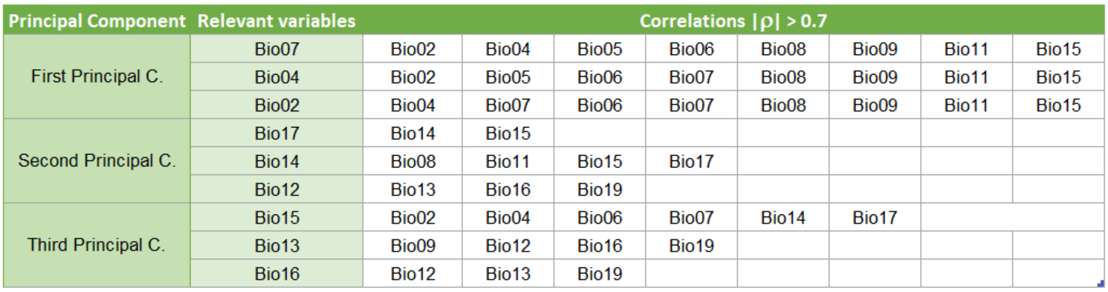
The output for this PCA produced a map with the first three components displayed (Figure 4.9. ArcGIS allows users to produce another output, the “Output data file” option on the menu: a text file containing the covariance matrix, correlation matrix, eigenvalues, and eigenvectors of this technique. Given the nature of these variables, and that some are derived from other ones in the dataset, it’s expected that the whole dataset is very correlated [30]. Extremely correlated variables are generally not useful and are discarded, so we chose to produce the text file to verify the coefficients and proceed with data cleaning.

The full correlation matrix for this PCA is depicted in Table A.1, on Appendix A. In a light red background are the cells with an absolute correlation value, henceforth $|\rho|$, superior to 0.7. Being a common threshold used in literature [30], $|\rho| > 0.7$ was tested as being a rigorous value for biologic variables, while still preserving a sufficiently high amount of information.

In Table A.1, we see that two pairs of features were found to be identical ($|\rho| = 1$): the pair Bio09 and Bio10, and the pair Bio17 and Bio18. Considering the meaning of these variables in Figure 3.5, this is not unexpected. As Portugal is a Mediterranean country, the driest quarter of the year, the Summer, is typically also the driest [31,32], similar to other countries in the same Koppen climate classification [33]. We therefore proceeded to eliminate two of these features, Bio10 and Bio18.

Bio03, which refers to isothermality, is a measure of how temperature oscillates on a daily basis and its relation to an annual variation. We found its purpose to not be clear and easily explainable, and with its very low significance in the first three Principal Components, it was not included in the final variable set.

Having eliminated the aforementioned variables, we analysed the matrix of eigenvectors, annexed on Table A.2 in Appendix A. As explained in the previous subsection, each principal component is an eigenvector of the original matrix, and each cell represents the weight of the respective feature on the PCs. A higher absolute value indicates a higher weight on that component. On the table fig. 4.8, we displayed the three most relevant variables of the first three PCs. Only the first three PCs were considered, as they explain 93% of all variability, achieving a balanced compromise of simplicity and small data loss. The figure represents the correlations of those relevant variables, ordering them per importance.



Principal Component	Relevant variables	Correlations |ρ| > 0.7							
First Principal C.	Bio07	Bio02	Bio04	Bio05	Bio06	Bio08	Bio09	Bio11	Bio15
	Bio04	Bio02	Bio05	Bio06	Bio07	Bio08	Bio09	Bio11	Bio15
	Bio02	Bio04	Bio07	Bio06	Bio07	Bio08	Bio09	Bio11	Bio15
Second Principal C.	Bio17	Bio14	Bio15						
	Bio14	Bio08	Bio11	Bio15	Bio17				
	Bio12	Bio13	Bio16	Bio19					
Third Principal C.	Bio15	Bio02	Bio04	Bio06	Bio07	Bio14	Bio17		
	Bio13	Bio09	Bio12	Bio16	Bio19				
	Bio16	Bio12	Bio13	Bio19					

Figure 4.8: Compiled correlation matrix, ordered by variable importance. Correlations of each variable with itself were not included.

Data was then cleaned in the following way. The first variable, Bio07, is closely related to Bio04 and Bio02, which is expected and will be mentioned in chapter 5. The statistical tool considered them to be the set of variables that describe the most variability of this dataset, therefore they were not eliminated. The next variables, Bio05, Bio06, Bio08, Bio09 and Bio11, being highly correlated to Bio07 and having low weights on all principal components (per A.2, were discarded, as they were redundant and did not contribute to the PCA. After analysing relationships of Bio04 and Bio02, no other variable could be eliminated. In the next set, Bio17 and Bio14 eliminated no variables, however a new feature, Bio19, could be discarded due to its correlation with Bio12. The third principal component had no contributions to the elimination of variables, so the process of data cleaning ended.

This final set of variables, the ones most relevant after cleaning, were used to create the final PCA. This PCA has also created a map output, describing the variables across the surface. The first PC is displayed in red, the second PC in green, and the third PC is displayed in blue, creating all colours in the RGB spectrum.

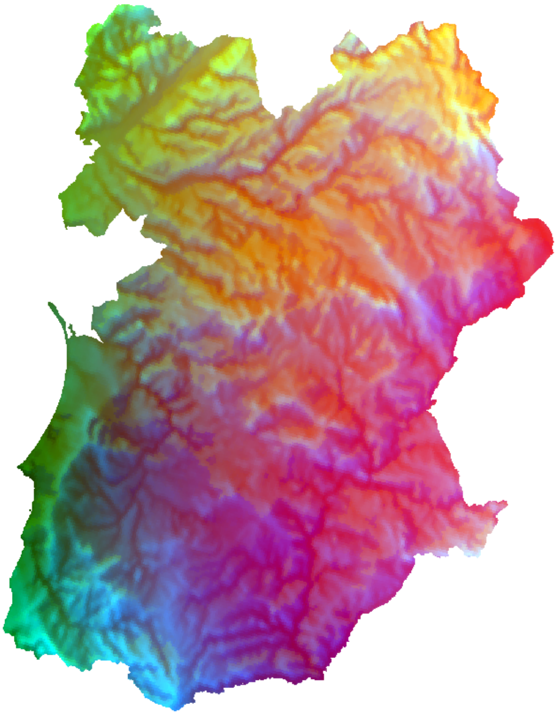


Figure 4.9: Output from Principal Components Analysis of the final ten bioclimatic variables, Alentejo Region. The three first components are pictured here in RGB format.

4.3.2 Regionalization

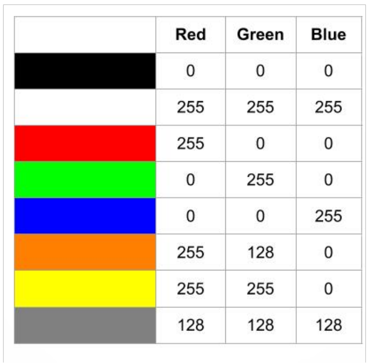
Describing the Alentejo region through principal components already allows some interpretation and insight, but it's necessary to demarcate regions and describe them climatically for future wine growing investment. It was therefore decided to experiment with different methods, for a more accurate result.

4.3.2.A Colour analysis hypothesis

When producing the PCA, it outputs a map with the three RGB components. Each colour exists in a range from 0 to 255, and specific values in all three components create a new colour. Some examples of RGB values and their corresponding colours are below, on figure fig. 4.10.

As per fig. 4.9, some colours can be grouped together through simple observation: the top right corner, referring to Serra de S. Mamede, is predominantly yellow, whereas the coast has a varying green colour, with blue undertones near the southernmost part of Alentejo. Given that these values of colour derive from the values of the first three PCs, it is a plausible hypothesis that similar RGB values present similar climate characteristics.

Histogram analysis is a usual approach. In literature, colour histograms are common for representing and segmenting images, in what is called histogram-based thresholding. For a certain image, at least



	Red	Green	Blue
⬛	0	0	0
⬜	255	255	255
🟥	255	0	0
🟩	0	255	0
🟦	0	0	255
🟧	255	128	0
🟨	255	255	0
▨	128	128	128

Figure 4.10: Some colours and their corresponding RGB values. Image from [34].

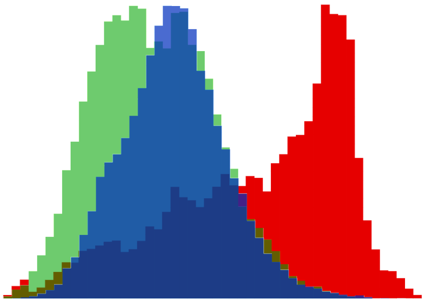


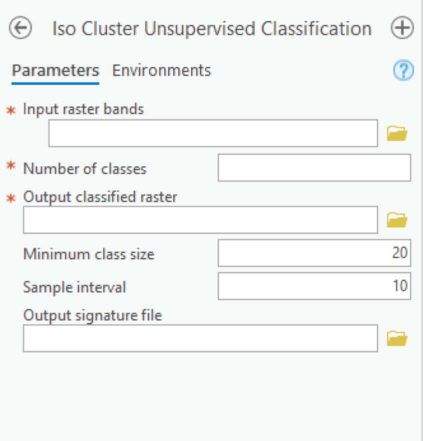
Figure 4.11: Histograms of the first three principal components.

three histograms, one for each RBG component, are built, then combined to calculate which value combinations are most frequent [35, 36]. A preliminary analysis was made to explore this technique, which were found out to be unsatisfactory.

4.3.2.B Clustering

Since the objective is to create regions of this area of Portugal considering climate data, and there are no previously labelled groups a point could belong to, it is appropriate to apply an unsupervised machine learning method. The algorithm applied was ISO Cluster, an algorithm mainly utilized on ArcGIS.

On the tool menu in fig. 4.12, in "Input Raster Bands", the second PCA performed with ten variables was chosen, as a summarized way to describe data variability with no correlation between variables (principal components). The number of classes is the final number of clusters, and the number of seeds the ISO Cluster algorithm will initialize. The minimum class size remained the default, as the number of observations in the dataset (number of raster cells) in relation to the number of asked clusters is sufficient to create statistically valid classes. The sample interval, an optional parameter seen on the tool menu, is the frequency with which the algorithm chooses a cell to perform calculations on each iteration. It was also run with the default value, a very small one in comparison with the dataset, which produced longer and more thorough iterations.



Iso Cluster Unsupervised Classification

Parameters Environments

* Input raster bands

* Number of classes

* Output classified raster

Minimum class size 20

Sample interval 10

Output signature file

Figure 4.12: Parameters to configure the ISO Cluster Unsupervised Classification algorithm.

In the next image, fig. 4.13, we see the outputs of the clustering algorithm, from 5 to 10 clusters. ISO Cluster created, for each image, *k* aleatoric cluster centres, and, based on the values of each point, they were annexed to the nearest one, which evolved until convergence.

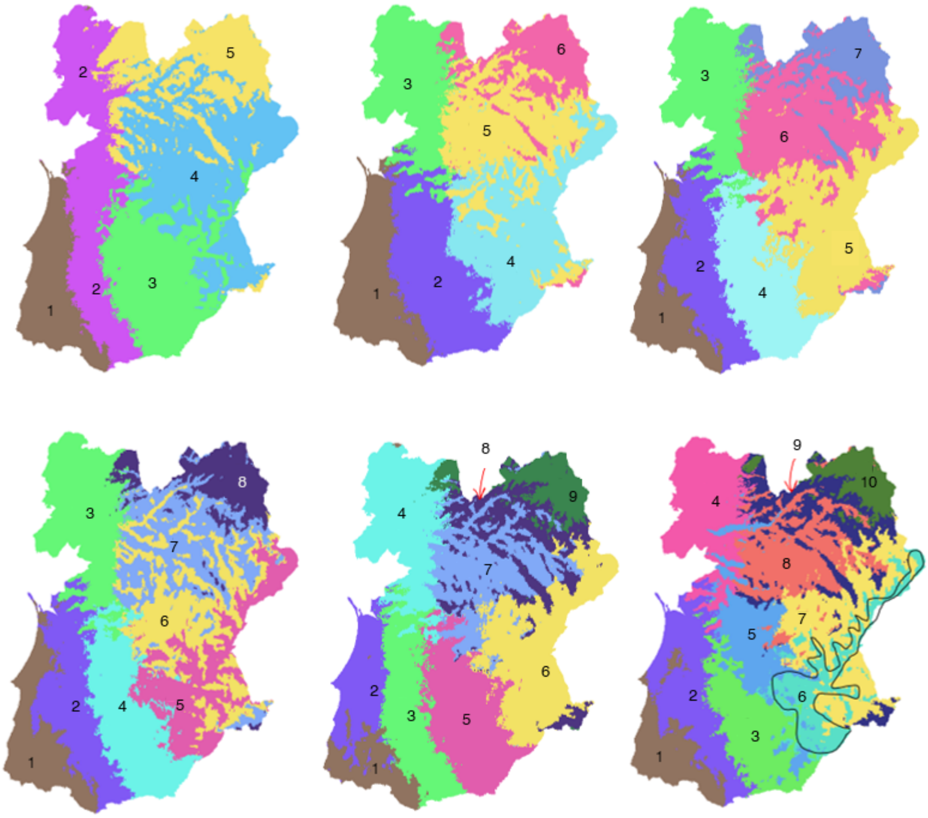


Figure 4.13: Composition of the clustered Alentejo map, from 5 through 10 classes.

A key parameter of a clustering algorithm is the number of clusters to create. The optimal number is largely situational, depending on the context and goal of performing a clustering classification. For a general sense of the major zones, it might suffice to categorize into 5 classes; a more specific setting could benefit from 10 classes or more, depending on the importance of changes in temperature or precipitation. One degree Celsius of difference between two clusters, for example, might be too high and need additional classes in some situations.

Nonetheless, in a general setting, to create too few clusters would be to misclassify the data, and to not provide as much information as the algorithm could. Too many would fall into a loss of significance over each cluster, as they would represent very small subsets of data and, at a maximum degree, would exist one cluster per observation.

After gathering data about the ecological context of Alentejo, we decided to perform the algorithm with 5, 6, 7, 8, 9 and 10 clusters. Less than 5 would provide little information, too broad for the scope of the project, while more than 10 classes was found, by consultation with experts at the NBI company, to be too specific for the area under study.

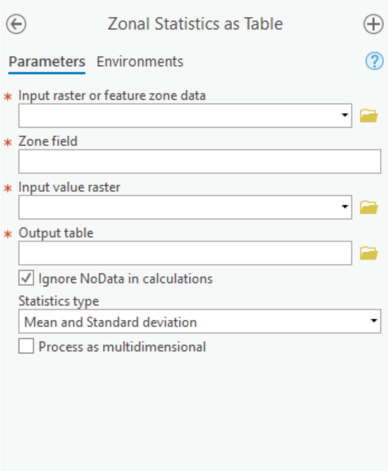
Through the tool that is explained further, Zonal Statistics as Table, we were able to gather data to apply the Student t-test, followed by the Analysis of Variance (ANOVA), statistical tests for cluster significance. Succinctly, both tests calculate the mean and variance of populations, compare them and test the null hypothesis. In this setting, the null hypothesis states that their means are approximately identical, therefore the cluster is not robust enough to be relevant, and may be dissolved. If it's found that this hypothesis may be rejected, then there is statistical significance in the tested clusters. The t-test may only be applied to two groups, while ANOVA applies to two or more populations. The chosen $\alpha = 0.05$ indicates that, for the null hypothesis to be rejected, a 95% or more Confidence Interval for the difference of means must be reached. The *p-value*, the value outputted by the statistical tests to prove the null hypothesis, must also verify *p-value* < 0.05 for it to be rejected [37, 38].

On Appendix A is a table, Table A.4, with mean values of all variables, plus the first and second principal components, for the 10 cluster file. After gathering the number of observations (pixels), means and standard deviation values of the first PC, the introductory t-test was applied. The groups chosen were Cluster 3 and 5, for their consistently similar values throughout all variables, which could indicate a dissolution of these specific clusters. The *p-value* for this test was 0.000, meaning that the tested clusters were significant.

To test all clusters amongst themselves, the ANOVA was applied. Through the same gathered values, it revealed that all 10 clusters are statistically significant, with *p-values* $= 0.000$ for each calculated comparison.

The ANOVA, a widely used method in statistical analysis, is considered a valid assessment of significance [39, 40]. Possibly due to the large number of observations per population, small changes in means or standard deviation values were considered sufficiently relevant to justify the existence of all classes.

To try to capture small, and possibly relevant, microclimates in this area, and not do a too simplistic analysis, the 10 class clustering iteration was the one chosen to proceed with the statistical description, and will be the one presented henceforth.



Zonal Statistics as Table

Parameters Environments

* Input raster or feature zone data

* Zone field

* Input value raster

* Output table

☑ Ignore NoData in calculations

Statistics type

Mean and Standard deviation

☐ Process as multidimensional

Figure 4.14: Zonal Statistics As Table, in ArcGIS.

The final step to analyse climate data across Alentejo is to be able to study bioclimatic variables according to the clusters. Two ArcGIS tools appropriate for this situation are “Zonal Statistics” and “Zonal Statistics As Table”, which have similar parameters. Given a raster that contains the regions with which to split the data, and another input raster with the data to be summarised, both tools will calculate statistics according to each cluster, one through a visual map and the other by creating a table.

Some statistics that the software may calculate are Sum, Mean, Standard Deviation or Range. Using “Zonal Statistics As Table” we chose Mean and Standard Deviation, both to analyse cluster validity, and to describe through the mean how each bioclimatic variable differs across the map.

An example of a Zonal Statistics table and the results of this analysis are presented in Chapter 5.

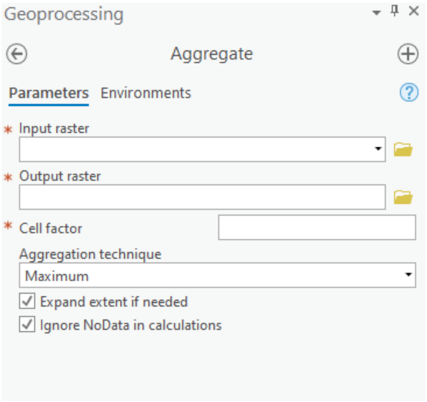
4.4 Land cover data

4.4.1 Initial PCA

COS data presents the location of different habitats on the map, such as water bodies or invasive plant species, and this project’s motivation is to describe the presence of certain habitats with regard to a specific climate region. One way to describe the presence of a habitat across a surface is through a Euclidean distance calculation for the entire map. Each cell indicates the nearest distance, in meters, to an instance of that habitat, which will create a map indicating areas of strong and weak presence of the variable. These distance rasters will then incorporate a PCA, to analyse how they are distributed across the map.

Therefore, we proceeded to calculate the Euclidean distance maps of all variables, for which base rasters are necessary.

Due to an error that will be mentioned in the final chapter, Chapter 6, the Euclidean distances were not being correctly calculated. The solution was found to be a conversion to considerably smaller resolution rasters, to correctly maintain all areas. The consequent upscaling to rasters with a larger cell size was made using the tool Aggregate, which aggregates a specified number of smaller pixels into a new one on the output raster, of a higher cell size. Although not mandatory, this allowed for both datasets, climate and of land cover, to have similar precision and scale.



Maximum
☑ Expand extent if needed
☑ Ignore NoData in calculations

Figure 4.15: Aggregate menu on ArcGIS Pro.

The Aggregation Technique parameter on Figure 4.15 was set to the "Maximum" option. A raster that is created from a vector data file, such as polygons, serves the main purpose of conveying features' location. Such kind of rasters are usually composed of pixels with only a value of 0 or 1, indicating that that specific cell is present, or not, similar to the example on Figure 3.2. "Maximum" allows for the upscaled (larger) pixel to automatically assume the value 1 (therefore being present in the output) if a single pixel that composed it presents the value 1, even if the majority is valued 0. Other options that the software provides are "Majority" or "Minimum" of the small sized pixel values.

This creates a raster with 1000 meter cell size, which decreases precision but improves file size and processing speed. A comparison of all three file types is represented in Figure 4.16 and will be commented further.

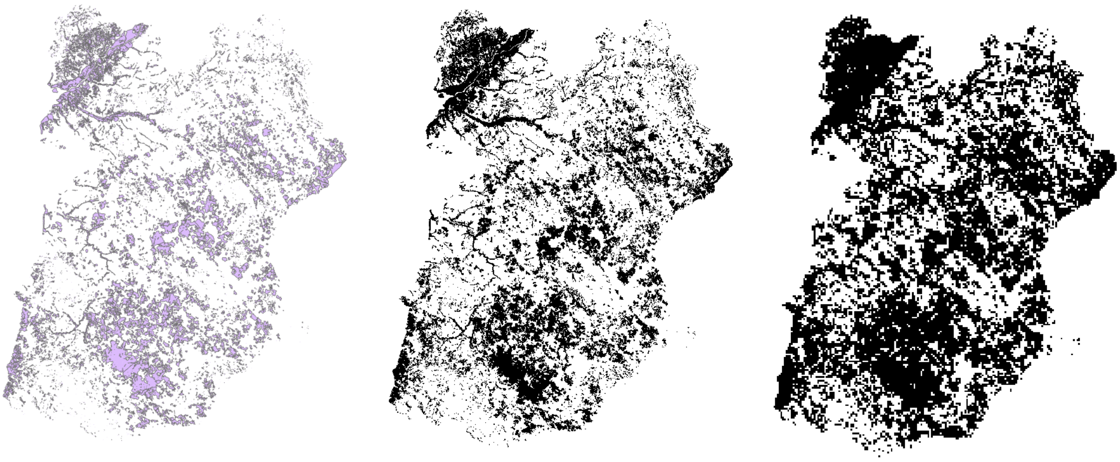


Figure 4.16: An example of a COS variable as (a) a shapefile, (b) a 25 meter cell raster and (c) a 1000 meter cell raster.

With the 1000 meter rasters, we calculated the Euclidean distance map for each variable. It signals 0 when the cell is coincident with an area of the habitat, and increments by 1000 as it gets farther; the value, should be noted, is referent to the nearest area of all those that impact a certain cell.

An example of a distance raster is below, on fig. 4.17, together with the respective shapefile for comparison.

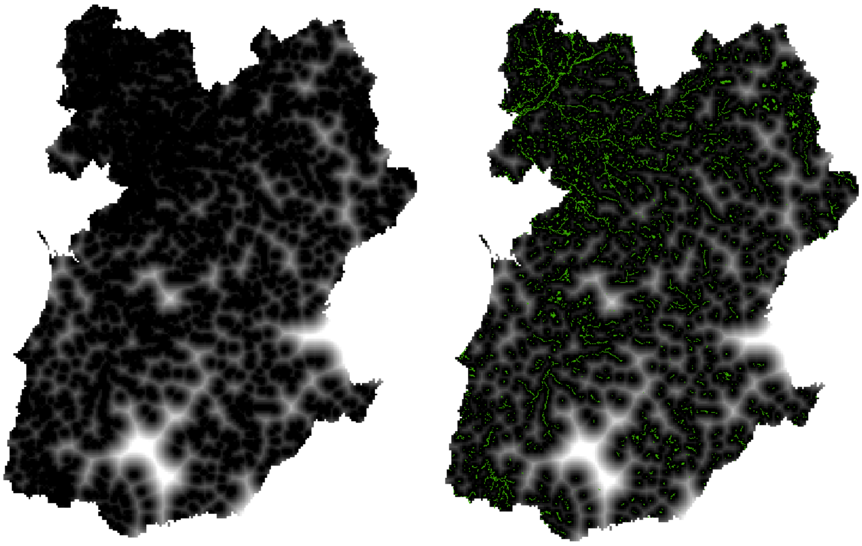


Figure 4.17: An example of a distance raster and respective shapefile.

Before applying a PCA the rasters were normalized, as explained in the previous subsection. After that procedure, the PCA was performed. It served as an exploratory analysis measure, to attempt to form some conclusions on potential areas of interest. It produced a map with the first three PCs, presented further.

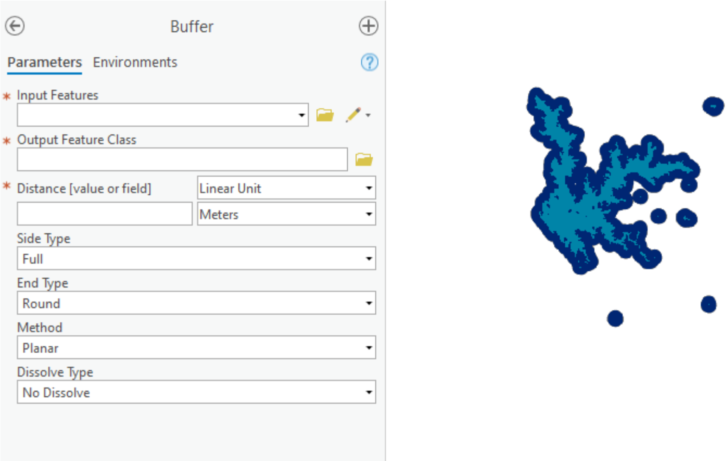
4.4.2 Buffer calculation

Apart from the PCA analysis, this project decided to explore a different approach on intersecting COS and climate data. Having the clustered map, we aim to describe how habitats are scattered throughout Alentejo, therefore we aimed to intersect COS variables with each region.

However, some habitats' impact is not limited exclusively to their area, but have a wider impact. The presence of some variables, such as water bodies, might completely alter their location, from the soil composition, plant health, to the habitats of pollinating insects and larger animals.

To cater for that, we created buffers. A buffer is an area created around all sides of a feature, utilized here to portray the reach of a habitat's influence. The decided area of influence was 300 meters, counting from the edges of the feature. A close zoom on a buffer is depicted on fig. 4.18. The buffer area is in dark blue, whilst the water body that is being buffered is the lighter blue polygon in front.

ArcGIS possesses the Buffer tool, on the Analysis Toolbox, to create these zones. The Input Features are the shapefiles of each habitat (it is necessary to be in vector form, not raster) and the Output Feature



Buffer

Parameters Environments

* Input Features

* Output Feature Class

* Distance [value or field] Linear Unit

Meters

Side Type: Full

End Type: Round

Method: Planar

Dissolve Type: No Dissolve

Figure 4.18: The Buffer tool (a) and an example of a buffer (b).

Class option is the name of the buffer output file. In the Distance field we chose 300 meters, and in the Dissolve Type choice, the "Not dissolve" option. If chosen "Dissolve all", all buffers would be considered one single polygon. This would greatly hinder future calculations, as there wouldn't be information about the size of each feature, and consequently wouldn't be possible to count how many instances of a habitat exist in a specific cluster, or the area that those instances occupy.

Because cluster areas are important for future statistics, if there are habitat polygons near each other with overlapping buffers, the buffer area in total would be miscounted. Overlapping areas of influence should count as a single one, but their areas would be summed, largely overstating the real value (see fig. 4.19 for further clarification). The Dissolve Boundaries, a more specific tool to handle boundaries of polygons, allowed for a join of overlapping buffers, flattening them and unifying their area.

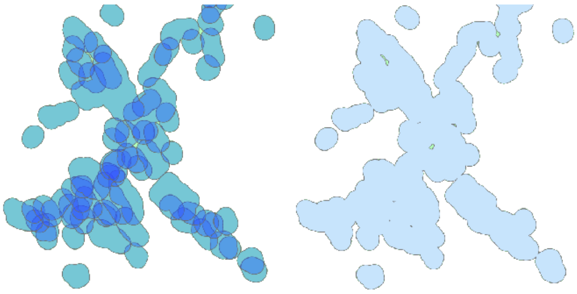


Figure 4.19: An example of (a) boundaries not dissolved and (b) dissolved.

For reasons stated earlier, these buffers were applied on specific variables: water bodies, native forest, riparian areas and scrubland.

An important detail of the riparian features is their possibly negligible area, because of their nature as usually thin streams of water, being length a more significant measure. On final calculations of area percentage on each cluster, riparian features would have a minimal presence, and their substantial relevance could be undervalued. To consider this issue, ArcGIS provides the Polygon to Centerline tool, particularly designed for hydrographic polygons. This tool creates lines to represent bodies of water, specifically through their centre; the tool Polygon to Line, that might be considered valid in this situation, creates lines around the polygons, which would be an error.

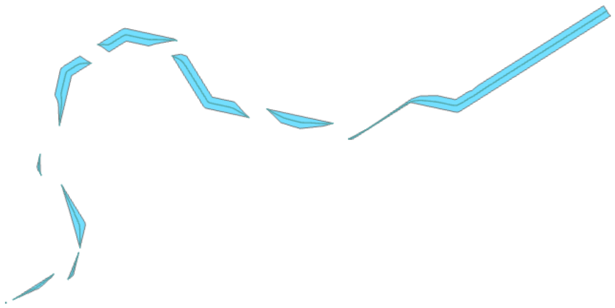


Figure 4.20: An example of a riparian area polygon, with a centerline.

Creating centerlines allowed for a study of riparian zones' length, in meters, in final descriptive statistics, besides the area provided by polygons.

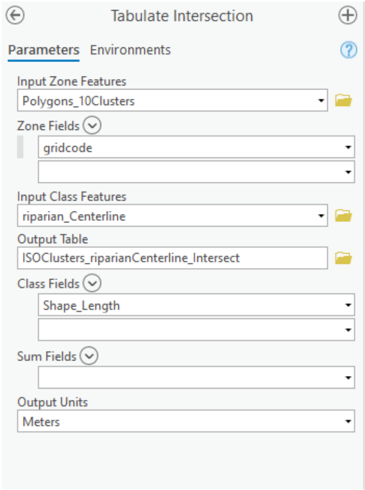
At this phase, we had shapefiles of buffers with dissolved boundaries of these four variables, and could proceed to the intersection between them and the climate clusters.

4.5 Criteria compilation

To use the necessary tool to calculate statistics, the clustering file had to be converted to polygons. Due to the nature of the clusters, that contain a significant amount of small areas spread throughout the map, the polygon conversion outputted the same small areas. Each cluster was then defined by multiple small polygons and a usual larger polygon, all of which had to be considered the same class. In the Edit tab of the main screen, all polygons with the same *gridcode* (a unique number assigned to each cluster during the ISO Clustering process) were merged, creating the final shapefile needed.

The “Tabulate Intersection” was the method used for the final statistics. This tool receives a shapefile that defines the zones, which refers to the previously calculated clusters as polygons; another feature file with the variable to calculate the desired statistics of, and it outputs a table with their data of intersection. This tool differs from the one utilized by the bioclimatic dataset: instead of means, sums or standard deviations, “Tabulate Intersection” focuses on the area occupied by the variable on each cluster, and calculates the percentage relative to the whole cluster area.

The outputs of this tool are the final statistical descriptions of this project, which will be analysed in



Tabulate Intersection

Input Zone Features: Polygons_10Clusters

Zone Fields: gridcode

Input Class Features: riparian_Centerline

Output Table: ISOClusters_riparianCenterline_Intersect

Class Fields: Shape_Length

Sum Fields:

Output Units: Meters

Figure 4.21: The Tabulate Intersection tool.

the next section.

5

Results and Discussion

Contents

5.1 Results and Discussion

5.1.1 Climate data

With the PCA we have gained insight on what variables explain the most variability of the Alentejo climate dataset, so it is useful to picture some, as well as others of general interest, for future interpretation. In this page we depicted Bio01, the annual mean temperature (meanings of bioclimatic variables in fig. 3.5), and Bio07, deemed the variable with the highest weight on the first PC, which regards to the annual temperature range – the lowest temperature of the year subtracted from the highest temperature, for the 1979-2013 period.

In the first image, we see that the central and south areas of Alentejo present a higher annual mean temperature than the northeast. Topographical features are very well detailed here, with the Guadiana river and other small rivers presenting, interestingly, the highest annual mean temperatures. Bio07 is displaying how different the highest temperature in the Summer is related to the lowest temperature in Winter. Near the coast the range is low, and it increases to the east, presenting higher variability.

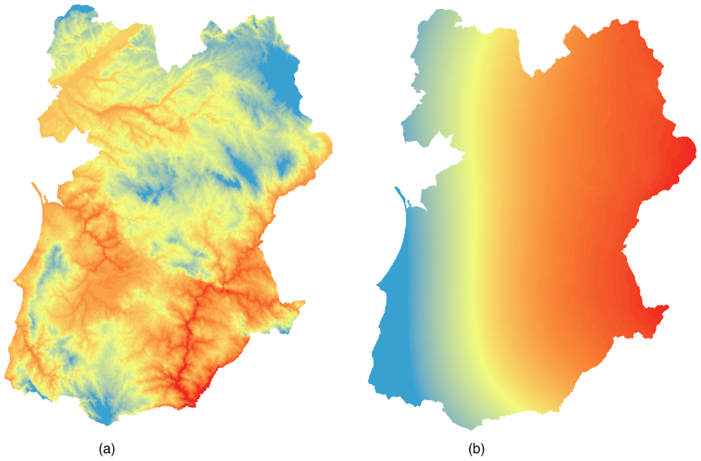


Figure 5.1: Bio01 (a) and Bio07 (b) across the Alentejo map. Red colour indicates higher values.

From Bio12 we see that North Alentejo, as well as the southernmost area, has the highest precipitation rates, while the interior, as expected, presents less cumulative precipitation. The centre and south areas of the map present higher degrees of variability, meaning that monthly precipitation values can vary greatly, while the north, coincident with Portalegre district, is more stable throughout the year. Bio17 highlights how the south of this region can reach extremely low precipitation levels during the driest quarter, where drought is a common situation. The northernmost region showcased, through these variables, a more stable flux of precipitation throughout the year, one of the highest yearly precipitation

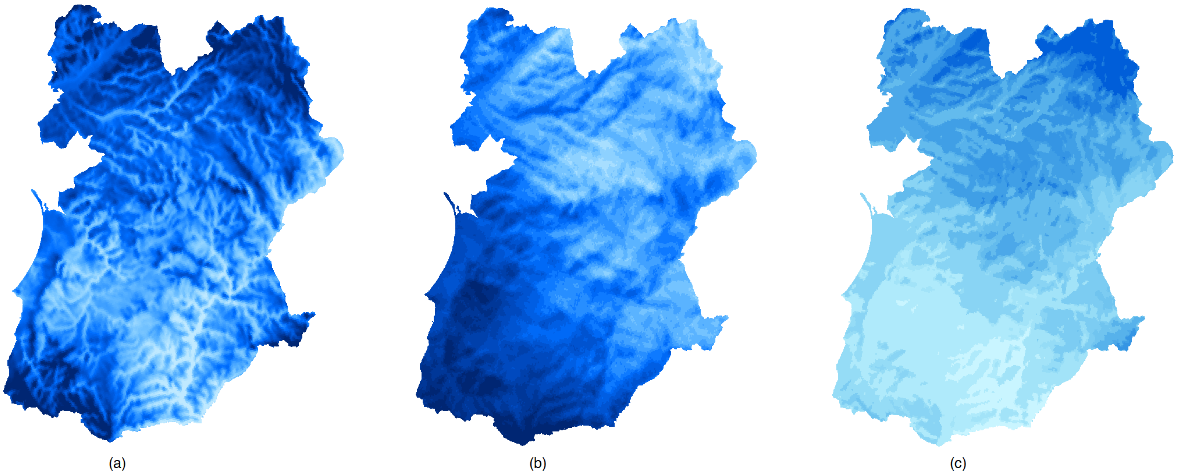


Figure 5.2: Bio12 (a), Bio15 (b) and Bio17 (c) across the Alentejo map. Darker colours indicates higher values.

levels, and the most humid dry season. We can then hypothesize that this area is more resilient to drought and favourable to species that may need higher water availability.

The map result of the PCA, previously shown on Figure 4.9, may be described by three individual bands (principal components), the first two relevant to present on their own. It is also pertinent to give special emphasis to the second component, when layered with the water habitat of COS data.

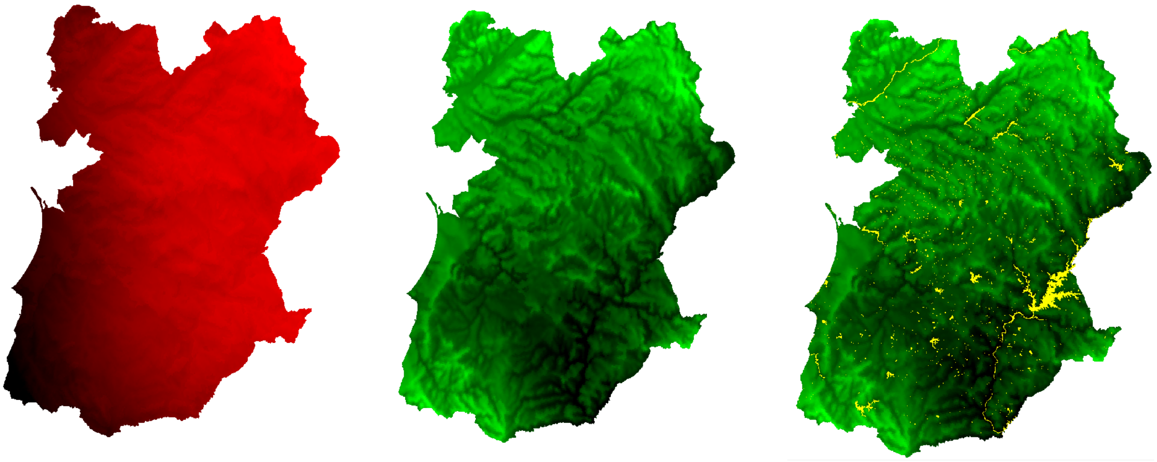
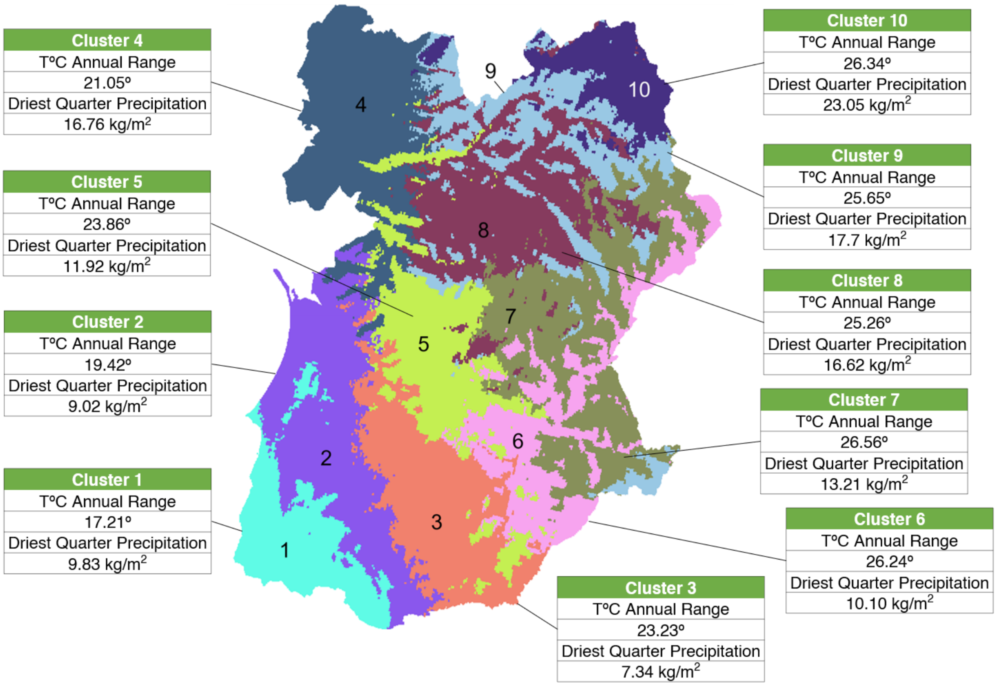


Figure 5.3: First two bands of PCA on climate data.

The first principal component, displayed in red, is a measure of temperature variations throughout different periods of time, and inclues Bio07 and of the most relevant. This map of temperature range is exhibiting the region's continentality. Continentality is an observed climate condition that states that temperatures have smaller variations near a large water body, such as an ocean, and larger variations as one gets farther from the coast [41]. It is a known property and prevalent in countries with a large coastal area, such as Portugal. Near the ocean, temperatures are consistently temperate and with a low range; near the opposite border, shared with Spain, temperatures can vary greatly between the day and the night, and between summer and winter.

According to the eigenvectors table, we can understand that the second principal component is an axis of combined precipitation variables. With no DEM (digital elevation model) data applied to the technique, PCA recognized topographical features, like changes in elevation, through precipitation. The water habitat layer overlaid on this map shows that the area with the lowest values, in darker colour, overlaps with rivers and riparian zones. This means they have a specific precipitation pattern, identifiable through this method.

Considering the clustering algorithm applied to the PCA variables, some statistics were calculated for each class, to better understand, on a cluster basis, temperature and precipitation values. These values were gathered using the two most descriptive variables, Bio07 and Bio17, and the average minimum and maximum temperatures registered, Bio05 and Bio06.



Cluster 4
TºC Annual Range
21.05º
Driest Quarter Precipitation
16.76 kg/m^2

Cluster 5
TºC Annual Range
23.86º
Driest Quarter Precipitation
11.92 kg/m^2

Cluster 2
TºC Annual Range
19.42º
Driest Quarter Precipitation
9.02 kg/m^2

Cluster 1
TºC Annual Range
17.21º
Driest Quarter Precipitation
9.83 kg/m^2

Cluster 10
TºC Annual Range
26.34º
Driest Quarter Precipitation
23.05 kg/m^2

Cluster 9
TºC Annual Range
25.65º
Driest Quarter Precipitation
17.7 kg/m^2

Cluster 8
TºC Annual Range
25.26º
Driest Quarter Precipitation
16.62 kg/m^2

Cluster 7
TºC Annual Range
26.56º
Driest Quarter Precipitation
13.21 kg/m^2

Cluster 6
TºC Annual Range
26.24º
Driest Quarter Precipitation
10.10 kg/m^2

Cluster 3
TºC Annual Range
23.23º
Driest Quarter Precipitation
7.34 kg/m^2

Figure 5.4: Clustered map of Alentejo with statistics for Bio07 and Bio17.

The continentality feature is observed comparing data of Cluster 1 and 10: the annual temperature range in Cluster 1 is 17.21 ℃, while in Cluster 10 it is 26.34 ℃. In a future implementation of a grape variety that needs constant temperate weather, coastal clusters present themselves as a better choice, temperature-wise.

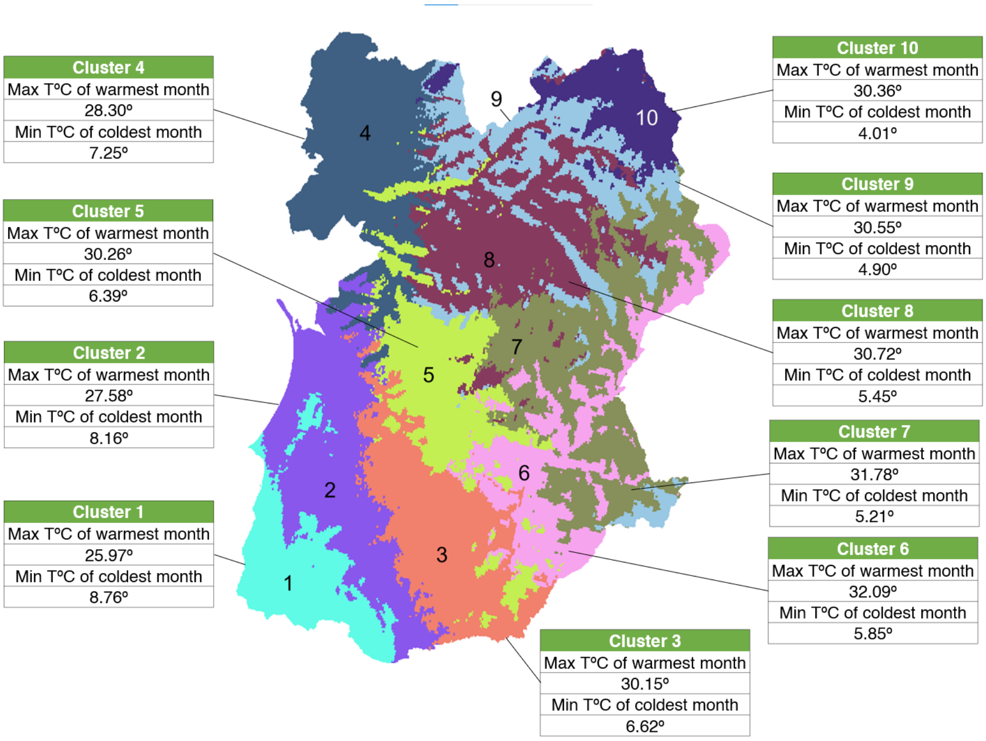


Figure 5.5: Clustered map of Alentejo with statistics for Bio05 and Bio06.

The temperature values for Clusters 1, 2 and 4 are similar, with Cluster 1 exhibiting the least temperature range. It should be noted that, from the three clusters along the coast, Cluster 1 is the smallest: both Cluster 2 and 4 cover area that might be considered as belonging to the centre of Alentejo, less affected by continentality.

On the eastern border, Cluster 6, 7, 8, 9 and 10 present very similar levels of temperature range, with only one degree of difference between them. That is clearly consistent with their annual minimum and maximum temperature values. Cultivars appropriate for hot climates, which tend to produce sweeter and stronger wine, would be a suitable choice for new wine growing investments; if paired, however, with bodies of water or riparian areas, as rainfall values only seem consistent on Cluster 10.

Cluster 3 is the area with less precipitation amount on the driest quarter, and even in cumulative annual precipitation.

5.1.2 Land cover data

After converting all habitat variables into new raster files and creating Euclidean distance maps, they were individually normalised to produce a PCA, whose output is displayed on fig. 5.6.

We understood with this output that PCA created artificial lines, which don't align with any previously recorded topographic feature or specific habitat placement. They derive from an artificialization of the

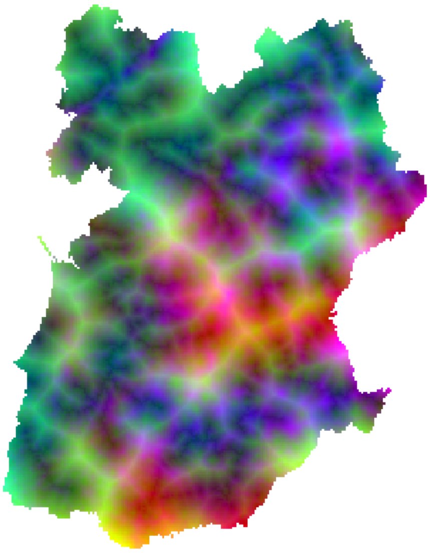


Figure 5.6: Principal Component Analysis on COS data.

Euclidean distance maps, with no geographical value, therefore the result was suboptimal. We have also detected a lack of patterns in land cover. While some habitats are semi-natural, their placement was most likely planned, driven by external decisions such as financial, social or political. We can conclude with this PCA application that a numerical manipulation and transformation of COS data doesn't allow for meaningful conclusions, and that technique descriptive, statistical approach is more useful in Alentejo.

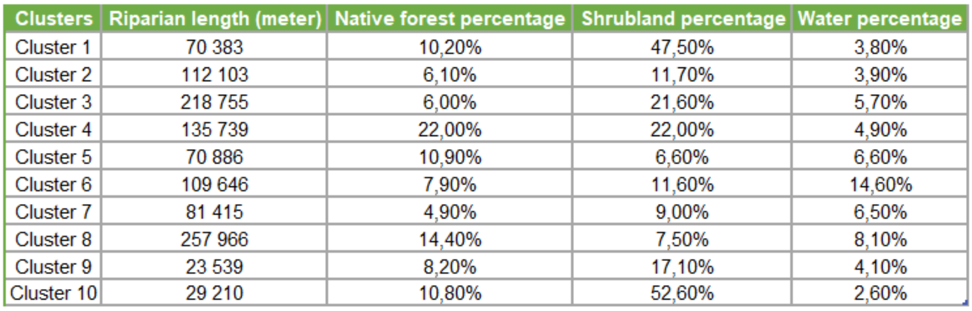
5.1.3 Integrated analysis

More than a climate description of Alentejo, in part already addressed in literature with respect to climate change challenges [3, 42], this project's motivation is to encompass that climatic description with land cover data, specifically with variables that are here hypothesized as semi-natural and positive (recall Chapter 2).

This semi-natural variable distribution across the map is displayed on Appendix B with accompanying maps; a summarized table version is below.

Cluster 3, and, to a lesser degree, 2 and 1, have considerably low precipitation values during the driest quarter, and precipitation amount is very inconsistent throughout the year. The three clusters have very substantial riparian length, but further research is needed to test if riparian areas might suffice for an extensive vineyard's water needs.

Cluster 1, which encompasses most of the Odemira municipality of Beja, is the second highest valued cluster in shrubland area, with 47.5%. From the article by Paredes et all [5], it might be beneficial to employ that shrubland area to provide more complex and diverse biodiversity for vineyards; however, it is the second less fertile in water availability, and, as seen earlier, also low on precipitation values.



Clusters	Riparian length (meter)	Native forest percentage	Shrubland percentage	Water percentage
Cluster 1	70 383	10,20%	47,50%	3,80%
Cluster 2	112 103	6,10%	11,70%	3,90%
Cluster 3	218 755	6,00%	21,60%	5,70%
Cluster 4	135 739	22,00%	22,00%	4,90%
Cluster 5	70 886	10,90%	6,60%	6,60%
Cluster 6	109 646	7,90%	11,60%	14,60%
Cluster 7	81 415	4,90%	9,00%	6,50%
Cluster 8	257 966	14,40%	7,50%	8,10%
Cluster 9	23 539	8,20%	17,10%	4,10%
Cluster 10	29 210	10,80%	52,60%	2,60%

Figure 5.7: Area percentage, per cluster, of each semi-natural variable.

Therefore we may say that this specific area is not for cultivars with high water needs, which is one of the most important factors for vine development [43].

The Cluster 10, which represents the Serra de S. Mamede Natural Park, is the cluster highest in shrubland area. Its water body percentage is low, and is the cluster with the second lowest value of riparian length. Despite these values, it is one of the clusters with higher precipitation values throughout the year, and its annual temperature range provides a comfortable interval for vineyards to have a proper dormancy period, where cold is a relevant factor. We may suggest this cluster, for all reasons described above, a potentially good candidate for new vineyards, proving the cultivars chosen are adequate for the high Summer temperatures.

Cluster 4 stands out as a seemingly good candidate for wine growing. It has a considerable riparian length and medium water availability; considerable precipitation amounts, which don't reach critical levels during the warmest quarter; and has a medium precipitation variability. It is also fertile in native forests and shrubland, which are being hypothesized as being beneficial for vineyards. By land cover data, this area is highly urban. Further analysis of this area's land planning might prove beneficial, to utilize this area to its full potential.

6

Conclusion and Future Work

Contents

In this chapter we will present this thesis' conclusions and suggested future work.

6.1 System Limitations

There is some lacking data that would have been beneficial to analyse, such as aquifers. When analysing merely water bodies at the surface, we may be wrongly estimating water availability, therefore it is suggested an aquifer analysis to complement this study.

In COS, the classification "Olive grove" was assumed to be referring to intensive olive groves, and therefore prejudicial for a vineyard, due to specific management practices that worsen soil health and surrounding biodiversity. That classification was not explicit regarding that management, and some traditional olive groves might have been misclassified, and could be a positive presence on Alentejo's vineyards. A clarification regarding that specific class should be sought with DGT.

6.2 Conclusions and Future Work

Climate suitability for wine growing is highly complex, and dependant on multiple factors. Per Gregory Jones, renowned viticulture expert and climatologist, the minimum temperature interval for proper wine growing is [-1, 18.9] for the growing period [44], which is, in the Northern Hemisphere, the Spring and Summer seasons. From temperature values alone, a producer would theoretically be able to produce wine on all clusters defined in this project. Water availability is, however, a very important factor for vines, and also, proven by Paredes et al, their surroundings.

Cluster 4 and Cluster 10 are, from the data collected and analysed, good candidates for future wine growing, due to their medium to high rainfall values, appropriate temperatures (although for different cultivars), and very high percentage of semi-natural variables.

We have described the presence of four semi-natural variables in Alentejo, and, within that description, elaborate on the area's climate. NBI, partnering with CVRA, may deepen this work by obtaining productivity data from wine producers. Having a stable climate, productivity and semi-natural habitat presence may be correlated, to further discover relationships between them. A positive correlation would not only aid in controlling pests and lower insecticide applications, but it might incentivize general habitat conservation, which might bring benefits to Alentejo's climate, but also flora, fauna and population.

Bibliography

[1] Eurostat, "Wine-grower holdings by production (dataset)," 2015. [Online]. Available: https://ec.europa.eu/eurostat/databrowser/view/vit_t1/

[2] N.Português, "Artwork: NUTS 2 and 3 of portugal in 2013." [Online]. Available: https://commons.wikimedia.org/wiki/File:NUTS_PT_2013.png

[3] Helder Fraga and Joao Santos, "Climate change projections for the Portuguese viticulture using a multi-model ensemble," *Ciência Téc. Vitiv*, vol. 27, no. 1, pp. 39–48, 2012.

[4] D. Santillán, V. Sotés, A. Iglesias, and L. Garrote, "Adapting viticulture to climate change in the mediterranean region: Evaluations accounting for spatial differences in the producers-climate interactions," vol. 12, 2019.

[5] D. Paredes, J. A. Rosenheim, R. Chaplin-Kramer, S. Winter, and D. S. Karp, "Landscape simplification increases vineyard pest outbreaks and insecticide use," *Ecology Letters*, vol. 24, pp. 73–83, 1 2021.

[6] A. Dalla Marta, D. Grifoni, M. Mancini, P. Storchi, G. Zipoli, and S. Orlandini, "Analysis of the relationships between climate variability and grapevine phenology in the nobile di montepulciano wine production area," vol. 148, pp. 657–666.

[7] H. Fraga, A. C. Malheiro, J. Moutinho-Pereira, R. M. Cardoso, P. M. M. Soares, J. J. Cancela, J. G. Pinto, and J. A. Santos, "Integrated Analysis of Climate, Soil, Topography and Vegetative Growth in Iberian Viticultural Regions," *PLoS ONE*, vol. 9, no. 9, p. e108078, Sep. 2014.

[8] G. Jones and G. Goodrich, "Influence of climate variability on wine regions in the western USA and on wine quality in the napa valley," vol. 35, pp. 241–254.

[9] G. Jones, M. White, and O. Cooper, "Climate change and global wine quality," vol. 85, pp. 504–+.

[10] J. Camps and M. Ramos, "Grape harvest and yield responses to inter-annual changes in temperature and precipitation in an area of north-east spain with a mediterranean climate," vol. 56, pp. 853–64.

[11] D. Paredes, D. S. Karp, R. Chaplin-Kramer, E. Benítez, and M. Campos, “Natural habitat increases natural pest control in olive groves: economic implications,” publisher: Springer Verlag.

[12] F. Gould, Z. S. Brown, and J. Kuzma, “Wicked evolution: Can we address the sociobiological dilemma of pesticide resistance?” vol. 360, no. 6390, pp. 728–732.

[13] C. Dempsey, “Types of GIS data explored: Vector and raster),” 05 2021. [Online]. Available: https://www.gislounge.com/geodatabases-explored-vector-and-raster-data/

[14] “A guide to principal component analysis (PCA) for machine learning.” [Online]. Available: https://www.keboola.com/blog/pca-machine-learning

[15] “Registo nacional de dados geográficos - direção-geral do território.” [Online]. Available: https://snig.dgterritorio.gov.pt/rndg/srv/por/catalog.search#/home

[16] “Land use and coverage area frame survey (LUCAS) —| land & water | food and agriculture organization of the united nations | land & water | food and agriculture organization of the united nations.” [Online]. Available: https://www.fao.org/land-water/land/land-governance/land-resources-planning-toolbox/category/details/en/c/1236417/

[17] EEA), European Environment Agency, “CLC 2018 — copernicus land monitoring service.” [Online]. Available: https://land.copernicus.eu/pan-european/corine-land-cover/clc2018

[18] DGT, “Especificações técnicas da carta de uso e ocupação do solo (COS) de portugal continental para 2018 - DGT.” [Online]. Available: https://www.dgterritorio.gov.pt/node/1045?language=en

[19] J.-F. Pekel, A. Cottam, N. Gorelick, and A. S. Belward, “High-resolution mapping of global surface water and its long-term changes,” vol. 540, no. 7633, pp. 418–422. [Online]. Available: http://www.nature.com/articles/nature20584

[20] European Commission’s Joint Research Centre, “Global surface water – data access.” [Online]. Available: https://global-surface-water.appspot.com/download

[21] D. N. Karger, O. Conrad, J. Böhner, T. Kawohl, H. Kreft, R. W. Soria-Auza, N. E. Zimmermann, H. P. Linder, and M. Kessler, “Data from: Climatologies at high resolution for the earth’s land surface areas,” 2018. [Online]. Available: https://doi.org/10.5061/dryad.kd1d4

[22] D. N. Karger and N. E. Zimmermann, “CHELSA climate downloads—technical description.” [Online]. Available: https://chelsa-climate.org/downloads/

[23] D. P. Dee, S. M. Uppala, A. J. Simmons, P. Berrisford, P. Poli, S. Kobayashi, U. Andrae, M. A. Balmaseda, G. Balsamo, P. Bauer, P. Bechtold, A. C. M. Beljaars, L. van de Berg,

J. Bidlot, N. Bormann, C. Delsol, R. Dragani, M. Fuentes, A. J. Geer, L. Haimberger, S. B. Healy, H. Hersbach, E. V. Hólm, L. Isaksen, P. Kållberg, M. Köhler, M. Matricardi, A. P. McNally, B. M. Monge-Sanz, J.-J. Morcrette, B.-K. Park, C. Peubey, P. de Rosnay, C. Tavolato, J.-N. Thépaut, and F. Vitart, "The ERA-interim reanalysis: configuration and performance of the data assimilation system," vol. 137, no. 656, pp. 553–597, _eprint: https://misclibrary.wiley.com/doi/pdf/10.1002/qj.828. [Online]. Available: https://misclibrary.wiley.com/doi/abs/10.1002/qj.828

[24] Sriram.aeropsn, "English: This diagram explains how the formula for calculating the angular diameter is formulated." [Online]. Available: https://commons.wikimedia.org/wiki/File:Angular_dia_formula.JPG

[25] J. P. Wilson and J. C. Gallant, Eds., *Terrain analysis: principles and applications*. Wiley.

[26] "Bioclimatic variables — WorldClim 1 documentation." [Online]. Available: https://www.worldclim.org/data/bioclim.html

[27] I. T. Jolliffe, *Principal Component Analysis*. Springer New York, OCLC: 853262127. [Online]. Available: https://doi.org/10.1007/978-1-4757-1904-8

[28] "Partitional clustering in r: The essentials." [Online]. Available: https://www.datanovia.com/en/courses/partitional-clustering-in-r-the-essentials/

[29] B. Qian, J. Su, Z. Wen, R. Yang, A. Zomaya, and O. Rana, *Orchestrating the Development Lifecycle of Machine Learning-Based IoT Applications: A Taxonomy and Survey*.

[30] C. F. Dormann, J. Elith, S. Bacher, C. Buchmann, G. Carl, G. Carré, J. R. G. Marquéz, B. Gruber, B. Lafourcade, P. J. Leitão, T. Münkemüller, C. McClean, P. E. Osborne, B. Reineking, B. Schröder, A. K. Skidmore, D. Zurell, and S. Lautenbach, "Collinearity: a review of methods to deal with it and a simulation study evaluating their performance," vol. 36, no. 1, pp. 27–46.

[31] C. Mora and G. Vieira, "The climate of portugal," in *Landscapes and Landforms of Portugal*, G. Vieira, J. L. Zêzere, and C. Mora, Eds. Springer International Publishing, pp. 33–46, series Title: World Geomorphological Landscapes. [Online]. Available: http://link.springer.com/10.1007/978-3-319-03641-0_2

[32] P. Miranda, M. F. Coelho, A. Tomé, M. Valente, A. Carvalho, C. Pires, H. Pires, and V. Pires, "20th century portuguese climate and climate scenarios in climate change in portugal: Scenarios, impacts and adaptation," pp. 27–83.

[33] IPMA, "IPMA - clima normais em portugal." [Online]. Available: https://www.ipma.pt/en/oclima/normais.clima/?print=true

[34] "Image representation — data quality explored." [Online]. Available: https://www3.tuhh.de/sts/hoou/data-quality-explored/2-1-1-image-representation.html

[35] M. Li, L. Wang, S. Deng, and C. Zhou, "Color image segmentation using adaptive hierarchical-histogram thresholding," vol. 15, no. 1, p. e0226345, 2020, publisher: Public Library of Science. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0226345

[36] G. Ramella and G. Sanniti di Baja, *Color Histogram-Based Image Segmentation*, pages: 83.

[37] A. F. Siegel and M. R. Wagner, "Chapter 15 - ANOVA: Testing for differences among many samples and much more," in *Practical Business Statistics (Eighth Edition)*, A. F. Siegel and M. R. Wagner, Eds. Academic Press, pp. 485–510. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780128200254000154

[38] "Analysis of variance (ANOVA)," in *Rank-Based Methods for Shrinkage and Selection*. John Wiley & Sons, Ltd, pp. 149–190, section: 4 _eprint: https://misclibrary.wiley.com/doi/pdf/10.1002/9781119625438.ch4. [Online]. Available: https://misclibrary.wiley.com/doi/abs/10.1002/9781119625438.ch4

[39] M. Peña, M. Cerrada, X. Alvarez, D. Jadán, P. Lucero, B. Milton, R. Guamán, and R.-V. Sánchez, "Feature engineering based on ANOVA, cluster validity assessment and KNN for fault diagnosis in bearings," vol. 34, no. 6, pp. 3451–3462. [Online]. Available: https://www.medra.org/servlet/aliasResolver?alias=iospress&doi=10.3233/JIFS-169525

[40] T. Calinski and L. C. A. Corsten, "Clustering means in ANOVA by simultaneous testing," vol. 41, no. 1, p. 39. [Online]. Available: https://www.jstor.org/stable/2530641?origin=crossref

[41] N. P. Molchanova, A. V. Letuchy, S. V. Morozova, K. S. Kondakov, and N. A. Shcherbakova, "The influence of the degree of climate continentality on the productivity of agricultural production," vol. 1010, no. 1, p. 012156, publisher: IOP Publishing. [Online]. Available: https://iopscience.iop.org/article/10.1088/1755-1315/1010/1/012156/meta

[42] H. Fraga, J. A. Santos, A. C. Malheiro, A. A. Oliveira, J. Moutinho-Pereira, and G. V. Jones, "Climatic suitability of Portuguese grapevine varieties and climate change adaptation," *International Journal of Climatology*, vol. 36, no. 1, pp. 1–12, Jan. 2016.

[43] H. Ojeda, A. Deloire, and A. CARBONNEAU, "Influence of water deficits on grape berry growth," vol. 40, pp. 141–145.

[44] G. Jones, "Climate, grapes, and wine." [Online]. Available: https://www.guildsomm.com/public_content/features/articles/b/gregory_jones/posts/climate-grapes-and-wine

A

Tables

As mentioned in chapter 4, the first table in this Appendix shows the correlation between the nineteen bioclimatic variables, which the PCA calculates automatically. Except for the correlation of each variable with itself, all cells of the matrix whose absolute value is superior to 0.7 are marked in red. The second table below shows the eigenvectors calculated by the same PCA. Eigenvectors produced by the technique are the principal components, and each value is the weight of the corresponding variable on that PC. On the first three PCs, the three variables with the higher weight value are marked in green. Lastly, on Figure A.1 we see how much data variability each Principal Component explains.

Table A.3 presents the eigenvalues of the second, and final, PCA, with the first three variables of the first three PCs also highlighted.

Table A.4 contains, for each of the 10 clusters produced with ISO Cluster, the number of observations per cluster and the mean values of the final ten bioclimatic variables. The mean values of the first and second PC are also listed.

Variable	Bio01	Bio02	Bio03	Bio04	Bio05	Bio06	Bio07	Bio08	Bio09	Bio10	Bio11	Bio12	Bio13	Bio14	Bio15	Bio16	Bio17	Bio18	Bio19
Bio01	1	-0,106550	0,067760	-0,167950	0,16264	0,513840	-0,139300	0,662430	0,395770	0,39577	0,664030	-0,654220	-0,563330	-0,544320	0,2103	-0,594340	-0,591500	-0,591500	-0,611180
Bio02	-0,106550	1	0,634410	0,959530	0,94626	-0,898410	0,986160	-0,750890	0,831420	0,83142	-0,786940	-0,220420	-0,477360	0,477200	-0,74053	-0,444980	0,327440	0,327440	-0,370360
Bio03	0,067760	0,634410	1	0,398520	0,50238	-0,431610	0,502630	-0,252160	0,399250	0,39925	-0,280160	-0,233250	-0,328330	0,213900	-0,41662	-0,341500	0,159600	0,159600	-0,264930
Bio04	-0,167950	0,959530	0,398520	1	0,93956	-0,921320	0,992680	-0,815330	0,837550	0,83755	-0,846390	-0,165520	-0,437940	0,506970	-0,73243	-0,393200	0,349580	0,349580	-0,327040
Bio05	0,162640	0,946260	0,502380	0,939560	1	-0,757480	0,953320	-0,587690	0,964120	0,96412	-0,622910	-0,403140	-0,634300	0,309960	-0,66312	-0,606700	0,138240	0,138240	-0,547360
Bio06	0,513840	-0,898410	-0,431610	-0,921320	-0,75748	1	-0,918990	0,948750	-0,571320	-0,57132	0,976070	-0,104050	0,170280	-0,661360	0,7342	0,121540	-0,547880	-0,547880	0,048670
Bio07	-0,139300	0,986160	0,502630	0,992680	0,95332	-0,918990	1	-0,793030	0,845780	0,84578	-0,826900	-0,195250	-0,461500	0,492680	-0,73942	-0,422330	0,336580	0,336580	-0,352850
Bio08	0,662430	-0,750890	-0,252160	-0,815330	-0,58769	0,948750	-0,793030	1	-0,390290	-0,39029	0,976160	-0,281300	0,002860	-0,733240	0,69329	-0,057280	-0,651450	-0,651450	-0,133860
Bio09	0,395770	0,831420	0,399250	0,837550	0,96412	-0,571320	0,845780	-0,390290	1	1,00000	-0,418870	-0,522210	-0,717410	0,155540	-0,5513	-0,695280	-0,018850	-0,018850	-0,646110
Bio10	0,395770	0,831420	0,399250	0,837550	0,96412	-0,571320	0,845780	-0,390290	1,000000	1	-0,418870	-0,522210	-0,717410	0,155540	-0,5513	-0,695280	-0,018850	-0,018850	-0,646110
Bio11	0,664030	-0,786940	-0,280160	-0,846390	-0,62291	0,976070	-0,826900	0,976160	-0,418870	-0,41887	1	-0,237730	0,029430	-0,700590	0,68757	-0,024770	-0,609790	-0,609790	-0,089050
Bio12	-0,654220	-0,220420	-0,233250	-0,165520	-0,40314	-0,104050	-0,195250	-0,281300	-0,522210	-0,52221	-0,237730	1	0,859680	0,530070	-0,12892	0,919480	0,643110	0,643110	0,956600
Bio13	-0,563330	-0,477360	-0,328330	-0,437940	-0,6343	0,170280	-0,461500	0,002860	-0,717410	-0,71741	0,029430	0,859680	1	0,111540	0,35141	0,988720	0,237050	0,237050	0,941300
Bio14	-0,544320	0,477200	0,213900	0,506970	0,30996	-0,661360	0,492680	-0,733240	0,155540	0,15554	-0,700590	0,530070	0,111540	1	-0,80059	0,220420	0,955510	0,955510	0,359530
Bio15	0,210300	-0,740530	-0,416620	-0,732430	-0,66312	0,734200	-0,739420	0,693290	-0,551300	-0,5513	0,687570	-0,128920	0,351410	-0,800590	1	0,246640	-0,738790	-0,738790	0,113030
Bio16	-0,594340	-0,444980	-0,341500	-0,393200	-0,6067	0,121540	-0,422330	-0,057280	-0,695280	-0,69528	-0,024770	0,919480	0,988720	0,220420	0,24664	1	0,347980	0,347980	0,972710
Bio17	-0,591500	0,327440	0,159600	0,349580	0,13824	-0,547880	0,336580	-0,651450	-0,018850	-0,01885	-0,609790	0,643110	0,237050	0,955510	-0,73879	0,347980	1	1,000000	0,479940
Bio18	-0,591500	0,327440	0,159600	0,349580	0,13824	-0,547880	0,336580	-0,651450	-0,018850	-0,01885	-0,609790	0,643110	0,237050	0,955510	-0,73879	0,347980	1,000000	1	0,479940
Bio19	-0,611180	-0,370360	-0,264930	-0,327040	-0,54736	0,048670	-0,352850	-0,133860	-0,646110	-0,64611	-0,089050	0,956600	0,941300	0,359530	0,11303	0,972710	0,479940	0,479940	1

Table A.1: Correlation of the bioclimatic variables.

Variables	Principal C. 1	Principal C. 2	Principal C. 3	Principal C. 4	Principal C. 5
Bio_01	-0,027958	0,167132	0,205055	-0,041703	0,383784
Bio_02	0,372245	0,083815	-0,134326	0,262455	0,022927
Bio_03	0,130150	0,054812	0,123826	0,900006	0,080351
Bio_04	0,388698	0,062022	-0,240154	-0,178825	-0,027109
Bio_05	0,320872	0,186521	-0,055962	-0,045436	0,233648
Bio_06	-0,223647	0,067567	0,217370	-0,056393	0,215819
Bio_07	0,389371	0,076732	-0,202872	0,011614	-0,003411
Bio_08	-0,188475	0,126604	0,220058	0,059224	0,284280
Bio_09	0,219265	0,204653	0,028331	-0,161353	0,347419
Bio_10	0,219265	0,204653	0,028331	-0,161353	0,347419
Bio_11	-0,186437	0,106662	0,237072	0,033240	0,294365
Bio_12	-0,026912	-0,334092	-0,173702	-0,021839	0,324542
Bio_13	-0,104914	-0,236093	-0,346981	0,090083	0,239297
Bio_14	0,236434	-0,370478	0,258049	-0,084156	0,000600
Bio_15	-0,285927	0,131887	-0,377571	0,056720	-0,058408
Bio_16	-0,091400	-0,264006	-0,301640	0,036547	0,261251
Bio_17	0,170676	-0,394198	0,279993	-0,022991	-0,006954
Bio_18	0,170676	-0,394198	0,279993	-0,022991	-0,006954
Bio_19	-0,074867	-0,315099	-0,248262	0,063595	0,317167

Variables	Principal C. 6	Principal C. 7	Principal C. 8	Principal C. 9	Principal C. 10
Bio_01	0,098191	0,057189	-0,146086	-0,115544	0,049869
Bio_02	-0,019133	-0,010674	0,136884	-0,030455	0,361951
Bio_03	0,052751	-0,012400	-0,119459	-0,010463	-0,111822
Bio_04	0,030739	0,040272	0,079208	0,135063	-0,197040
Bio_05	0,027902	0,029828	0,050812	-0,070657	0,012732
Bio_06	0,023585	-0,002093	-0,092515	-0,158819	-0,048826
Bio_07	0,001549	0,021123	0,104497	0,070568	0,044800
Bio_08	-0,004263	-0,063052	0,769960	0,464252	-0,041043
Bio_09	0,111595	0,070816	-0,101483	-0,022184	-0,068120
Bio_10	0,111595	0,070816	-0,101483	-0,022184	-0,068120
Bio_11	0,048485	0,007924	-0,148986	-0,153608	0,144746
Bio_12	-0,300415	-0,091955	-0,065375	0,097994	0,683626
Bio_13	0,006334	0,021612	0,278074	-0,487738	-0,327962
Bio_14	0,491734	-0,693012	0,050671	-0,087198	0,022278
Bio_15	0,770544	0,231860	0,015116	0,093133	0,289974
Bio_16	-0,037727	0,015947	0,188854	-0,266671	-0,073058
Bio_17	0,121497	0,465155	0,063144	0,012589	0,012309
Bio_18	0,121497	0,465155	0,063144	0,012589	0,012309
Bio_19	0,029582	-0,059092	-0,391980	0,597383	-0,337018

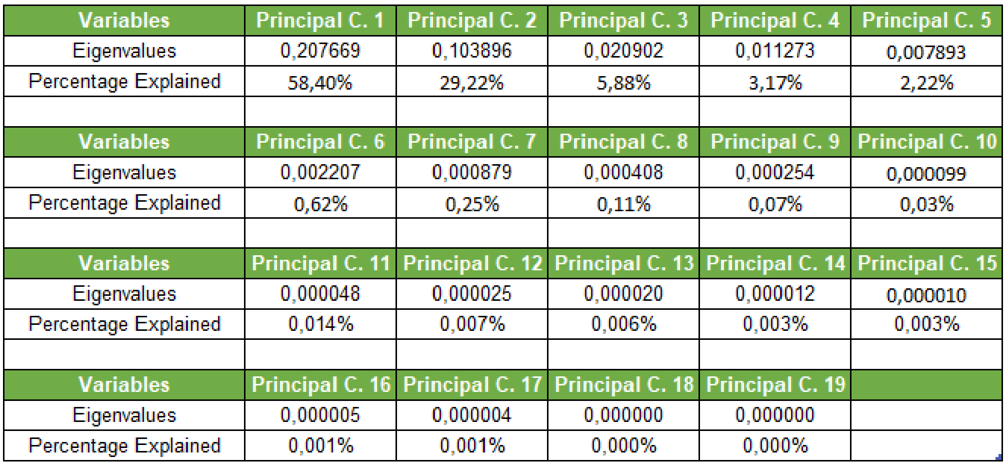
Variables	Principal C. 11	Principal C. 12	Principal C. 13	Principal C. 14	Principal C. 15
Bio_01	0,057823	-0,253081	0,640116	0,464873	-0,187236
Bio_02	0,592106	-0,315871	-0,003903	-0,256977	-0,233851
Bio_03	-0,322073	0,111790	-0,009489	0,016810	-0,025897
Bio_04	-0,258853	0,150080	0,172031	0,114877	-0,027769
Bio_05	0,222066	0,630846	0,145234	-0,043766	0,225635
Bio_06	0,151274	0,478146	0,076938	-0,384570	-0,361063
Bio_07	0,033282	0,048992	0,024933	0,203897	0,302305
Bio_08	-0,018439	-0,009468	-0,026930	0,020124	-0,004880
Bio_09	-0,219346	-0,191839	-0,302897	-0,157088	-0,126558

Table A.2 continued from previous page

Variables	Principal C. 1	Principal C. 2	Principal C. 3	Principal C. 4	Principal C. 5
Bio_10	-0,219346	-0,191839	-0,302897	-0,157088	-0,126558
Bio_11	0,199593	-0,061270	-0,261040	0,100571	0,666504
Bio_12	-0,284579	0,190125	-0,101059	0,157496	-0,135077
Bio_13	0,187132	0,048671	-0,308207	0,387682	-0,202448
Bio_14	-0,018542	0,005575	0,001257	0,004047	0,005050
Bio_15	-0,052922	0,087442	-0,004713	0,012480	-0,023165
Bio_16	-0,217702	-0,217305	0,418294	-0,526391	0,315610
Bio_17	0,023574	0,017860	-0,013160	-0,002193	0,004802
Bio_18	0,023574	0,017860	-0,013160	-0,002193	0,004802
Bio_19	0,313002	-0,025751	0,003685	-0,027366	0,020349

Variables	Principal C. 16	Principal C. 17	Principal C. 18	Principal C. 19	
Bio_01	-0,032625	0,027343	0,000000	0,000000	
Bio_02	0,158918	-0,139452	0,000000	0,000000	
Bio_03	0,028268	-0,019143	0,000000	0,000000	
Bio_04	0,634634	-0,391051	0,000000	0,000000	
Bio_05	-0,460367	-0,244790	0,000000	0,000000	
Bio_06	0,415121	0,321978	0,000000	0,000000	
Bio_07	0,125450	0,793735	0,000000	0,000000	
Bio_08	0,007371	-0,004093	0,000000	0,000000	
Bio_09	-0,104497	0,036448	0,001937	-0,707104	
Bio_10	-0,104497	0,036448	-0,001937	0,707104	
Bio_11	0,382176	-0,169159	0,000000	0,000000	
Bio_12	0,028800	-0,023033	0,000000	0,000000	
Bio_13	0,016526	-0,016274	0,000000	0,000000	
Bio_14	-0,003631	0,002331	0,000000	0,000000	
Bio_15	0,003432	-0,002338	0,000000	0,000000	
Bio_16	-0,025899	0,024981	0,000000	0,000000	
Bio_17	0,000859	0,000431	-0,707104	-0,001937	
Bio_18	0,000859	0,000431	0,707104	0,001937	
Bio_19	-0,019218	0,013863	0,000000	0,000000	

Table A.2: Eigenvalues of the first Principal Component Analysis applied to bioclimatic variables.



Variables	Principal C. 1	Principal C. 2	Principal C. 3	Principal C. 4	Principal C. 5
Eigenvalues	0,207669	0,103896	0,020902	0,011273	0,007893
Percentage Explained	58,40%	29,22%	5,88%	3,17%	2,22%
Variables	Principal C. 6	Principal C. 7	Principal C. 8	Principal C. 9	Principal C. 10
Eigenvalues	0,002207	0,000879	0,000408	0,000254	0,000099
Percentage Explained	0,62%	0,25%	0,11%	0,07%	0,03%
Variables	Principal C. 11	Principal C. 12	Principal C. 13	Principal C. 14	Principal C. 15
Eigenvalues	0,000048	0,000025	0,000020	0,000012	0,000010
Percentage Explained	0,014%	0,007%	0,006%	0,003%	0,003%
Variables	Principal C. 16	Principal C. 17	Principal C. 18	Principal C. 19	
Eigenvalues	0,000005	0,000004	0,000000	0,000000	
Percentage Explained	0,001%	0,001%	0,000%	0,000%	

Figure A.1: Percentage of variability explained by each of the 19 Principal Components.

Variables	Principal C. 1	Principal C. 2	Principal C. 3	Principal C. 4	Principal C. 5	Principal C. 6
Bio01	-0,04402	-0,18431	-0,17376	0,61295	0,70994	-0,04956
Bio02	0,456180	-0,13731	0,25671	0,10956	0,04619	0,72996
Bio04	0,477530	-0,11276	0,32858	-0,08657	0,06013	-0,61045
Bio07	0,477530	-0,13025	0,30927	0,00503	0,05059	-0,03988
Bio12	-0,0168	0,415050	0,17524	0,34179	-0,10625	-0,17836
Bio13	-0,12348	0,30784	0,373200	0,06573	0,0057	0,12415
Bio14	0,32261	0,409020	-0,30579	-0,3435	0,46993	-0,05872
Bio15	-0,37247	-0,11886	0,394690	-0,5484	0,48277	0,08757
Bio16	-0,10404	0,33932	0,31791	0,11478	0,00978	-0,02536
Bio17	0,24155	0,441010	-0,334070	-0,16093	-0,01089	0,16925
Eigenvalues	0,13244	0,07253	0,01594	0,00288	0,00171	0,00128
% Explained	58,149%	31,845%	6,999%	1,264%	0,751%	0,562%
Variables	Principal C. 7	Principal C. 8	Principal C. 9	Principal C. 10	Principal C. 11	
Bio01	0,18904	0,1238	-0,02106	0,01468	0,00172	
Bio02	-0,08956	-0,03109	0,15227	-0,0685	0,35114	
Bio04	0,17993	0,00352	-0,10108	0,00401	0,47425	
Bio07	0,06619	-0,01457	0,01963	0,05863	-0,80421	
Bio12	-0,0672	-0,09279	0,75549	0,22863	0,02574	
Bio13	-0,02642	0,5428	-0,36167	0,55103	0,03428	
Bio14	-0,53224	0,09411	0,02568	-0,00071	-0,00689	
Bio15	0,22786	-0,07486	0,30142	0,03865	0,00275	
Bio16	0,01415	0,34794	-0,03222	-0,79601	-0,05585	
Bio17	0,76057	0,03915	0,01697	0,02439	-0,00247	
Eigenvalues	0,00058	0,00028	0,00009	0,00002	0,00001	
% Explained	0,255%	0,123%	0,040%	0,009%	0,004%	

Table A.3: Eigenvalues and percentage of variability explained, for each Principal Component on the second PCA.

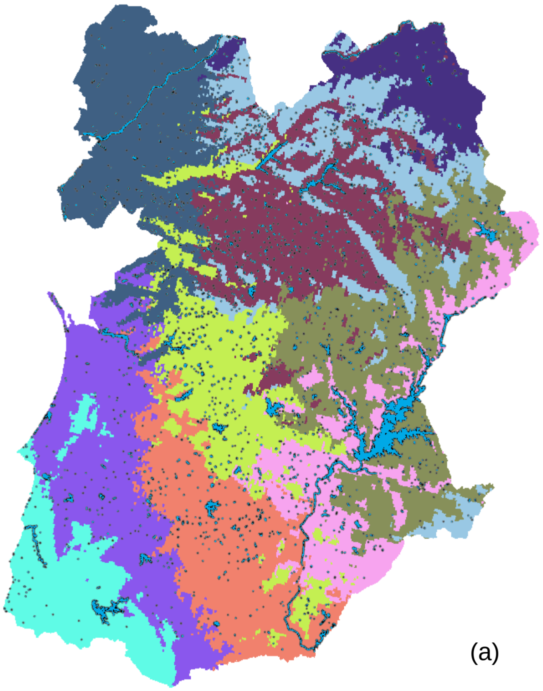
Cluster ID	*N*	Mean PC1	Mean PC2	Mean Bio01	Mean Bio02	Mean Bio04	Mean Bio07	Mean Bio12	Mean Bio13	Mean Bio14	Mean Bio16	Mean Bio17
1	3413020	0,40241	1,14867	165,37852	58,72637	3753,77270	17,20666	741,17282	118,42658	2,26832	342,28264	9,83084
2	6179274	0,71755	0,76998	168,53609	69,68472	4096,98160	19,41862	621,92021	100,53726	2,38459	288,83014	9,02204
3	5452419	1,13362	0,42625	171,74549	86,49603	4749,80176	23,23232	555,99067	90,58820	2,06377	258,40465	7,33628
4	6148873	1,14746	1,17492	167,40679	77,96245	4349,81263	21,05495	705,20730	103,62988	4,50675	305,84732	16,76407
5	4856372	1,34853	0,71080	168,86683	89,69458	4858,61238	23,86261	611,59335	95,29955	3,35927	273,87960	11,91894
6	4356304	1,61035	0,41743	172,54504	96,23686	5396,39433	26,23890	554,34409	83,69796	3,05887	243,30234	10,10293
7	6781882	1,71676	0,70521	167,36914	96,99322	5483,50711	26,56353	618,13280	91,30944	4,09579	266,55282	13,21190
8	6116184	1,70274	0,92827	165,41942	94,00159	5177,43367	25,26111	646,44029	91,76599	5,19475	270,77261	16,62262
9	4693056	1,68490	1,19431	161,67408	94,33337	5291,55364	25,65014	726,70111	105,58820	5,48181	311,38535	17,69910
10	2439424	1,91903	1,46719	155,87713	95,33694	5482,78653	26,34252	764,60087	106,75334	7,12424	317,41005	23,05267

Table A.4: Number of observations and mean values of relevant variables for each cluster, for ANOVA calculation.

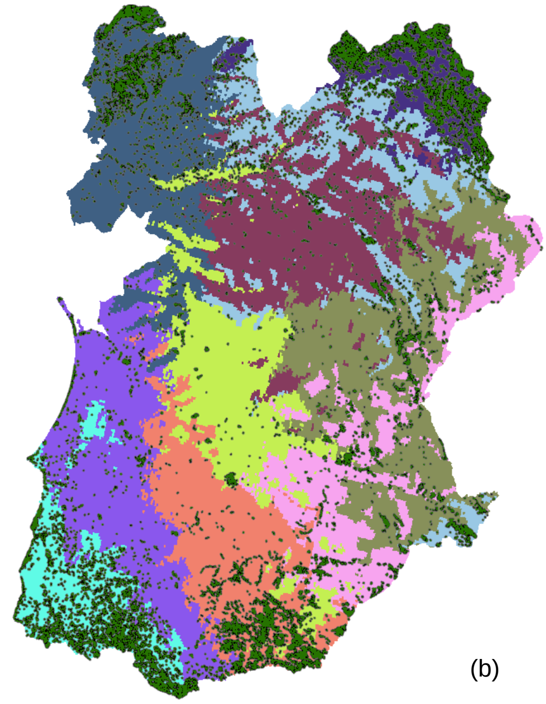
B

Supplementary Images

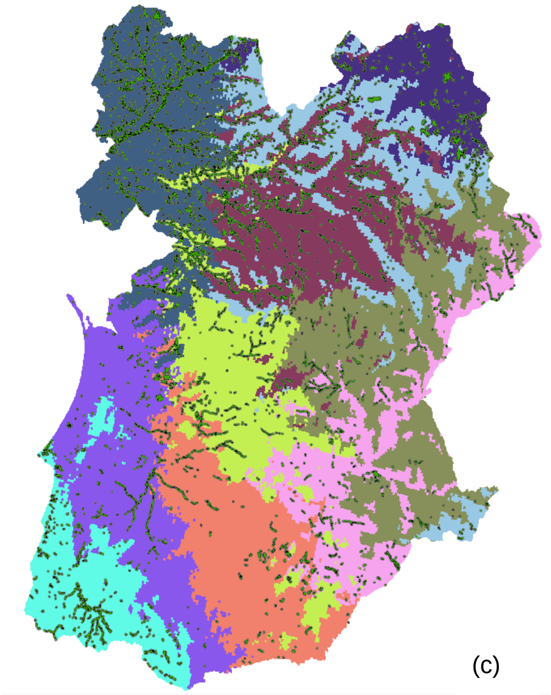
On this appendix we present, for the clustered map of Alentejo, the geographical distribution of the four semi-natural habitats, along with a table of area percentages. Image (a) shows the presence of water bodies on Alentejo, specifically when their presence is significant 50% or more of the year. (b) highlights the presence of shrubland, mostly distributed in the northernmost and southernmost parts of Alentejo as per the table. (c) refers to native forest presence, and, lastly, the image (d) shows the presence of riparian zones, described by their length in meters. The final image, (e), is a compilation of mean values of bioclimatic variables Bio07 and Bio17, along with a labelled cluster map.



Clusters	Water percentage
Cluster 1	3,8%
Cluster 2	3,9%
Cluster 3	5,7%
Cluster 4	4,9%
Cluster 5	6,6%
Cluster 6	14,6%
Cluster 7	6,5%
Cluster 8	8,1%
Cluster 9	4,1%
Cluster 10	2,6%

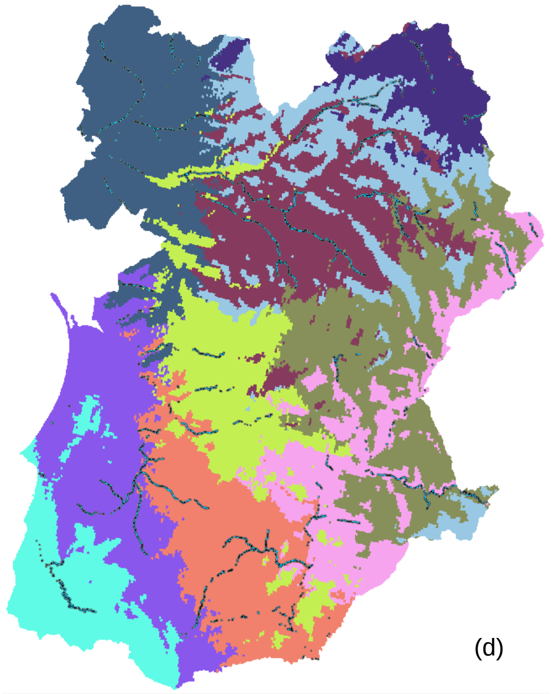


Clusters	Shrubland percentage
Cluster 1	47,5%
Cluster 2	11,7%
Cluster 3	21,6%
Cluster 4	22,0%
Cluster 5	6,6%
Cluster 6	11,6%
Cluster 7	9,0%
Cluster 8	7,5%
Cluster 9	17,1%
Cluster 10	52,6%



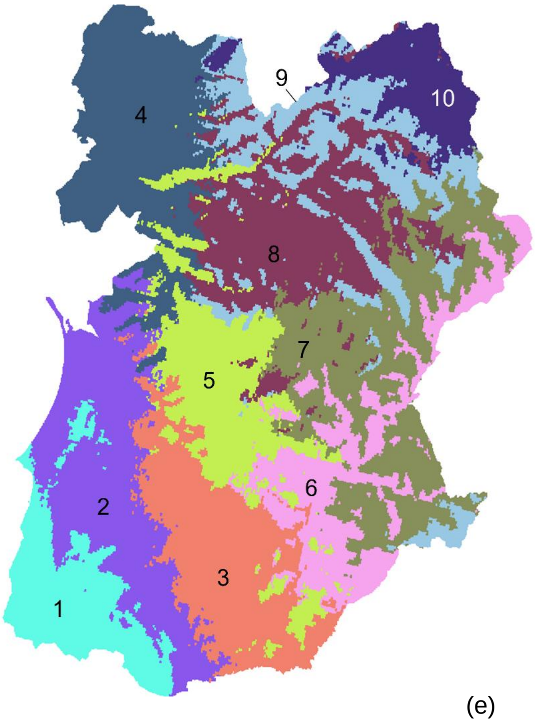
(c)

Clusters	Native forest percentage
Cluster 1	10,2%
Cluster 2	6,1%
Cluster 3	6,0%
Cluster 4	22,0%
Cluster 5	10,9%
Cluster 6	7,9%
Cluster 7	4,9%
Cluster 8	14,4%
Cluster 9	8,2%
Cluster 10	10,8%



(d)

Clusters	Riparian length (meter)
Cluster 1	70 383
Cluster 2	112 103
Cluster 3	218 755
Cluster 4	135 739
Cluster 5	70 886
Cluster 6	109 646
Cluster 7	81 415
Cluster 8	257 966
Cluster 9	23 539
Cluster 10	29 210



Clusters	Bio07 (°C)	Bio17 (kg/m2)
Cluster 1	17,2°	9,8
Cluster 2	19,4°	9,0
Cluster 3	23,2°	7,3
Cluster 4	21,1°	16,8
Cluster 5	23,9°	11,9
Cluster 6	26,2°	10,1
Cluster 7	26,6°	13,2
Cluster 8	25,3°	16,6
Cluster 9	25,7°	17,7
Cluster 10	26,3°	23,1

(e)