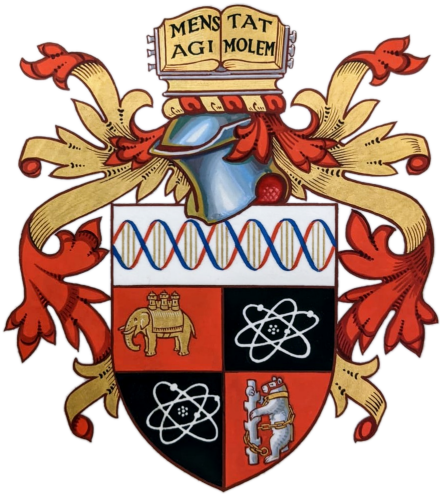
WARWICK
THE UNIVERSITY OF WARWICK



PREDICTIVE MODELING OF OLIVE YIELD IN SPAIN: A MACHINE LEARNING APPROACH TO ANALYZING THE IMPACT OF CLIMATE AND DROUGHT CONDITIONS

by

MHD RASLAN ALTINAWI

A thesis submitted in partial fulfilment

of the requirements for the degree of

MASTER OF SCIENCE IN DATA ANALYTICS

UNIVERSITY OF WARWICK

DEPARTMENT OF COMPUTER SCIENCE

September 2024

Contents

List of figures

1 Introduction

Spain stands as the world's leading producer of olives, contributing nearly 45% of the global olive oil production. This dominant position underscores the profound importance of the olive industry to Spain's economy, culture, and agricultural landscape. Olive cultivation in Spain is not merely a matter of economic significance; it is deeply intertwined with the nation's cultural heritage, particularly through its central role in the Mediterranean diet. Olive oil, often referred to as "liquid gold," is celebrated for its numerous health benefits, including its anti-inflammatory properties and its role in reducing the risk of chronic diseases. This has led to the global recognition and adoption of the Mediterranean diet, in which olive oil is a fundamental ingredient, further amplifying the value of olives beyond Spain's borders.

In recent years, the olive industry in Spain has faced unprecedented challenges, largely driven by climatic changes. These changes have manifested in several ways, including fluctuations in temperature, altered precipitation patterns, increased solar radiation, variations in humidity, and changes in soil moisture levels. Moreover, drought indices such as the Standardized Precipitation Index (SPI) and the Standardized Precipitation Evapotranspiration Index (SPEI), along with potential evapotranspiration rates, have shown significant deviations from historical norms. These climatic variables, collectively referred to as agro-climatic factors, have a direct and profound impact on agricultural productivity, with the olive sector being particularly vulnerable due to the crop's sensitivity to environmental conditions.

The recent climatic anomalies have had a tangible impact on olive yields across Spain, leading to considerable fluctuations in production. For instance, extreme temperatures during critical growing periods, coupled with irregular precipitation, have caused stress on olive trees, resulting in reduced fruit set and lower yields. Additionally, prolonged periods of drought, as reflected in the SPI and SPEI indices, have exacerbated water scarcity issues, further straining olive cultivation. The cumulative effect of these factors has led to a significant reduction in olive output in recent harvest seasons, which in turn has driven olive oil prices to record highs. This volatility in both production and pricing underscores the critical need to deepen our understanding of the relationship between agro-climatic variables and olive yields.

Understanding this relationship is not just a matter of academic interest; it has profound implications for the future of the olive industry in Spain. As climate change continues to alter weather patterns and exacerbate extreme weather events, the ability to predict

and mitigate the impacts on olive production becomes increasingly vital. This research aims to explore the intricate connections between agro-climatic factors and olive yield in Spain, with the goal of developing predictive models that can aid in the management and adaptation strategies for the olive industry. By examining historical climate data alongside olive yield records, this study will provide insights into how specific climatic variables influence production outcomes and identify the most vulnerable regions and periods.

Ultimately, the findings of this research could contribute to the development of more resilient agricultural practices, ensuring the sustainability of olive cultivation in Spain in the face of a changing climate. As Spain continues to play a pivotal role in the global olive market, safeguarding this industry is not only crucial for the country's economy but also for preserving a key component of the Mediterranean cultural and dietary heritage. Therefore, this study addresses a pressing need to align agricultural practices with the realities of climate variability, ensuring that the olive industry remains robust and capable of meeting both domestic and international demand.

2 Related Work

2.1 Introduction

The impact of climate change on agricultural production has become a critical area of research due to the increasing frequency and intensity of extreme weather events [1, 2]. In the Mediterranean region, where the climate is particularly variable, understanding how climatic factors affect crop yields is crucial for developing strategies to mitigate the negative impacts of climate change [3, 4]. Sustainable agriculture [5], which aims to meet current agricultural demands without compromising the ability of future generations to meet their own needs, is particularly important in this context. This is especially true for olive production in Spain, a country that leads the world in olive oil production and where olives are a key cultural and economic resource. Recently, the skyrocketing prices of olive oil [6, 7] due to yield losses highlight the necessity to study and address these challenges. Achieving sustainable agriculture, particularly in olive cultivation, requires ensuring that crop yields are sufficient to meet demand despite the adverse effects of climate change. This literature review explores the current understanding of the relationship between climatic factors and olive yield in Spain, with a focus on how drought, temperature, precipitation, and other meteorological variables influence olive production, and the implications for sustaining this vital agricultural sector.

2.2 Factors Influencing Crop Yield: The Role of Climatic Conditions

Crop yield is influenced by a complex interplay of various determinants, including genetic factors, soil quality, agricultural practices, and environmental conditions. Among these, environmental or climatic factors such as temperature, precipitation, and drought are particularly critical. These climatic factors have long been recognized as major determinants of crop yields due to their direct and indirect impacts on plant growth and development. The sensitivity of crops to these environmental variables varies by species, but for most agricultural systems, extreme weather conditions can lead to significant yield losses.

In the context of global climate change, the frequency and intensity of extreme events like heatwaves and droughts are expected to increase, posing a serious threat to food security worldwide. Understanding the impact of these climatic factors on crop yields is therefore essential for developing effective agricultural management practices and policy interventions

aimed at ensuring food security. For instance, temperature and precipitation patterns play a significant role in determining the timing and progression of various stages in a crop's life cycle, a concept known as crop phenology. These stages include germination (when the seed begins to grow), vegetative growth (when the plant develops leaves and stems), flowering (when the plant produces flowers), fruit set (when flowers develop into fruits or seeds), and maturity (when the crop is ready for harvest). Each of these stages is sensitive to climatic conditions; for example, unseasonal high temperatures during flowering can lead to poor fruit set, while insufficient rainfall during vegetative growth can stunt the plant's development. Drought, in particular, can have devastating effects by reducing soil moisture availability, leading to water stress and, consequently, lower yields. In light of these challenges, there is a growing body of research focused on assessing the impact of climatic variability on crop production using advanced data analysis techniques, including machine learning models that can handle complex, non-linear relationships between climate variables and crop yields.

2.3 Olive Production in Spain: Climatic Vulnerabilities and Cultural Significance

Olive trees (*Olea europaea*) are deeply integrated into the Mediterranean agricultural landscape, with Spain being the largest producer of olive oil globally. The cultivation of olives is not only economically significant but also culturally embedded, with olive oil being a staple of the Mediterranean diet, known for its health benefits, including reduced risks of cardiovascular diseases [8]. However, the productivity of olive groves is highly sensitive to climatic conditions, particularly to the Mediterranean climate's inherent variability.

The Mediterranean region, including Spain, is expected to face increasingly severe climatic challenges due to global warming. These challenges include higher temperatures, reduced and more erratic rainfall, and more frequent and intense droughts [9]. Such changes in climate are likely to have significant impacts on olive yields, making it imperative to study the relationship between olive crop and the wide array of environmental factors.

2.4 Impact of Climatic Factors on Olive Yield

2.4.1 Temperature and Heat Stress

Temperature is a crucial climatic factor affecting olive production. Olive trees are well-suited to the Mediterranean climate, characterized by mild winters and hot, dry summers. However, extreme temperatures, particularly heat stress during critical periods such as flowering and fruit set, can significantly reduce yields [10]. Studies have shown that prolonged exposure to high temperatures can lead to poor fruit set and increased fruit drop, thereby reducing the overall yield [11].

2.4.2 Precipitation and Water Availability

Water availability is another critical factor for olive production. Olive trees are relatively drought-tolerant, but their yield is still highly dependent on adequate water supply, particularly during the flowering and fruit development stages [9]. Insufficient rainfall, particularly during the growing season, can lead to water stress, which negatively impacts the quantity and quality of the olive yield. The relationship between precipitation and olive yield has been well-documented, with studies indicating that rainfall patterns in spring and early summer are particularly important [12].

2.4.3 Drought and Olive Yield

Drought is generally defined as a prolonged period of deficient precipitation that results in extensive damage to crops and other vegetation, disrupting agricultural production. It is a complex phenomenon that can be classified into various types, such as meteorological drought (lack of precipitation), agricultural drought (insufficient moisture for crops), and hydrological drought (reduced water levels in lakes, rivers, and reservoirs). Understanding and quantifying drought is essential for assessing its impacts on agriculture and developing strategies to mitigate its effects.

To quantify drought, researchers and meteorologists use drought indices, which are numerical tools that signify the intensity or degree of drought or wet conditions over a specific period. These indices are crucial for identifying drought onset, duration, and intensity, making them vital for agricultural planning and risk assessment.

One of the most widely used drought indices is the Standardized Precipitation Index (SPI) [13]. SPI is based solely on precipitation data and measures the deviation of precipitation from its long-term mean for a specified time scale (e.g., 1 month, 3 months, 12 months). It is a flexible index because it can be used to monitor drought on multiple time scales, reflecting different types of drought (e.g., agricultural drought on short scales and hydrological drought on longer scales). The SPI is calculated by fitting historical precipitation data to a probability distribution, which is then transformed into a standard normal distribution, where positive SPI values indicate wetter-than-average conditions and negative values indicate drier-than-average conditions.

However, drought is not only about precipitation; temperature also plays a significant role in determining drought intensity. To incorporate the effects of temperature, the Standardized Precipitation-Evapotranspiration Index (SPEI) was developed [14]. SPEI is similar to SPI but also considers potential evapotranspiration (PET), which represents the amount of water that would evaporate and be transpired by vegetation if sufficient moisture were available. PET is influenced by factors such as temperature, humidity, wind speed, and solar radiation, making it a critical measure of the atmospheric demand for moisture [15, 16]. By accounting for both precipitation and PET, SPEI provides a more comprehensive assessment of drought conditions, especially in the context of climate change, where rising temperatures can increase evapotranspiration and exacerbate drought impacts

even if precipitation levels remain unchanged.

For instance, SPEI can show that a region might experience a more severe drought under higher temperatures due to increased evapotranspiration, even if precipitation levels are similar to those in previous years. This makes SPEI particularly useful in regions where temperature fluctuations significantly influence water availability.

Utilizing these indices, research has established a significant link between drought conditions and the sensitivities of major crops like wheat, maize, rice, and soybeans in leading agricultural countries. These studies have found that these crops exhibit varying levels of sensitivity, with the risks expected to increase under future climate scenarios. Notably, the response of crop yields to increasing drought conditions is non-linear, meaning that even small increases in drought intensity can lead to disproportionately large declines in crop yields. This finding underscores the critical importance of using drought indices like SPI and SPEI to enhance the accuracy of predictions and risk assessments, thereby informing more effective agricultural strategies [17].

Likewise, in Spain, Drought is one of the most significant climatic challenges facing olive production. The frequency and severity of drought events are expected to increase due to climate change, posing a significant threat to olive yields [9]. Drought stress can affect olive trees in several ways, including reducing the growth of new shoots, decreasing the number of flowers, and leading to early fruit drop. Additionally, prolonged drought conditions can lead to permanent damage to the trees, further reducing yields in subsequent years [11].

Research has highlighted the need to develop drought-resistant olive cultivars and implement water-saving irrigation techniques to mitigate the impact of drought on olive production. Furthermore, predictive models that incorporate drought indices and other climatic variables are essential for forecasting the potential impacts of drought on olive yields and for developing effective adaptation strategies [10].

2.4.4 Integration of Climate Data in Olive Yield Studies

The use of comprehensive climate data, such as those provided by TerraClimate and the Spanish State Meteorological Agency (AEMET), has become increasingly important in studying the impact of climatic factors on olive yield. These datasets offer high-resolution information on temperature, precipitation, and drought conditions, which are crucial for understanding the complex interactions between climate and olive production [18]. By integrating these data sources with machine learning models, researchers can more accurately assess the impact of climatic variability on olive yields and develop more robust predictive models.

2.5 Methodologies in Analyzing Climatic Impact on Olive Yield

Recent studies have employed various methodologies to analyze the impact of climatic factors on crop yields. Machine learning techniques, particularly deep learning, have proven effective in modeling the complex relationships between climate variables and crop performance. These models can handle large datasets and capture non-linear interactions, making them particularly suited for studying the impact of climate on olive yields [18].

Additionally, traditional statistical methods, such as regression analysis, continue to be widely used to identify key climatic variables that influence olive production. These methods are often combined with machine learning techniques to enhance the accuracy of predictions and to better understand the underlying mechanisms driving yield variability [10].

2.6 Conclusion

The impact of climatic factors on olive yields in Spain is a critical area of research that has significant implications for both the economy and food security. Understanding how temperature, precipitation, and drought affect olive production is essential for developing strategies to mitigate the negative impacts of climate change. The integration of climate data from sources like TerraClimate and AEMET with advanced modeling techniques offers a promising approach to studying these impacts. As the climate continues to change, research in this area will be vital for ensuring the sustainability of olive production in Spain and for preserving the cultural and economic value of this important crop.

Future research should focus on refining the models used to predict the impact of climatic factors on olive yields, incorporating more detailed climate data, and exploring the potential of adaptive strategies such as the development of drought-resistant cultivars and improved irrigation techniques. By advancing our understanding of these relationships, we can better prepare for the challenges posed by climate change and ensure the continued success of olive production in Spain.

3 Methodology

3.1 Software and Tools

To conduct the research for this dissertation, several software tools were utilized to manage data collection, preprocessing, and modeling. The choice of these tools was guided by their specific capabilities in handling large-scale climate data and performing advanced machine learning tasks.

1. Jupyter Notebook: An interactive environment that was essential for integrating various tools and ensuring reproducibility.

2. R (with spei and climaemet packages): Used for climate data analysis and drought index calculation, leveraging specialized packages for processing and analyzing climate-related data.

3. Google Earth Engine (GEE): Google Earth Engine is a cloud-based platform designed for the processing and analysis of large-scale geospatial data, particularly satellite imagery and climate datasets. It provides an online JavaScript code editor that enables users to write and execute scripts without the need for local software installation or data downloads. In the context of my research, GEE was used to generate zonal statistics—quantitative summaries of data within specified geographic regions—using pre-defined maps. This functionality allows for efficient spatial analysis, where specific metrics (such as mean, sum, or distribution of pixel values) are calculated for designated zones, facilitating detailed environmental and geographical studies.

4. Python (scikit-learn, XGBoost): Employed for predictive modeling, including feature selection, model training. Sktime was used for handling time-series related tasks such as data splitting and imputing missing values.

Each of these tools played a critical role in different stages of the research process, contributing to the robustness and reliability of the findings presented in this dissertation.

3.2 Data Collection

This section outlines the steps and processes undertaken to collect, pre-process, and analyze the data for the study. The primary goal was to assess the relationship between climate

factors, drought indices, vegetation health, and olive yield in Spain using machine learning models. The data was sourced from various remote sensing platforms, government agencies, and climatic databases, which were then processed to create a comprehensive dataset spanning multiple years. The detailed methodology is described below. We will also discuss decisions made along the way to include or exclude certain datasets.

3.2.1 Spanish Ministry of Agriculture

Given our study's focus is on Spain olive yield, the most credible and reliable resource which provide a comprehensive data about the agriculture sector in Spain is The Spanish Ministry of Agriculture. We'll highlight the most relevant datasets for our study.

3.2.1.1 Annual Crop Areas and Productions

The annual crop areas and production provide annual reports covering all planted crops in Spain, including olive, at the province and national level. These reports include varying information such as the total planted area, the total yield, the extracted products from crops and so on. We focus specifically on olive crop, and we utilize the following data sources:

1. **Provincial Analysis of Olive Production by Destination**: This dataset provides an annual analysis of olive production across all provinces of Spain, focusing on the destination of the olives. The data are categorized into two main types:

   - **Table Olives** (Aceituna para aderezo): Olives destined for direct consumption as table olives.
   - **Olive Oil** (Aceituna para almazara): Olives processed for oil production.

   The dataset aggregates the production figures without distinguishing between different farming practices or cultivation methods, offering a broad overview of provincial contributions to both table olive and olive oil production. This data is part of an annual series, available from 1998 to 2022.

2. **Provincial Analysis of Olive Production by Cultivation Method and Production Status**: This source consists of two datasets, one focusing on table olives (Aceituna total de mesa) and the other on olives destined for oil production (Aceituna para almazara). Both datasets provide detailed information at the provincial level, differentiating between:

   - **Cultivation Method**
     - **Secano**: Refers to dry farming practices where olives are grown without irrigation, relying solely on natural rainfall.
     - **Regadío**: Refers to irrigated farming, where olives are grown with the aid of artificial irrigation systems.
   - **Production Status**

- **En plantación regular**: Represents the area under regular olive plantations, irrespective of whether they are currently producing olives.
- **En producción**: Refers specifically to areas that are actively producing olives during the recorded period.

These datasets are available for both table olives and olives destined for oil production, offering a nuanced understanding of how different farming practices and the status of plantations contribute to the overall production within each province. It will enable us to understand how olive is planted in general (rainfed vs. irrigated), and if there's a relation between the province and the type of irrigation. Like the first dataset, these data are available annually from 1998 to 2022, facilitating detailed temporal analysis across 25 years.

All of the above datasets are available as excel files and has been downloaded manually. They are available only in Spanish and require some pre-processing which will be covered in later sections. Although there is data available before 1998, however, it's only available as scanned PDF files, which requires manual processing to extract the information and using OCR tools to recognize text from images, so the decision has been made to consider only data from 1998 onward.

3.2.1.2 Monthly Advances in Agricultural Surfaces and Production

In addition to annual reports, the ministry also provides more-granular monthly reports called **Monthly advances in agricultural surfaces and production** (Avances de superficies y producciones de cultivos). These reports provide, on a monthly basis, the estimated annual crop yield based on observed advances in areas and production. This is covered for all crops in Spain. Such data reflects the evolution of the surface areas and production of each crop, which can be crucial for our study to understand how monthly climatic variables correlate and influence olive crop advancement. However, this data covers only the years 2007-2024, and it's available as PDF files only for most of the years. Due to challenges in preparing this data, a decision has been made to omit this data.

3.2.2 Climatic and Environmental Variables

Central to our study is the use of climatic and environmental variables as well as drought indices, to investigate their effects on the overall olive yield.

Here, we'll discuss the most relevant sources that can be used and our assessment of the suitability of each source regarding its accuracy and ease of access.

3.2.2.1 SPEIbase

[19]

SPEIBase is a global dataset providing Standardized Precipitation Evapotranspiration Index (SPEI) values, which are essential for analyzing drought patterns. One of the

key strengths of SPEIBase is its ability to provide SPEI values calculated across multiple timescales, such as SPEI1, SPEI3, SPEI6, SPEI12, and beyond. These different timescales represent the number of months over which the index is calculated, allowing researchers to assess drought conditions over various durations. For instance, SPEI1 represents a short-term index calculated over one month, useful for identifying immediate water stress, while SPEI12 reflects a long-term index calculated over twelve months, which is more suitable for understanding prolonged droughts or wetter periods.

Additionally, SPEIBase has a monthly temporal resolution, meaning that the SPEI values are updated and available for every month, providing a consistent and regular dataset that is ideal for ongoing monitoring and analysis of drought trends. This monthly availability is particularly beneficial for detecting seasonal variations and understanding how drought conditions evolve over time. Moreover, SPEIBase offers extensive temporal coverage, typically spanning from 1901 to the present. This long-term coverage is invaluable for historical analysis, enabling researchers to track and compare drought events over more than a century.

Initially, we considered SPEIBase as a primary data source for this study due to its comprehensive drought index, flexible timescales, frequent updates, and extensive time coverage. However, two significant limitations became apparent. First, SPEIBase only offers SPEI values, which restricts its applicability when additional climate variables, such as temperature, precipitation, or soil moisture, are needed for a more robust and comprehensive analysis. Second, its relatively coarse spatial resolution—approximately 55.66 kilometers—poses challenges for conducting studies at the province level in Spain. Accurate modeling in such regions requires fine-scale spatial details, especially in areas with diverse microclimates where climate conditions can vary significantly over short distances.

These limitations led us to explore alternative datasets that are better suited to the specific requirements of this study.

3.2.2.2 TerraClimate [20]

To enhance the spatial resolution and accuracy of climate data used in this study, data was sourced from TerraClimate, a high-resolution climate dataset that is well-suited for modeling localized climate impacts on agricultural yield. TerraClimate provides monthly observations of several key climate variables, including actual evapotranspiration, maximum and minimum temperature, potential evapotranspiration, precipitation, and soil moisture, covering the period from 1998 to 2023.

For this study, data covering the specific period from 1998 to 2023 was accessed and extracted using Google Earth Engine (GEE). TerraClimate, which is satellite-based, offers a spatial resolution of approximately 4,638.3 meters, allowing for a granular analysis of climatic conditions at a more localized level. This level of detail is crucial for capturing the microclimatic variations that significantly affect olive yield, particularly in the diverse landscapes where olive groves are cultivated.

Given that TerraClimate does not directly provide the Standardized Precipitation-Evapotranspiration Index (SPEI) or the Standardized Precipitation Index (SPI), these indices were manually calculated using the R package spei. This package is specifically designed for drought analysis, enabling the integration of high-resolution climate data into the drought indices. The manual calculation of SPEI and SPI from TerraClimate data, although methodologically complex, provides a more accurate reflection of localized drought conditions.

In summary, the use of TerraClimate data, coupled with the manual calculation of relevant drought indices, ensures that the climate variables incorporated into the predictive models are of high spatial and temporal resolution. This approach enhances the accuracy and reliability of the models in assessing the impact of climate and drought on olive yields over the study period.

3.2.2.3 AEMET Open Data

Here is the map of AEMET stations in Spain:

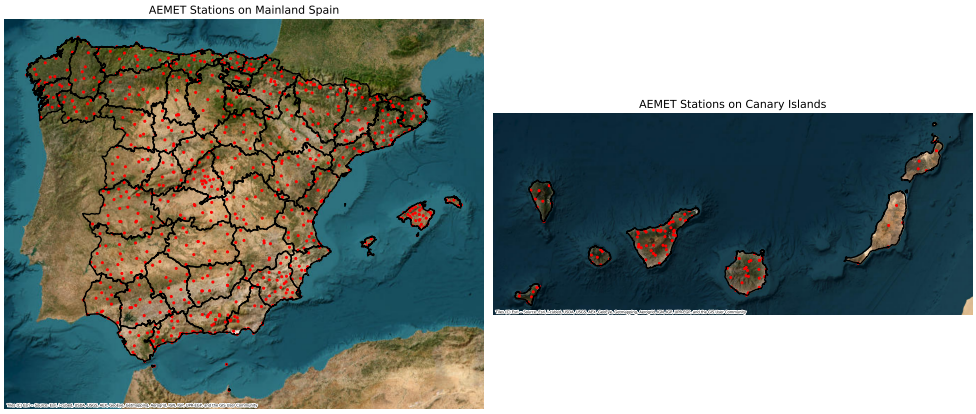


Figure 3.1: Map of AEMET Stations on Mainland Spain and Canary Islands

In addition to the TerraClimate data, we considered using climate data from the *Agencia Estatal de Meteorología* (AEMET) as an alternative source for our analysis. While TerraClimate provides satellite-based data with global coverage, AEMET offers ground-based, high-resolution climate data specifically for Spain. This difference in data sources provides an opportunity to compare the effectiveness of predictive models using satellite-based versus ground-based observations.

AEMET's dataset consists of daily meteorological records collected from a network of weather stations distributed across mainland Spain and the Canary Islands. Each province is covered by multiple stations, ensuring detailed spatial resolution. The locations of these stations are illustrated in Figure 3.1, where the red dots indicate the distribution of AEMET stations.

The AEMET data includes various meteorological variables, such as the date of observation, unique identifiers for each weather station, and the station's name and location. It

also records the altitude of the station, daily average temperature, daily precipitation, minimum and maximum temperatures with their respective times, wind speed and direction, sunshine duration, atmospheric pressure, and relative humidity. These variables provide comprehensive daily records that are critical for analyzing climate impacts on olive yield.

To facilitate the testing of machine learning models on this dataset, we used the `climaemet` R package to efficiently download daily meteorological data from AEMET's servers. This data provides an opportunity to explore how predictive models perform when trained on ground-based observations, in contrast to the satellite-based TerraClimate data.

By comparing the outcomes from models using these two distinct data sources, we aim to assess the relative strengths and weaknesses of each dataset in predicting olive yields under varying climatic conditions across Spain.

3.2.3 Vegetation Data

The study employs vegetation data to examine the relationship between climate variables and olive yield across different provinces in Spain. Vegetation indices, such as the Normalized Difference Vegetation Index (NDVI) and the Enhanced Vegetation Index (EVI), are critical proxies for assessing vegetation health, photosynthetic activity, and overall biomass productivity. These indices are particularly valuable in agricultural studies as they reflect the greenness and vitality of crops, which are directly influenced by climatic conditions.

For this research, NDVI and EVI data were sourced from the MODIS/061/MOD13Q1 dataset, covering the period from 2000 to 2022. The monthly average values of these indices were calculated for each province to create a detailed temporal and spatial dataset. By analyzing these indices over time, the study aims to capture the dynamic response of olive trees to varying climatic conditions, thus providing insights into how climate variability influences olive yield across different regions. This temporal data is crucial for understanding long-term trends and patterns that may affect agricultural productivity in Spain.

3.2.4 Maps

The spatial framework for this study was established using vector maps obtained from the Instituto Geográfico Nacional (IGN) of Spain. The primary map utilized was the "Database of Administrative Divisions of Spain," which provides comprehensive vector data delineating the boundaries of Spain's administrative units, including provinces. This high-resolution map was essential for visualizing spatial patterns, performing spatial aggregations, and integrating the vegetation data with other geographic datasets within the Google Earth Engine environment.

In addition to its use for spatial aggregation of the MODIS vegetation indices, the vector map was also employed in the calculation of the Standardized Precipitation-Evapotranspiration Index (SPEI) and the Standardized Precipitation Index (SPI), as detailed in previous sections. These drought indices were calculated at the provincial level, necessitating accurate spatial delineation to ensure that the climatic data were correctly aligned with the administrative

boundaries. The precise mapping provided by the IGN dataset allowed for the accurate aggregation and analysis of these climatic indices, which are crucial for understanding the impacts of drought conditions on olive yield.

Furthermore, the vector map facilitated the production of visual representations of the spatial distribution of vegetation indices, SPEI, and SPI across Spain. These visualizations were instrumental in interpreting the complex relationships between climate variables, vegetation dynamics, and agricultural productivity, particularly in the context of olive cultivation.

3.3 Exploratory Data Analysis (EDA)

3.3.1 Introduction: Olive Yield Trends in Spain

Spain is the world's largest producer of olives, and most of this production is concentrated in just a few provinces, particularly in the autonomous community of Andalucía. Figure 3.2 illustrates the yearly total olive yield in Spain from 1998 to 2022. The chart shows a general upward trend with significant year-to-year variability. Peak production years such as 2011, 2013, and 2018 exceeded 9 million tons, while in other years, such as 2012 and 2022, yields dropped dramatically below 5 million tons. This variability is primarily driven by external factors, especially climate conditions, which affect production capacity.

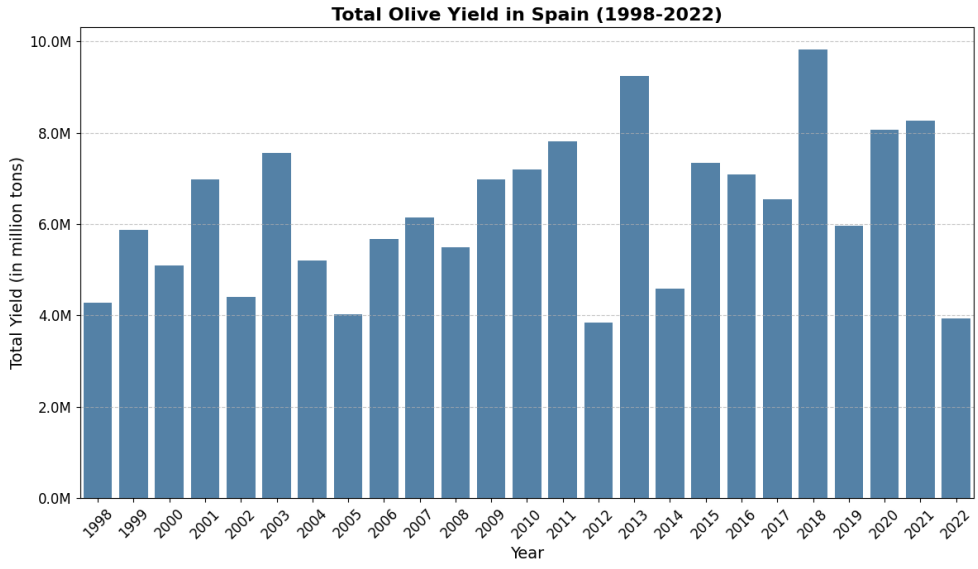


Figure 3.2: Total Olive Yield in Spain (1998–2022)

3.3.2 Regional Concentration of Olive Production

Olive production in Spain is highly concentrated in Andalucía. Figure 3.3 shows the top 10 olive-producing provinces in 2022, with Jaén accounting for 22.60% of the total national production, followed by Córdoba and Sevilla. The map in Figure 3.4 further emphasizes

this regional dominance, with the southern provinces producing the majority of olives, while the northern provinces contribute very little.

Figure 3.5 examines the yearly olive yield trends for the top 5 producing provinces from 1998 to 2022. The figure highlights how Jaén consistently outproduces the other provinces, with Córdoba and Sevilla also being significant contributors. This further supports the observation that olive production is geographically concentrated and dependent on a few key provinces in Andalucía.

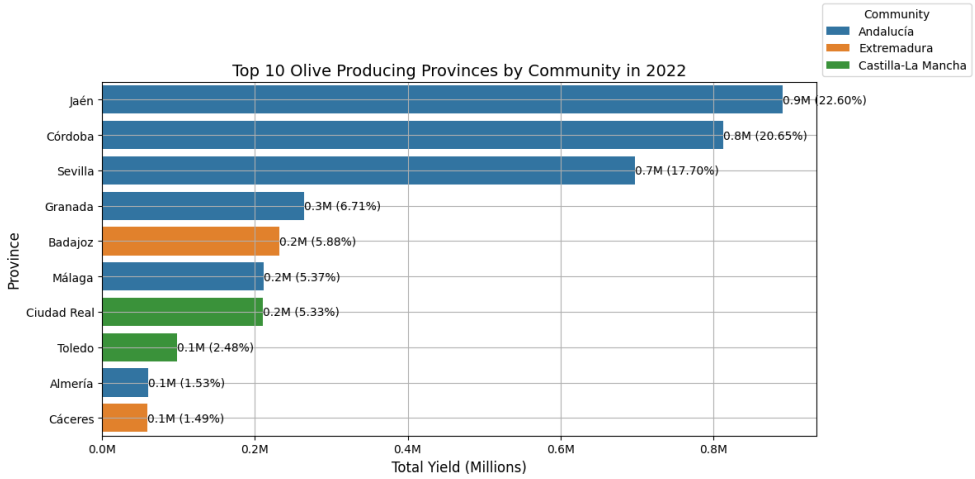


Figure 3.3: Top 10 Olive Producing Provinces by Community (2022)

3.3.3 Olive Cultivation Systems: Rainfed vs. Irrigated

A large portion of Spain's olive cultivation is rainfed, with irrigation being used less frequently. Figure 3.6 shows that both olive oil and table olive production rely predominantly on rainfed systems. Irrigated systems produce a significant amount of olive oil, but rainfed systems contribute consistently more, especially in terms of table olives, which remain heavily rainfed.

This reliance on rainfed systems underscores the importance of adequate precipitation for maintaining high production levels. Figure 3.6 highlights that rainfed production remains the backbone of the olive industry in Spain, making it highly sensitive to changes in precipitation patterns. Adequate rainfall during key phenological stages such as the flowering and harvesting periods is crucial for sustaining yields.

Figure 3.7 further contrasts table olive and olive oil production over time. While olive oil production fluctuates significantly, often driven by climate factors, table olive production remains relatively stable. This distinction highlights the different climatic requirements for these two crops, with olive oil production being more sensitive to temperature and precipitation variability.

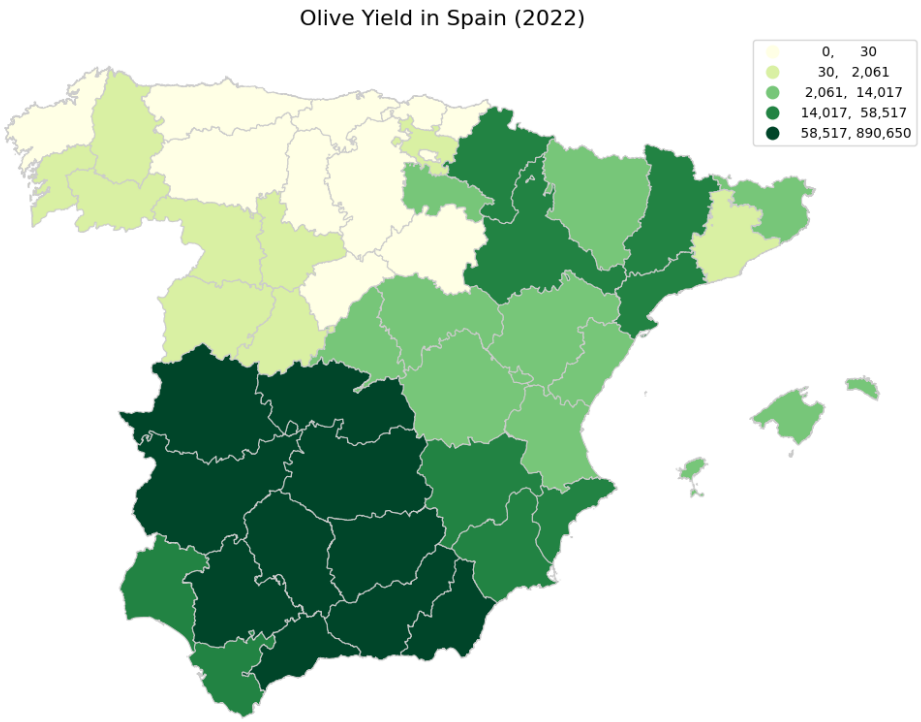


Figure 3.4: Olive Yield in Spain by Province (2022)

3.3.4 Olive Yield Correlation with Climate Variables

3.3.4.1 Correlations with TerraClimate Variables

The TerraClimate data provides insight into how climate variables influence olive yield during different phenological periods. In Figure 3.8, the correlation matrix shows that maximum temperature and standardized precipitation-evapotranspiration index (SPEI) are moderately correlated with olive yield during the flowering season (April, May, June). SPEI_6M and SPEI_9M show particularly strong positive correlations, indicating that drought conditions, as captured by these indices, have a significant impact on yield. Water balance and potential evapotranspiration, on the other hand, show weaker or negative correlations, highlighting the complexity of interactions between water availability and crop development during this stage.

During the harvesting season (October, November, December), Figure 3.9 shows that the correlations between climate variables and yield become less pronounced, with maximum temperature and precipitation maintaining some influence, but drought indices like SPEI losing their predictive strength.

3.3.4.2 Stronger Correlations with AEMET Data

The AEMET dataset, with its localized climate information, shows stronger and more consistent correlations with olive yield, especially during the flowering season (Figure 3.10).

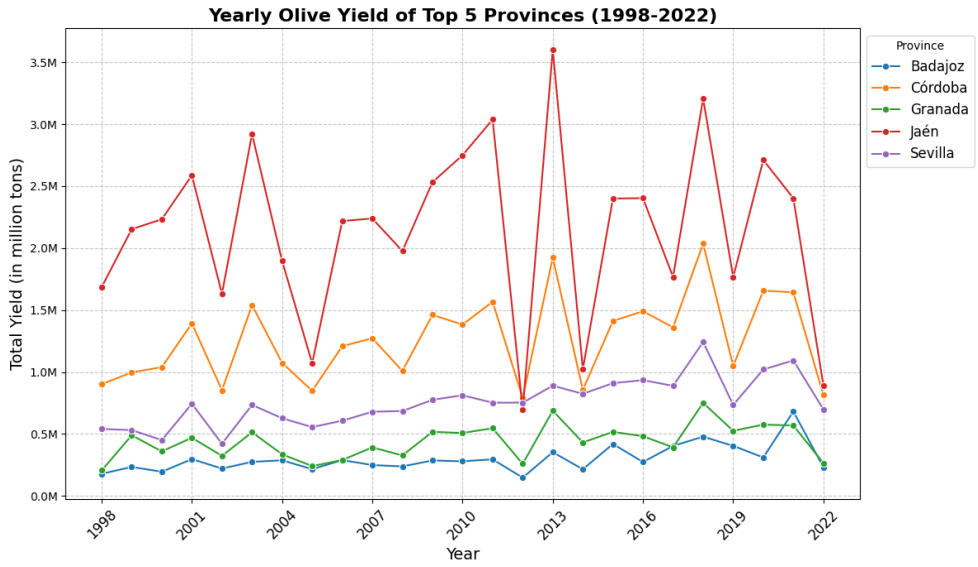


FIGURE 3.5: Yearly Olive Yield of Top 5 Provinces (1998–2022)

FIGURE 3.6: Irrigated vs. Rainfed Olive Production (1998–2022)

Variables such as maximum temperature and precipitation are highly correlated with yield, with wind speed showing a strong negative correlation. These correlations remain similarly strong during the harvesting season (Figure 3.11), suggesting that the AEMET dataset provides a more accurate reflection of microclimatic conditions affecting yield.

3.3.5 CASE STUDY: CLIMATE TRENDS AND OLIVE YIELD IN JAÉN

Jaén, the top olive-producing province, serves as a case study for examining how long-term climate trends affect olive yield. Figures 3.12 and 3.13 show the evolution of minimum and

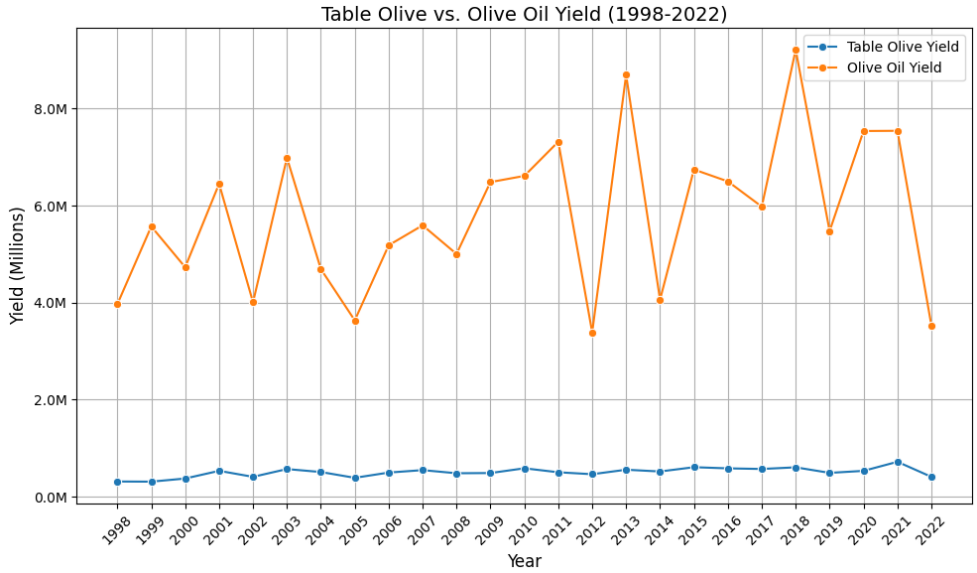


FIGURE 3.7: Table Olive vs. Olive Oil Yield (1998–2022)

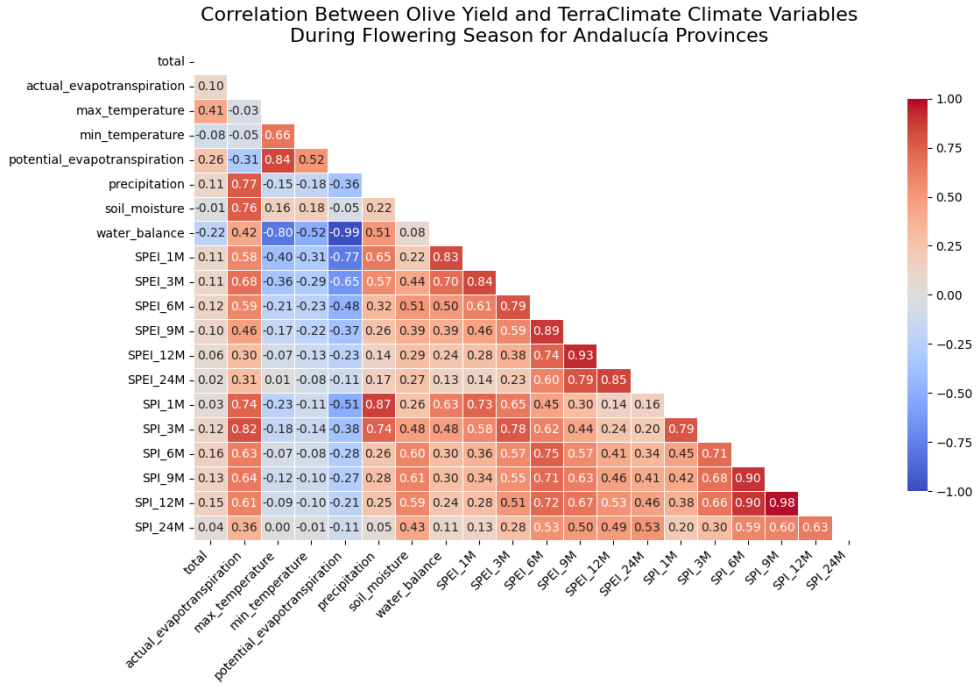


FIGURE 3.8: Correlation Between Olive Yield and TerraClimate Variables During Flowering Season (April-June) in Andalucía Provinces

maximum temperatures in Jaén from 1998 to 2023. These figures reveal a steady increase in temperature, particularly during the flowering season (April-June) and harvesting season (October-December). The temperature trends suggest that heat stress may become a more frequent issue, impacting yields, especially during flowering.

The precipitation patterns in Jaén also show significant variability, as seen in Figure 3.14. Notable droughts occurred in 2006, 2012, and 2017, years where olive yield was

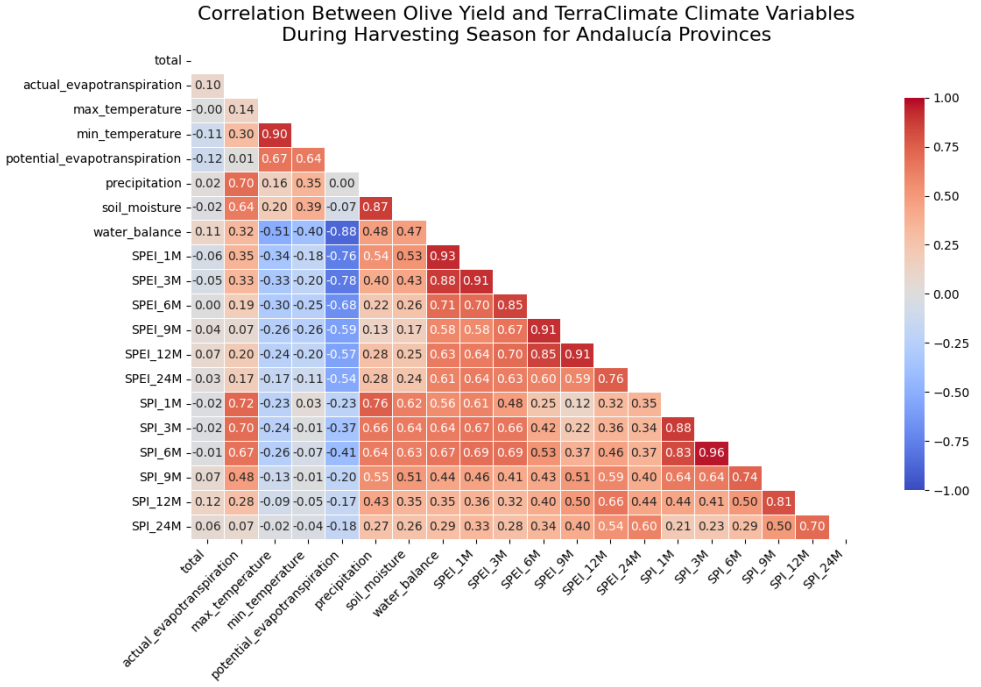


Figure 3.9: Correlation Between Olive Yield and TerraClimate Variables During Harvesting Season (October-December) in Andalucía Provinces

significantly lower. In contrast, years like 2010 and 2014 experienced above-average precipitation, particularly in key months such as April and October. These fluctuations in rainfall, combined with the prevalence of rainfed cultivation, underscore the critical role that adequate precipitation plays in maintaining consistent yields.

In addition, the Standardized Precipitation-Evapotranspiration Index (SPEI), as illustrated in Figure 3.15, highlights prolonged periods of drought stress, particularly in 2015 and 2019. These negative SPEI values correspond to the sharp declines in olive yield, reaffirming the vulnerability of olive production to drought conditions. The trends in Jaén provide a clear case of how temperature and precipitation fluctuations, driven by climate variability, directly influence yield outcomes.

3.3.6 Challenges in Predictive Modeling: Yield Skewness and Variability

Predictive modeling of olive yield faces significant challenges due to the highly skewed distribution of yield data. As shown in Figure 3.16, the yield data is positively skewed, with a few high-yield years disproportionately affecting the overall trend. The large standard deviation and high kurtosis values further highlight the spread of the data, indicating the presence of outliers that make accurate predictions difficult.

Additionally, the variability of climate conditions, as seen in the calendar plots and correlation heatmaps, adds to the complexity of prediction. Extreme weather events, such as droughts and heatwaves, further exacerbate this variability. Thus, future models must

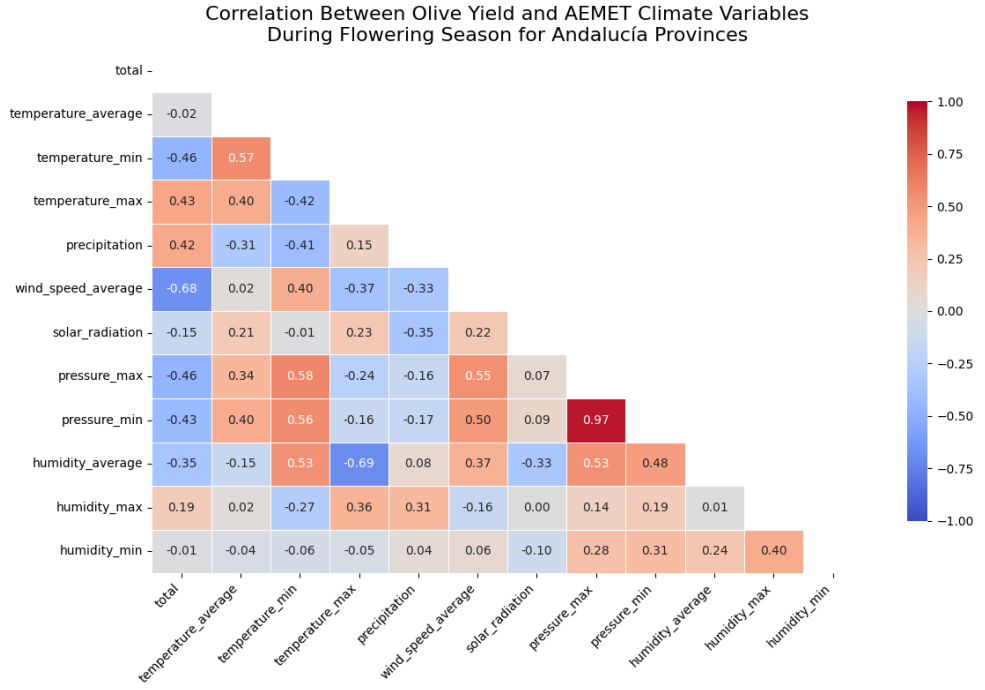


Figure 3.10: Correlation Between Olive Yield and AEMET Climate Variables During Flowering Season in Andalucía Provinces

account for non-linear relationships between climate variables and olive yield, and they must be capable of handling the skewed nature of the yield data to improve prediction accuracy.

3.4 Data Pre-processing and Transformations

This section outlines the steps taken to prepare the data for predictive modeling. The focus is on converting various datasets to a monthly frequency and structuring them in a tabular format suitable for machine learning models. Monthly aggregation is motivated by the need to capture the temporal dynamics of climate, vegetation, and environmental conditions over the course of a year, which are crucial for accurately predicting annual olive yield. The following subsections detail the data preparation process.

3.4.1 Data Transformation and Aggregation

3.4.1.1 Motivation for Monthly Aggregation and Data Transformation

The predictive modeling approach adopted in this study estimates annual olive yield for each province based on monthly climatic and vegetative variables. By aggregating the data to a monthly level, we capture seasonal variations and specific climatic events that affect olive yield. Additionally, transforming the data into a tabular format with monthly variables as features allows the model to learn from temporal patterns within each year.

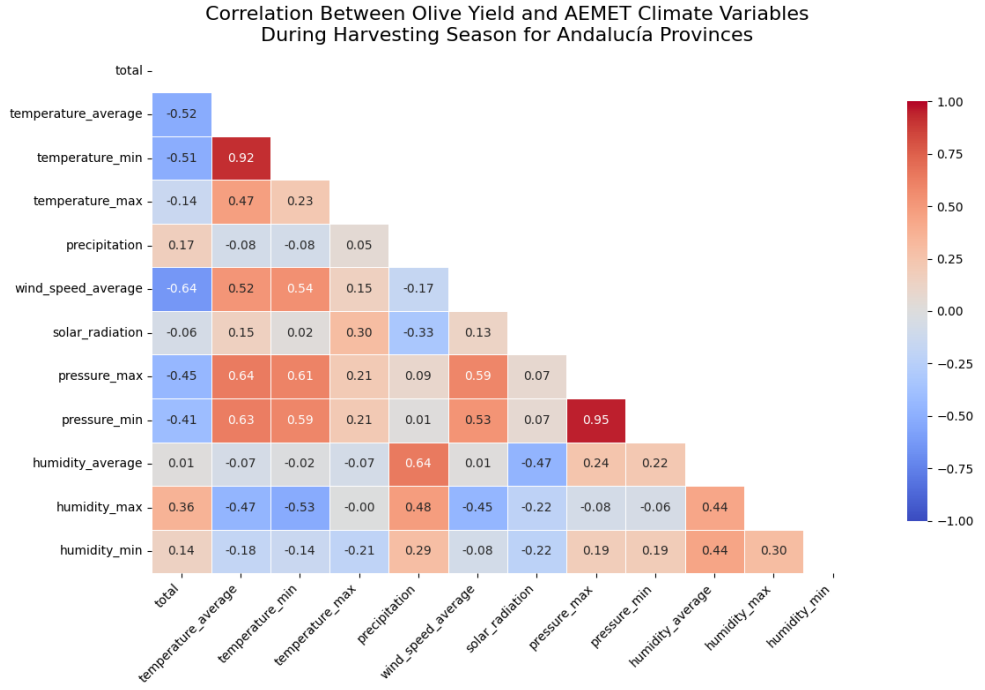


Figure 3.11: Correlation Between Olive Yield and AEMET Climate Variables During Harvesting Season in Andalucía Provinces

This granularity is essential to understanding how intra-annual climate dynamics influence agricultural outcomes.

3.4.1.2 Data Splitting

The dataset spans the years 1998 to 2022. To ensure robust model development and evaluation, the data was split into training and held-out test sets:

- **Training Set:** 80% of the data, consisting of the years 1998 to 2017, was used for model development. This training set was further split internally into training and validation subsets during model development to tune hyperparameters and prevent overfitting.
- **Held-Out Test Set:** The remaining 20%, covering the years 2018 to 2022, was reserved as a held-out test set. This set was not used during model training or hyperparameter tuning and was solely utilized for final model evaluation. This ensures that the test set provides an unbiased assessment of the model's generalization performance on unseen data.

This time-based split simulates a real-world scenario where past data is used to predict future outcomes, while ensuring that the final evaluation does not suffer from data leakage.

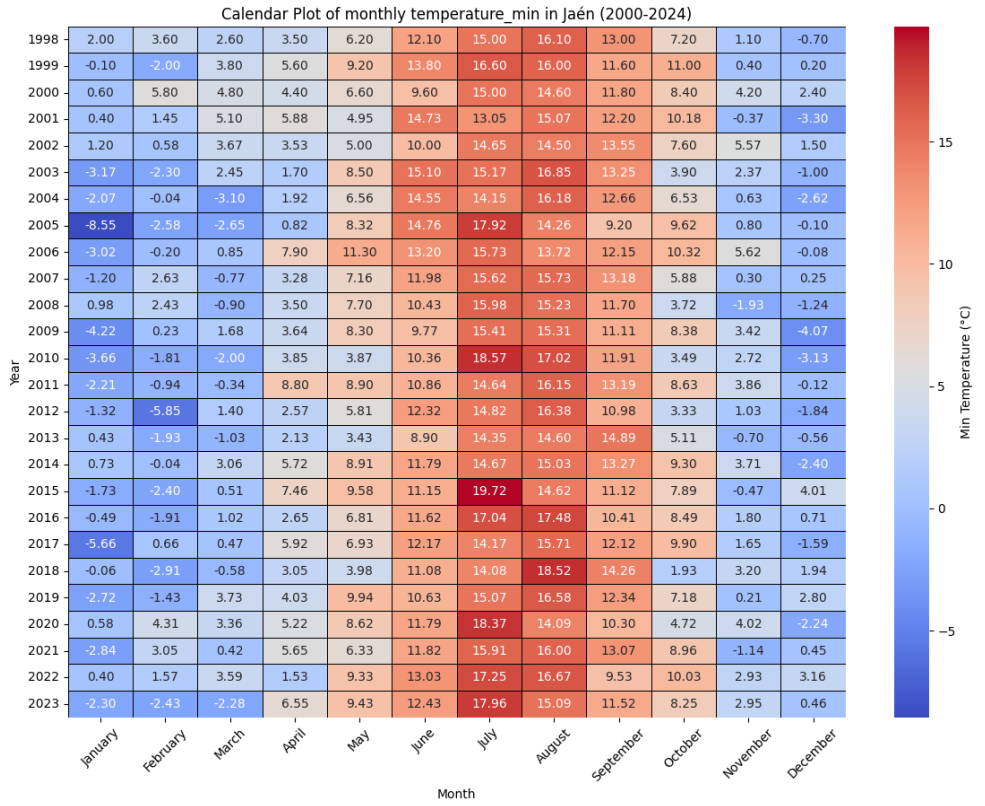


Figure 3.12: Calendar Plot of Monthly Minimum Temperature in Jaén (2000-2024)

3.4.2 Data Sources

3.4.2.1 Olive Yield Data

The olive yield data was provided in Excel format with the following columns:

- **province_name:** The name of the province.
- **total:** The total olive yield.
- **olive_oil:** The yield specific to olive oil production.
- **table_olive:** The yield specific to table olives.

In the original dataset, rows containing aggregated yield values for each autonomous community were included after the province-level data. To ensure the dataset contained only province-level information, these aggregated rows were removed. This step maintained the appropriate granularity, ensuring the model focuses on individual provinces.

3.4.2.2 TerraClimate Data

The TerraClimate dataset provides key climate variables relevant to olive yield. An additional feature, the **water balance**, was engineered to capture the difference between

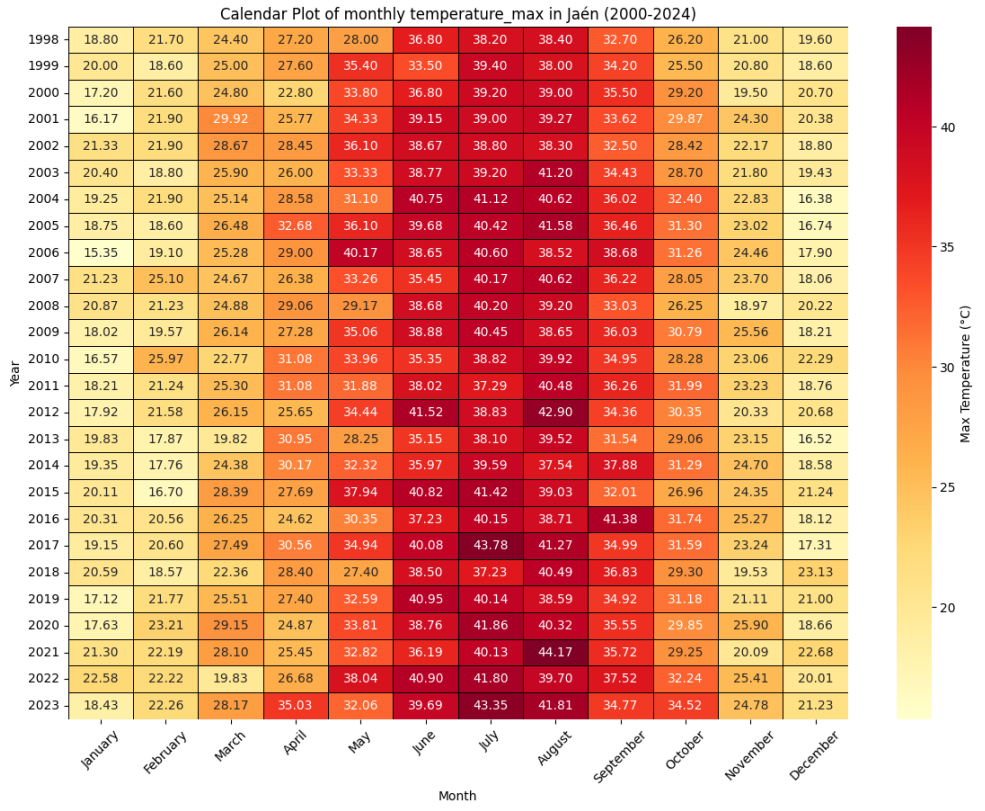


Figure 3.13: Calendar Plot of Monthly Maximum Temperature in Jaén (2000-2024)

precipitation and potential evapotranspiration. This variable reflects moisture availability, which is a critical factor in agricultural productivity, particularly in regions prone to drought. Including this feature improves the model's ability to predict olive yield by directly accounting for the impact of climate conditions on crops.

3.4.2.3 AEMET Data

The AEMET dataset includes daily meteorological data recorded at multiple stations within each province. To align this data with the model's structure, the following steps were taken:

- **Data Aggregation:** Daily data was aggregated to a monthly frequency for each province. This was necessary to create a dataset where each province-year pair could be analyzed based on monthly climatic variables.
- **Aggregation Methods:** The following aggregation methods were applied across all stations within each province:
  - **temperature_average:** Mean of daily average temperatures.
  - **temperature_min:** Minimum of daily minimum temperatures.
  - **temperature_max:** Maximum of daily maximum temperatures.
  - **precipitation:** Sum of daily precipitation.

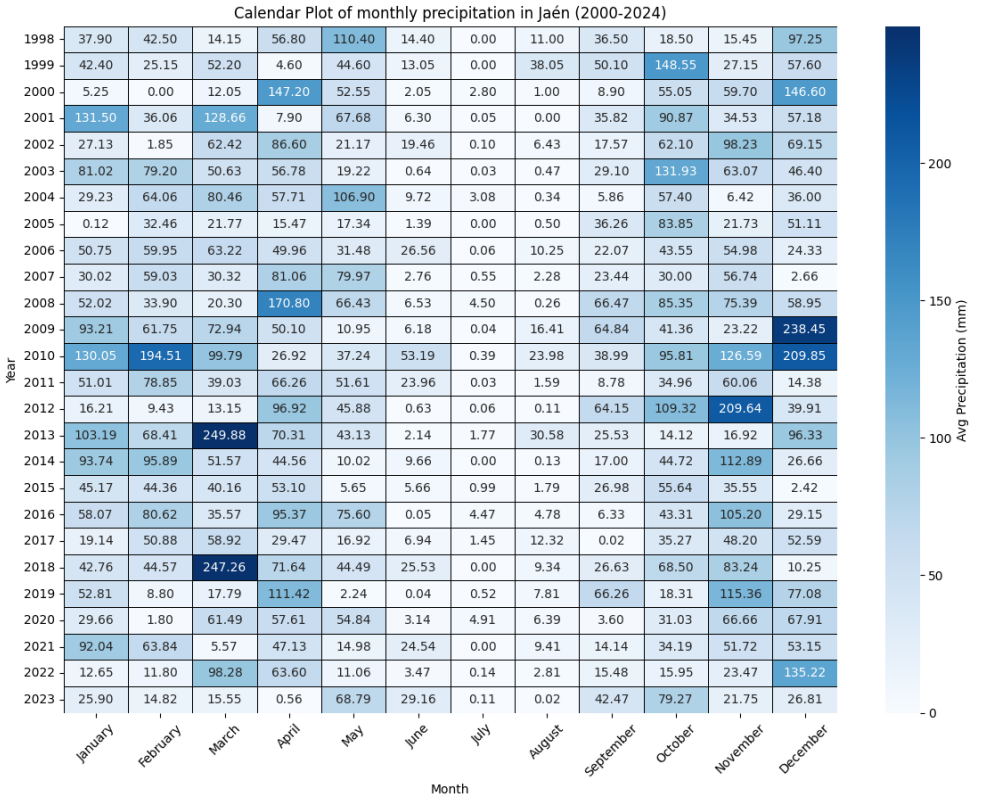


Figure 3.14: Calendar Plot of Monthly Precipitation in Jaén (2000-2024)

- **wind_speed_average:** Mean of daily wind speeds.
- **solar_radiation:** Mean of daily solar radiation.
- **pressure_max:** Maximum of daily maximum pressures.
- **pressure_min:** Minimum of daily minimum pressures.
- **humidity_average:** Mean of daily average humidity.
- **humidity_max:** Maximum of daily maximum humidity.
- **humidity_min:** Minimum of daily minimum humidity.

This aggregation process ensured that the data was compatible with the model's input structure, which relies on monthly climatic variables to predict annual yield outcomes.

3.4.2.4 Vegetation Indices

The vegetation data, derived from the MODIS/061/MOD13Q1 dataset, was initially available with a 16-day temporal frequency. To maintain consistency with the other datasets, the vegetation indices were aggregated to a monthly frequency using mean aggregation. This ensures that vegetation dynamics are captured in a format compatible with the model, enabling it to reflect the influence of vegetative conditions on olive yield.

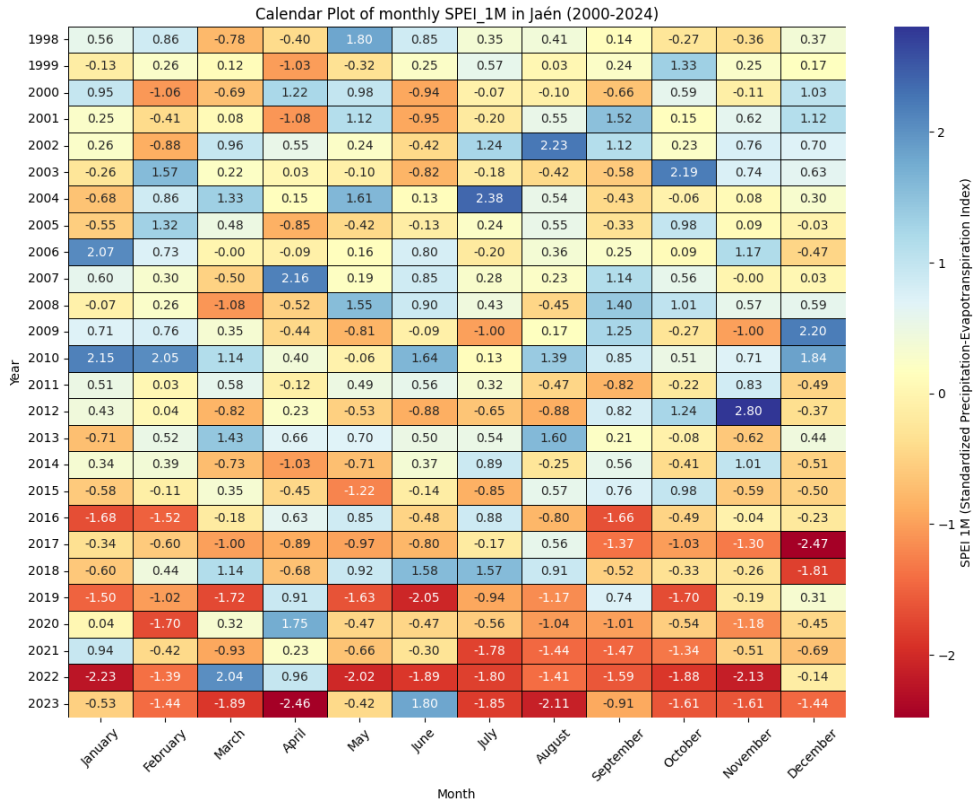


Figure 3.15: Calendar Plot of Monthly SPEI_1M in Jaén (2000-2024)

3.4.3 Data Pre-processing

3.4.3.1 Missing Data Imputation

After splitting the data into training and test sets, missing values in the TerraClimate data (and similarly for other datasets, if applicable) were imputed. The imputation model was fitted on the training data to avoid data leakage and then applied to the test data. This approach ensures that the integrity of the predictive modeling process is preserved, preventing any information from the test set from influencing the model during training.

3.4.3.2 Data Transformation (Pivoting)

To facilitate the modeling process, the data was transformed from a long format, where each row represented a month within a year, into a tabular format where each row corresponds to a province-year pair. The transformation followed this structure:

- **Original Structure:**

  province | date (year and month) | var_1 | var_2 | ... | var_n

- **Transformed Structure:**

  province | year | var_1_jan | var_1_feb | ... | var_1_dec | ... | var_n_dec

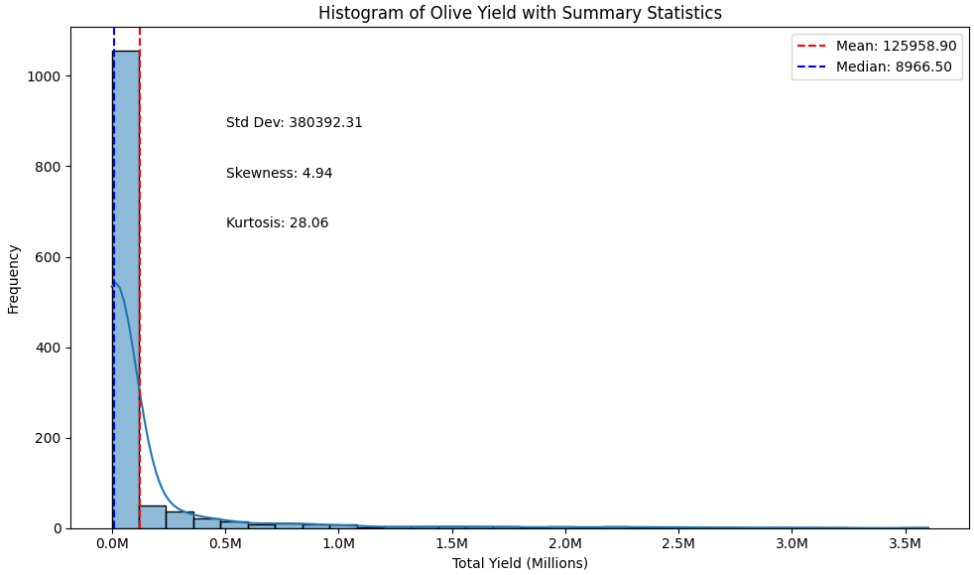


Figure 3.16: Histogram of Olive Yield Distribution with Summary Statistics

This pivoting process was crucial for structuring the data so that the model could use all monthly variables for a given year as features. This setup allows the model to capture the full temporal dynamics within a year, providing the necessary input to predict the annual olive yield.

3.5 Regression Analysis

This section describes the regression analysis methodology used to explore the relationships between various climatic, meteorological, and vegetation health variables and olive yield. The primary goal is to identify which factors most strongly influence olive yield and understand their impact over time.

To achieve this, two approaches were followed:

1. **Direct Feature Regression Analysis:** This approach uses raw monthly features to explore which individual variables are most strongly associated with olive yield.

2. **Window-Summarized Feature Regression Analysis:** This second approach focuses on a subset of promising variables identified from the first approach and generates window-summarized features to explore whether temporal patterns over longer periods, such as multi-year trends, provide deeper insights into how these factors influence olive yield.

Both approaches follow the same data preparation, cross-validation, model selection, and evaluation processes, but differ in the features used, allowing us to explore both immediate and cumulative effects.

3.5.1 Rationale for Two Approaches

The two approaches—direct feature regression and window-summarized feature regression—were chosen to explore different ways of understanding the relationship between olive yield and environmental factors:

- **Direct Feature Regression:** This approach helps identify which variables (e.g., temperature, precipitation, vegetation indices) are most correlated with olive yield on a month-by-month basis. The goal here is to evaluate the direct, short-term effects of these factors.
- **Window-Summarized Feature Regression:** After identifying promising variables from the direct feature regression, this approach generates window-summarized features (e.g., averages or rolling statistics over 3 or more years). This allows us to explore whether longer-term cumulative patterns—such as multi-year temperature averages or extreme conditions—have a more significant impact on olive yield. This is particularly useful for identifying lag effects or understanding whether sustained trends (e.g., droughts, warming) influence the yield.

3.5.2 Approach 1: Direct Feature Regression Analysis

The first approach uses the raw monthly data to understand the short-term relationships between climatic, meteorological, and vegetation health variables and olive yield. This provides a baseline understanding of which factors appear to have the strongest influence.

3.5.2.1 Data Splitting and Cross-Validation

Given the time-dependent nature of the data, we employed an *Expanding Window Splitter* for cross-validation. This method ensures that past data is always used to predict future data, preserving the temporal structure. For instance, the first split uses the years 1998–2009 as training data and tests on the years 2010–2012. The window expands in subsequent splits, as follows:

- **First Split:** Train on 1998–2009, test on 2010–2012.
- **Second Split:** Train on 1998–2010, test on 2011–2013.
- **Third Split:** Train on 1998–2011, test on 2012–2014.

This method ensures that each model is trained on a growing amount of past data, preventing data leakage and simulating a realistic forecast-like scenario.

3.5.2.2 Model Selection and Hyperparameter Tuning

A variety of regression models were used to assess the influence of each variable, including:

- **Linear Regression**: Provides a baseline for understanding linear relationships.
- **Ridge Regression**: Adds L2 regularization to limit the influence of less significant features and prevent overfitting.
- **Random Forest Regressor**: Captures non-linear relationships between variables.
- **Gradient Boosting Regressor**: Sequentially improves model accuracy by focusing on errors.
- **XGBoost Regressor**: A highly efficient form of gradient boosting.
- **Support Vector Regressor (SVR)**: Captures non-linear relationships but can be computationally expensive.

Because of the large number of features (monthly values for multiple variables), performing a full *Grid Search* for hyperparameter tuning was infeasible. Instead, *Randomized Search* was used, which efficiently samples from the hyperparameter space without requiring exhaustive testing of all combinations.

3.5.2.3 Evaluation Metrics

The following metrics were used to evaluate the models:

- **Mean Absolute Error (MAE)**: The average absolute difference between the predicted and actual yield values.
- **Mean Squared Error (MSE)**: Squares the errors, penalizing larger deviations more heavily.
- **Root Mean Squared Error (RMSE)**: The square root of MSE, making the error more interpretable in the context of the data.
- **R^2 Score**: Indicates how much of the variation in olive yield is explained by the model, with a score of 1 representing perfect explanatory power.

3.5.2.4 Insights from Approach 1

This analysis provides insights into which specific variables are most closely associated with olive yield. The strongest predictors—those variables that consistently have the highest explanatory power—are identified from the TerraClimate, AEMET, and vegetation health indices datasets. These findings form the basis for the second approach, where window-summarized features are created to further explore the relationships.

3.5.3 Approach 2: Window-Summarized Features Regression Analysis

The second approach focuses on generating window-summarized features for a selection of promising variables identified from the first approach. This method is used to explore whether temporal patterns—such as sustained trends or extreme conditions over several months or years—have a more significant impact on olive yield than individual monthly values.

3.5.3.1 Window Feature Generation

For this approach, window-based summary statistics are generated for selected variables. For example, rolling averages over 3 or more years are calculated for variables like temperature, precipitation, and vegetation health indices. The idea is to capture cumulative effects or longer-term trends that might influence yield outcomes. The following window-based statistics are generated:

- **Rolling Mean**: Captures the average value over a specified window (e.g., 3 years).
- **Rolling Min/Max**: Measures extreme values within the window, providing insights into droughts or extreme weather events.
- **Standard Deviation**: Measures variability over the time window, indicating how consistent or volatile a variable is over time.

By focusing on only the most promising features, we limit the number of window-summarized features to those that show the most potential based on the first approach.

3.5.3.2 Model Selection, Cross-Validation, and Hyperparameter Tuning

The same set of models (Linear Regression, Ridge, Random Forest, etc.) is used for this approach, following the same process for cross-validation and hyperparameter tuning. The *Expanding Window Splitter* is again employed to ensure the temporal structure is respected, and *Randomized Search* is used to optimize hyperparameters.

3.5.3.3 Evaluation Metrics

As with the first approach, the models are evaluated using the same set of metrics: MAE, MSE, RMSE, and R^2. This consistency in evaluation allows for a direct comparison between the direct feature regression and the window-summarized feature regression.

3.5.3.4 Exploration of Temporal Effects

The purpose of this second approach is to investigate whether longer-term temporal patterns (such as multi-year temperature trends or prolonged periods of low precipitation) provide deeper insights into how environmental variables influence olive yield. By summarizing

variables over windows, we explore whether sustained climatic conditions have stronger predictive power than individual monthly values.

3.5.4 Conclusion

By employing both the direct feature regression and window-summarized feature regression approaches, we aim to gain a comprehensive understanding of the relationships between environmental variables and olive yield. The first approach identifies the strongest individual monthly predictors, while the second approach explores whether cumulative or long-term trends enhance our understanding of these relationships. Together, these approaches provide a holistic view of the factors that most influence olive yield.

4 Results

4.1 Results

This section presents a detailed analysis of the machine learning models used to uncover the relationship between climate variables and olive yield in Spain, with a focus on the province of *Jaén*. The results are analyzed across several dimensions, including feature importance, error plots, and time series comparisons between actual and predicted yields. The primary goal is to explore how climate data, especially from *AEMET* (station-based data), impacts olive yield, with a secondary focus on prediction accuracy.

4.1.1 Model Evaluation

Several models were tested across different datasets, including *AEMET* (station-based climate data), *TerraClimate* (satellite-based gridded data), and window features that aggregate climate variables over specific periods. *AEMET*-based models showed superior predictive performance compared to *TerraClimate*, highlighting the importance of local, station-based data in capturing the specific climate conditions that influence olive yield.

Model	Dataset	R^2 Score	Correlation	MAE	MSE	RMSE
XGBoostRegressor	window-features	0.89	0.95	50036.93	1.84e+10	135661.1
GradientBoostingRegressor	AEMET	0.64	0.81	113798.2	6.08e+10	246489.0
RandomForestRegressor	AEMET	0.63	0.80	117331.2	6.21e+10	249278.2
XGBoostRegressor	AEMET	0.63	0.82	114430.3	6.31e+10	251141.5
Ridge	AEMET	0.47	0.70	187469.3	8.95e+10	299082.8
XGBoostRegressor	TerraClimate	0.46	0.71	140455.4	9.61e+10	309974.0

Table 4.1: Summary of model performance across datasets. AEMET-based models outperform TerraClimate-based models, with XGBoost trained on window features performing the best.

4.1.2 Feature Importance Analysis

Feature importance provides critical insights into how climate variables influence olive yield. Figures 4.1 through 4.6 present the feature importance for the six best-performing models trained on different datasets. A key finding is that *AEMET* data consistently leads to better performance, while *TerraClimate* models underperform, likely due to the generalized nature of satellite-based data.

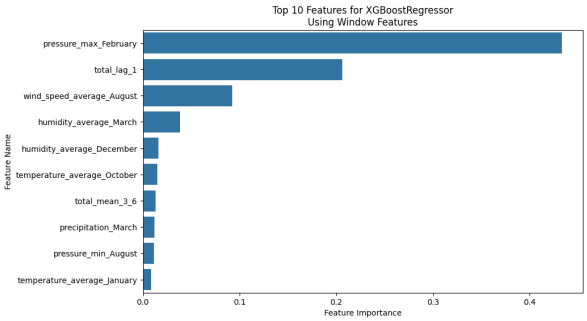


FIGURE 4.1: Feature importance for XGBoost trained on window features.

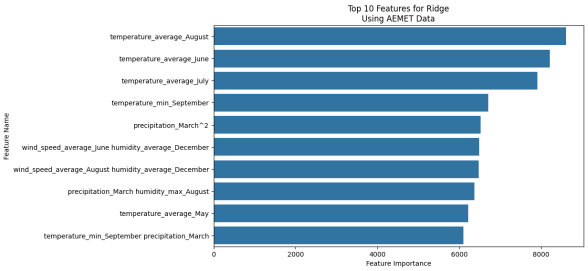


FIGURE 4.2: Feature importance for Ridge trained on AEMET data.

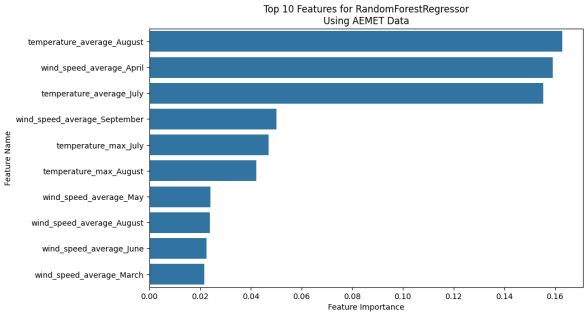


FIGURE 4.3: Feature importance for RandomForest trained on AEMET data.

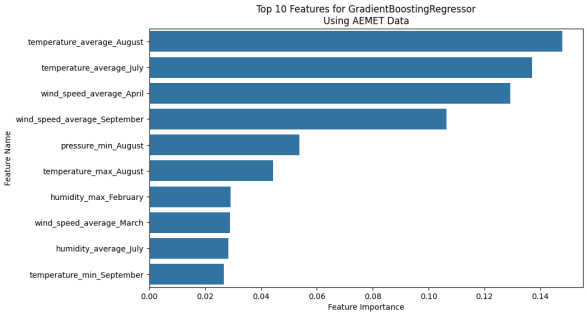


FIGURE 4.4: Feature importance for GradientBoosting trained on AEMET data.

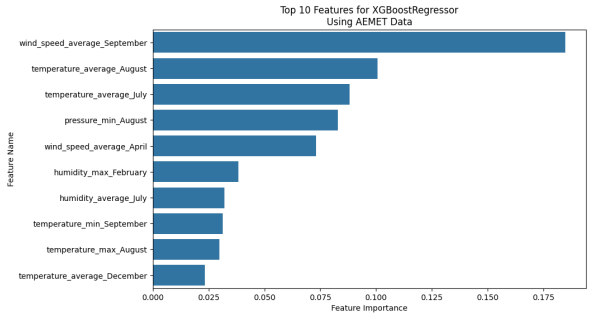


FIGURE 4.5: Feature importance for XGBoost trained on AEMET data.

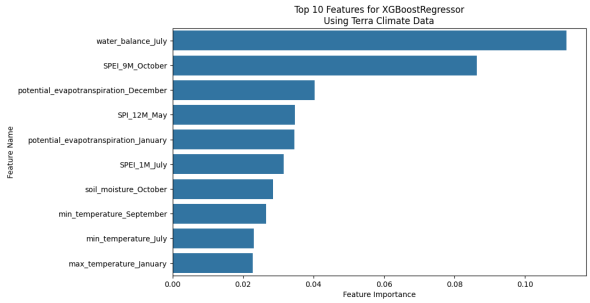


FIGURE 4.6: Feature importance for XGBoost trained on TerraClimate data.

4.1.3 PREDICTION ERROR DISPLAY

The prediction error plots (Figures 4.7 through 4.9) provide insights into how well the models generalize across different yield ranges. In Figure 4.7, the XGBoost model trained on window features shows tight residual clustering for moderate yield values but scatters more for extreme values, suggesting underperformance in high-yield cases.

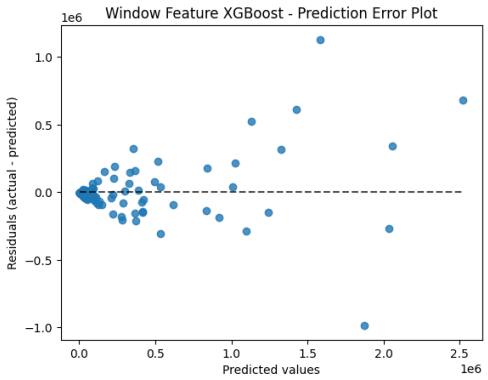


FIGURE 4.7: Prediction error plot for XGBoost trained on window features.

4.1.4 PREDICTED VS. ACTUAL OLIVE YIELD (TIME SERIES PLOT)

The time series plots in Figures 4.10 through 4.13 compare actual and predicted olive yields for *Jaén*. The XGBoost model trained on window features (Figure 4.10) closely aligns with

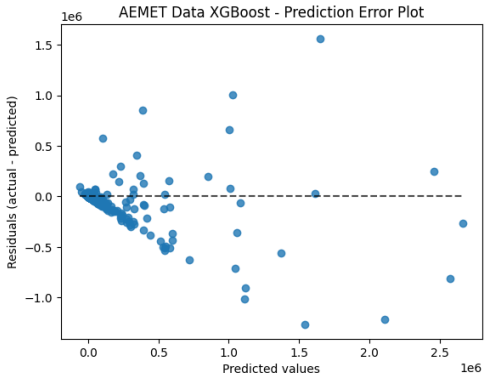


Figure 4.8: Prediction error plot for XGBoost trained on AEMET data.

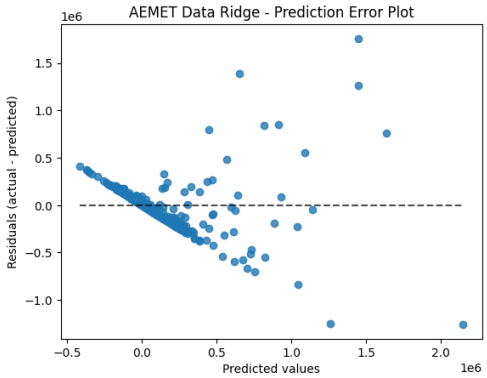


Figure 4.9: Prediction error plot for Ridge trained on AEMET data.

the actual values during the training period but underpredicts in extreme conditions during the test period (2016-2022). This suggests that while the auto-regressive features help in predicting average yields, the model underestimates sharp fluctuations caused by extreme weather conditions.

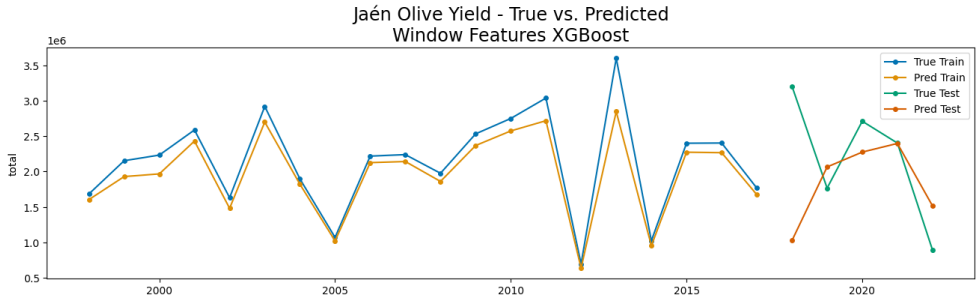


Figure 4.10: Time series plot of actual vs. predicted olive yield for Jaén using XGBoost trained on window features.

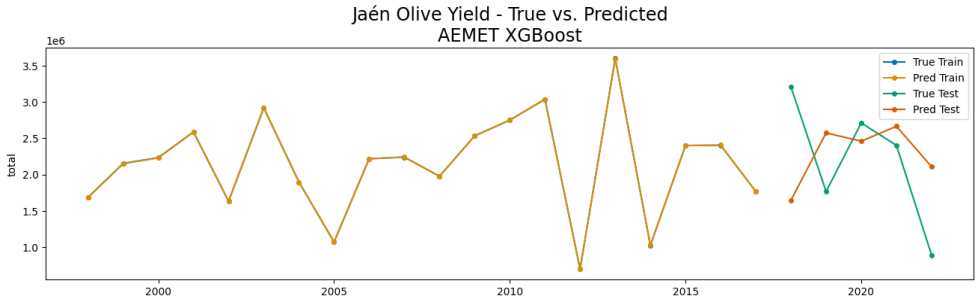


FIGURE 4.11: Time series plot of actual vs. predicted olive yield for Jaén using XGBoost trained on AEMET data.

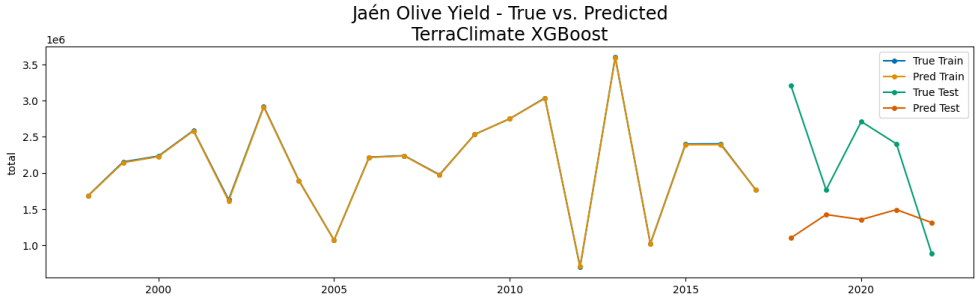


FIGURE 4.12: Time series plot of actual vs. predicted olive yield for Jaén using XGBoost trained on TerraClimate data.

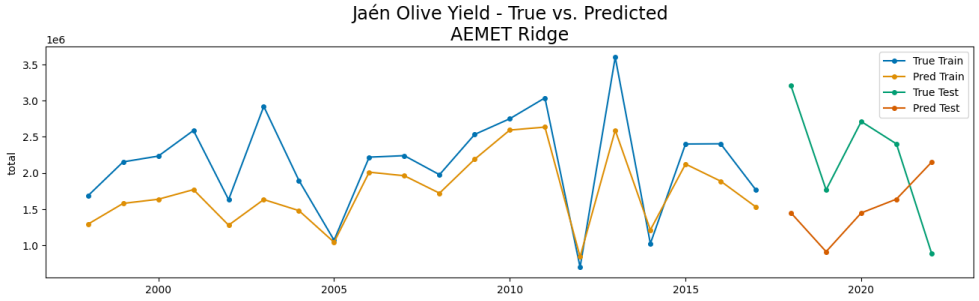


FIGURE 4.13: Time series plot of actual vs. predicted olive yield for Jaén using Ridge trained on AEMET data.

5 Discussion

5.1 Discussion

This study aimed to explore the relationship between climate variables and olive yield in Spain, using several machine learning models across different datasets, such as AEMET (station-based data) and TerraClimate (satellite-based data). While the results provided valuable insights, there are several limitations and considerations that need to be addressed, as well as areas for future improvement.

5.1.1 Limitations of the Current Methods

One key limitation of the current methodology is the use of monthly climate data. Olive yield is highly sensitive to specific climate conditions during key phenological stages, such as flowering, fruit setting, and ripening. By aggregating the data at a monthly level, much of the day-to-day variability that could affect yield is lost. For instance, short-term heatwaves, unseasonal frosts, or extreme weather events that occur within a month could have significant impacts on the yield but may not be fully captured by monthly averages. A more granular approach, such as using daily climate data, would provide a clearer picture of the microclimatic conditions that influence olive yield.

Additionally, olive trees are perennial crops, and their yield is influenced by long-term factors such as tree age, health, and accumulated stress from previous years. The models used in this study, which assume yield is largely independent from year to year, do not account for such carryover effects. Although the auto-regressive nature of yield (captured by the feature *total_lag_1*) reflects some persistence in yield, it is essential to acknowledge that the climate conditions in one year can impact yields in subsequent years through complex biological mechanisms. More advanced models that incorporate long-term tree health data and climate history could provide better yield predictions.

5.1.2 Opportunities for Improvement

Incorporating daily climate data and more granular information would likely improve model accuracy. Complex temporal models such as Long Short-Term Memory (LSTM) networks could be beneficial for modeling the sequential nature of climate events and their impacts on olive yield. LSTMs can handle time-series data with varying lags between cause (climate conditions) and effect (yield), which is particularly relevant for perennial

crops like olives. By working with daily data, future efforts could capture more of the short-term variations in weather patterns, which are not evident in monthly summaries but may significantly affect the crop's phenology and yield.

Furthermore, introducing variables such as tree age, soil conditions, irrigation practices, and pest or disease incidence would help to build a more comprehensive model of olive productivity. These factors, often omitted in freely available climate datasets, play a critical role in determining olive yield and should be included in future studies to improve yield prediction.

5.1.3 Challenges in Quantifying Olive Yield Dynamics

Olive yield dynamics are complex due to the perennial nature of the crop. Unlike annual crops, where yield is highly dependent on the current season's conditions, olive trees exhibit biannual bearing and yield patterns that fluctuate due to factors accumulated over multiple years. This introduces a significant challenge when attempting to model yield based only on climate data from the same year. The assumption that yield is solely influenced by current-year conditions oversimplifies the biological processes involved.

This study's models, particularly those using window features and lagged yield values, reveal some persistence in olive yield across consecutive years. However, these models still operate under the assumption that climate conditions in the same year primarily drive yield. To accurately capture the long-term dynamics of olive yield, future studies should explore how previous years' climate and yield data, as well as tree age and health data, influence productivity in subsequent seasons.

5.1.4 Key Observations and the Role of AEMET Data

Despite these limitations, this study has shown that freely available AEMET station-based data provides a valuable resource for monitoring and predicting olive crop health and productivity. The models trained on AEMET data outperformed those based on TerraClimate data, highlighting the importance of using local, station-specific climate data to capture the variability in environmental conditions that affect olive yield.

By quantifying the relationship between climate variables and yield, this analysis demonstrates the potential of using AEMET data to track climate impacts on agricultural productivity at a regional scale. With further improvements, such as incorporating more granular data and more advanced models, AEMET data could be a powerful tool for farmers and policymakers to monitor and predict crop health, plan for irrigation needs, and anticipate the effects of climate change on agriculture.

5.1.5 Conclusion

In summary, this study has taken important steps in modeling the relationship between climate and olive yield using machine learning models, particularly in the province of Jaén. However, more work remains to fully capture the complexities of olive production.

Future efforts should focus on incorporating more comprehensive datasets, using higher temporal resolutions, and considering long-term biological factors such as tree health and age. Moreover, advanced modeling techniques like LSTMs offer promising avenues for future research to better understand the intricate dynamics of climate and yield in perennial crops like olives.

6 Project Management

6.1 Project Management

The dissertation project was managed using a combination of formal and informal methods to ensure steady progress and the reproducibility of the work. GitHub was primarily used as a repository for the project, ensuring that the code and data were safely stored and could be reproduced, but it was not utilized for versioning or collaboration. This provided a backup mechanism to avoid loss of work and maintain reproducibility, which was especially important as the project evolved and new datasets were introduced.

Weekly meetings with the supervisor provided regular progress updates and a platform for discussing challenges. These meetings played a crucial role in keeping the project on track, allowing for necessary adjustments to be made as new challenges and insights emerged.

Over the course of the project, several unexpected shifts in focus occurred. Initially, the primary focus was on using drought indices to predict olive yield, but it soon became evident that drought data alone were insufficient to capture the complexity of the yield dynamics. As a result, the focus shifted towards incorporating broader climate data, which led to the acquisition of additional datasets from sources like AEMET station data and TerraClimate. This evolution in scope introduced unforeseen challenges, such as the need for new data processing techniques and feature engineering methods.

These shifts made it difficult to anticipate future steps and plan accordingly. However, decisions were made throughout the project to ensure the final deliverables were sound and based on rigorous analysis. For instance, although initial experiments included vegetation indices, their limited predictive power led to their exclusion from the final model suite. The focus was redirected to ensuring the robustness of the analyses and models that were ultimately chosen.

In conclusion, while the project faced continuous challenges and shifting objectives, the strategies employed, such as maintaining a reproducible codebase on GitHub and adapting to new insights through regular meetings with the supervisor, were essential for managing the project. Despite the limited time and evolving scope, the project was able to deliver accurate and reliable results by focusing on what was achievable within the constraints.

Bibliography

1. Arora, N. K. Impact of climate change on agriculture production and its sustainable solutions. *Environmental sustainability* **2,** 95–96 (2019).
2. Fraga, H., Pinto, J. G., Viola, F., Santos, J. A., *et al.* Climate change projections for olive yields in the Mediterranean Basin. *International Journal of Climatology* **40,** 769–781 (2020).
3. Seager, R. *et al.* Climate variability and change of Mediterranean-type climates. *Journal of Climate* **32,** 2887–2915 (2019).
4. Alrteimei, H. A., Ash'aari, Z. H. & Muharram, F. M. Last decade assessment of the impacts of regional climate change on crop yield variations in the Mediterranean region. *Agriculture* **12,** 1787 (2022).
5. Shah, F. & Wu, W. Soil and crop management strategies to ensure higher crop productivity within sustainable environments. *Sustainability* **11,** 1485 (2019).
6. BBC News. *Olive oil price skyrockets as Spanish drought bites* https://www.bbc.co.uk/news/world-europe-67565503. Accessed: 2024-08-19. 2024.
7. The Guardian. *Extra virgin olive oil prices surge as global production drops* https://www.theguardian.com/business/article/2024/may/07/extra-virgin-olive-oil-prices-global-production. Accessed: 2024-08-19. 2024.
8. Delgado-Lista, J. *et al.* Long-term secondary prevention of cardiovascular disease with a Mediterranean diet and a low-fat diet (CORDIOPREV): a randomised controlled trial. *The Lancet* **399,** 1876–1885 (2022).
9. Fernandez-Carrillo, A., Rivas-Gonzalez, F. & Revilla-Romero, B. *Satellite imagery and climate variables suggest variations in the phenology of olive groves in Southern Spain* in *Remote Sensing for Agriculture, Ecosystems, and Hydrology XXIII* **11856** (2021), 118560M.
10. Galán, C., Vázquez, L., García-Mozo, H. & Domínguez, E. Forecasting olive (Olea europaea) crop yield based on pollen emission. *Field Crops Research* **86,** 43–51 (2004).
11. Ortiz-Bobea, A., Knippenberg, E. & Chambers, R. G. Unpacking the climatic drivers of US agricultural yields. *Environmental Research Letters* **14,** 064003 (2019).
12. Galán, C., Vázquez, L., García-Mozo, H. & Domínguez, E. Forecasting olive (Olea europaea) crop yield based on pollen emission. *Field Crops Research* **86,** 43–51 (2004).

13. McKee, T. B., Doesken, N. J., Kleist, J., *et al. The relationship of drought frequency and duration to time scales* in *Proceedings of the 8th Conference on Applied Climatology* **17** (1993), 179–183.

14. Vicente-Serrano, S. M., Beguería, S. & López-Moreno, J. I. A multiscalar drought index sensitive to global warming: the standardized precipitation evapotranspiration index. *Journal of climate* **23,** 1696–1718 (2010).

15. Allen, R. G. Crop evapotranspiration. *FAO irrigation and drainage paper* **56,** 60–64 (1998).

16. Thornthwaite, C. W. An approach toward a rational classification of climate. *Geographical review* **38,** 55–94 (1948).

17. Leng, G. & Hall, J. Crop yield sensitivity of global major agricultural countries to droughts and the projected changes in the future. *Science of the Total Environment* **654,** 811–821 (2019).

18. Khaki, S. & Wang, L. Crop yield prediction using deep neural networks. *Frontiers in Plant Science* **10,** 621 (2019).

19. Beguería, S., Vicente-Serrano, S. M. & Angulo-Martínez, M. A multiscalar global drought dataset: the SPEIbase: a new gridded product for the analysis of drought variability and impacts. *Bulletin of the American Meteorological Society* **91,** 1351–1354 (2010).

20. Abatzoglou, J. T., Dobrowski, S. Z., Parks, S. A. & Hegewisch, K. C. TerraClimate, a high-resolution global dataset of monthly climate and climatic water balance from 1958–2015. *Scientific data* **5,** 1–12 (2018).