Improved Prediction of Olive Crop Yield Using Satellite Imagery

A Case Study in Andalusia (Spain)

M. Isabel Ramos
Dept. Cartographic Engineering, Geodesy and Photogrammetry
University of Jaen
Jaen, Spain
miramos@ujaen.es

Lidia Ortega
Dept. Computer Science
University of Jaen
Jaen, Spain
lidia@ujaen.es

Angel Calle
Researcher hired for project
PREDIC I-GOPO-JA-20-0006
University of Jaen
Jaen, Spain
acalle@ujaen.es

Ruth M. Cordoba
Researcher hired for project
PREDIC I-GOPO-JA-20-0006
University of Jaen
Jaen, Spain
rcortega@ujaen.es

Juan J. Cubillas
Dept. Information and Communication Technologies Applied to Education
International University of La Rioja
Logrono, Spain
juanjose.cubillas@unir.net

***Abstract*—Agriculture is one of the strategic economic and social sectors in many countries. In Spain, agriculture has been and continues to be a fundamental sector on which the development of the rest of the sectors depends to a large extent. However, there are many factors that influence its performance, some are dependent on the farmer and others, such as the effects of climate change, do not depend on this. Therefore, crop yields are not always easy to forecast, especially not at an early stage, i.e. before any investment is made for the new cropping season. A case example is presented here in the olive grove in Andalusia, Spain. The objective is to analyze the most influential predictor variables on an early predictive model and the contribution, in this case, of data from satellite images.**

Keywords-olive crop; predictive modeling; multi-source data; satellite imagery.

I. INTRODUCTION

Agriculture plays an important role worldwide, but in specific areas, such as the Mediterranean basin, it has a special economic and social contribution. The climate of the countries in the Mediterranean area has conditioned the type of agriculture, mainly centered on vines, cereals and olives. However, there is currently growing concern about the lack of rainfall and abnormally high temperatures, which are determining factors in the fall in agricultural production, especially in the olive sector.

The productivity of olive crops depends on various factors such as soil fertility, weather conditions, the type of tillage carried out and also the amount of last year's crop [1]–[4]. Some of these factors are controllable by the farmer, but others, such as weather conditions, are not.

High olive harvest values are a key objective not only for growers but also for olive mills, distribution and insurance companies, as well as for the food industry. In this sense, it is reasonable to think that strategic decisions in this sector are made on the basis of the expected harvest. Thus, an early season of crop yield prediction, i.e. before planning the investment of resources, is essential in this sector. In the particular case of olive production, it makes it possible to establish market strategies, to know in advance the income of farmers and companies, the number of workers to be hired, or the machinery that will be needed. This information would make it possible to adjust production costs and contribute to environmental sustainability by optimizing the use of phytosanitary products or tillage.

In this sense, a few years ago, crop predictions relied heavily on the experience of farmers, but today much agricultural research focuses on improving crop forecasting. This shift has been complemented by other economic sectors using the predictive potential of Artificial Intelligence (AI) techniques, in particular Machine Learning (ML) and Deep Learning (DL). These techniques are now widely used for crop yield prediction [5]–[8]. However, crop yield prediction is one of the challenging problems in precision agriculture and, as indicated by Xu et al. [9] it is not a trivial task. There are several studies in olive crop yield prediction and most of them are based on the predictive value of pollen emission levels [10]–[14]. These studies take into account basic parameters of pollen levels, as well as other factors such as temperature, rainfall and relative humidity. In the latest study, a regression analysis is performed with results with an error of 0.96% in July. It should be borne in mind that in Spain the harvest is normally harvested between December and January. Therefore, in spite of the latter being a very reliable model the prediction is made after pollination, which occurs very late in the agricultural year, in April, May or June. It is therefore a very reliable model, but it provides a very late prediction, only 5 months before the olive harvest. These results provide useful information for the farmer, to help establish optimal harvesting periods. However, the purpose of our prediction is not that, but to help plan at the beginning of the crop year the investments and resources to be used for ploughing that year.

The importance of early crop prediction is a hot topic in the scientific literature [15], [16]. In the particular case of the olive

 ISBN: 978-1-68558-168-8

grove, most work focuses on the study of variables such as the amount of pollen in the air, vegetative indices that indicate the health of the plant or rainfall maps [11], [12], [17]. However, some of these values are collected in spring, between April and May. This forecasting period is too late to make strategic decisions of the type discussed above, as the investment of resources has already been made at that time. The challenge in this sector is an early forecast in the months of February, just when the previous season's olive harvest has finished.

The aim of this work is to generate an early predictive model of the harvest quantity of olives, our target. For this purpose, ML algorithms are applied using variables from different sources. Although not all the factors that will affect the next harvest are known at this time, it would be advisable to anticipate what the general trend will be if other negative factors do not intervene. In any case, this prediction model will be fed with different variables as the seasons progress in order to better adjust to the time of harvest. The current state of the work is in the analysis of the predictor variables used. At this phase the aim is to determine which are the most influential and to discard those that do not have a significant influence or introduce noise. In this sense, it is observed that the integration of variables from satellite imaginary improves the predictive error so far.

The document is structured as follows. Section 2 describes the methodology carried out and the data processing carried out using Google Earth Engine and QGIS together, as well as the ML algorithms used. Section 3 shows the first results with the different combinations of predictor variables, analyzing how the results improve with the incorporation of satellite image data. Section 4 describes the conclusions and future work to be developed.

II. METHODOLOGY

This section describes each of the phases comprising the workflow followed. The objective in this phase of the state of the research is to analyze how the predictor variables used influence the level of accuracy of the predictive model.

A. Research Area

The study area covers all the municipalities of the province of Jaen, located in Andalusia, southern Spain, between co-´ ordinates 38°N 3°W (WGS84). This province extends over a total area of 13,496 km², characterized by diverse topographies ranging from mountainous areas to wide valleys. In total, it has 97 municipalities (see Figure 1). The climate is Mediterranean, with partial variations depending on the configuration of the terrain and the proximity to the Guadalquivir River, which flows through the province. The average annual temperature varies between 15-17 °C, with marked differences in temperature between day and night, especially in areas close to the river. Rainfall, around 500 mm, is mainly concentrated in the coldest period of the year, although this pattern varies according to the geography of each area.

B. Dataset

At the core of any predictive analysis is reliable data to ensure an accurate prediction. Here, the data comes from official sources and entities, which supports its reliability. Examples of these data sources are: the Area of the Central Registry of Cartography of the National Geographic Institute of Spain which has provided the municipality boundary files, or the State Meteorological Agency (AEMET), dependent on the Spanish Government, meteorological data have been obtained. To achieve accurate results, some pre-processing is often necessary, such as outlier cleaning, data cleaning, categorization and normalization of information, among other steps.

In general terms, the yield of the olive grove depends on the climate, the quality of the soil, the presence of pests, tillage, olive tree varieties and planting density [1]–[4]. However, this study is conducted at municipal aggregation level, where several of these factors are inherently embedded in the harvest data or the municipality's geographical context, making it unnecessary to explicitly insert them into the dataset. For the research, influential data that do not conform to a recognizable yearly pattern are employed as training data. Therefore, environmental and meteorological variables are considered, as well as variables that indicated the state of the crop during the crucial seasons.

The olive tree follows a vegetative cycle composed of two growth stages, where proper care and feeding are crucial to ensure optimal yields after harvest. Olive cultivation is influenced by a variety of factors, some of which are specific to the local environment, such as environmental and weather conditions. Other factors are a direct result of the agricultural practices employed by the farmer on the farm. Within these stages, olive trees go through various phases, including bud break (February-April), flowering (April-May), fertilization and fruit set (May-June), fruit growth (June-September), fruit ripening (October-December) and dormancy (November-February). In this study, these stages are grouped into three seasons, coinciding with the seasons of the year in which the olive tree responds uniformly to the external factors affecting it. Table 1 shows the months covered by each season considered.

The yield of the olive grove depends on the climate, the quality of the soil, the presence of pests, tillage, olive tree varieties and planting density. This study is conducted at municipal aggregation level, where several of these factors are inherently embedded in the harvest data or the municipality's geographical context, making it unnecessary to explicitly insert them into the dataset. It is obvious that meteorological factors are decisive for crop production. Taking into account the level of aggregation of the crop yield data used in this research, the data are grouped at municipal level of aggregation and the dates to which the data refer are the seasons indicated in Table 1. In short, the predictive variables used are:

• Olive crop yield data from each municipality/year. The training data contain a total of eight years. • Weather information from official weather stations. The Regional Government of Andalusia publishes annual olive crop yield data at different levels of aggregation. In this sense, from weather stations are used rainfall data.

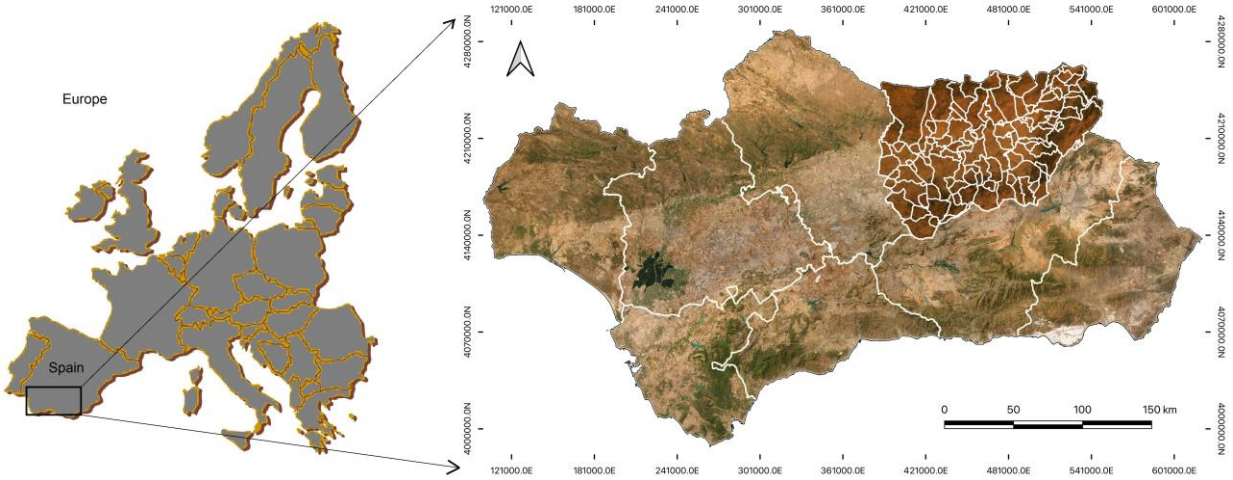

Figure 1. Research area. The left image represents the location of Andalusia in Europe. The right image is the map of Municipalities of Jaen inside Andalusia.´ The coordinates (m) are UTM, zone 30 referred to ETRS89.

TABLE I. SEASON DIVISION

Year	Dates
Winter	01-December to 29-February
Spring	01-March to 30 - June
Summer - Autumn	01-July to 30-November

• Satellite data. Taking into account the information to be obtained, the choice of satellite type becomes a critical decision. Consequently, different satellites are used to acquire data related to different fields, as it is temperature or precipitations.

C. Modeling process

The methodology followed consists of the following phases: (1) The understanding of the data, identify the information it provides, the format and the meaning of the values; (2) Its acquisition and preparation process, Once captured, the data requires prior preparation before being inserted into ML models, some require categorisation, others labelling, detection of outliers, null values, etc.; (3) The generation of the predictive models using different algorithms and, finally, (4) The validation or analysis of the accuracy obtained, in this sense the k-fold cross validation technique was used to evaluate results in statistical analyses. The output data of the model is the amount of olive harvest (the number of Kg of olives). This information is collected for each campaign and for each one of all the province municipalities.

Once the dataset is available, the following step is the detection of data anomalies. This is to identify unusual cases in apparently homogeneous data. A classification algorithm is used for this purpose because these anomalies can be considered as a particular case of classification. Specifically, in this phase, the algorithm used has been the Support Vector Machine (SVM) algorithm. The next phase consists of determining the level of influence of the variables used on the target attribute. In this case, the Minimum Description Length (MDL) algorithm [18]. Finally, regression analysis algorithms are used in the creation of the model. In this study, SVM with Gaussian kernel and Linear kernel, respectively, have been used. SVM regression supports two types of kernels: Gaussian kernel, which is used for non-linear regressions, and the linear kernel, which is suitable for linear regressions. SVMs work effectively on datasets containing many features, even if the number of cases to train the model is very small. Another algorithm used was Neural Networks (NN) which also offers good precision but not bettern than SVM which best fits the prediction in all cases.

III. RESULTS

The first results of this research are the analysis of the contribution of the use of satellite imagery to improve predictive models. We first analyze the results considering only the data from meteorological stations and then we analyze the improvement that the incorporation of satellite images entails.

A. Predictive models using only weather station data

The advantage of using weather stations is that they are daily data from stations maintained by official agencies. However, in this case, the particular disadvantage is the level of aggregation we are using and the distance between stations. Figure 2 shows the network of meteorological stations in Jaen.

Nevertheless, we do this analysis to see to what extent satellite imagery can overcome this drawback Municipalities, which have a station nearby, provide reliable rainfall data.

B. Predictive models using satellite imagery data

Temperature is monitored using MODIS constellation, given that it has spectral bands that allow the detection of surface temperature (see Figure 3). Finally, for rainfall information, we use ECMWF data, based on Copernicus satellite images and Google information. Specifically, ERAS5 (Earth Climate Reanalysis) project.

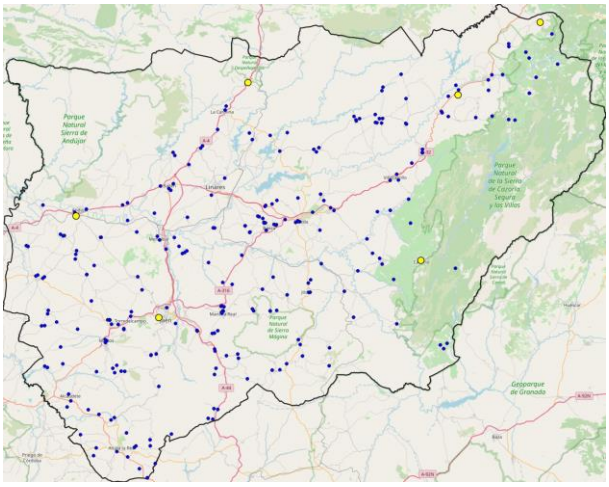

Figure 2. Location of the network of meteorological stations in the province of Jaen (yellow dots) and the oil cooperatives (blue dots).

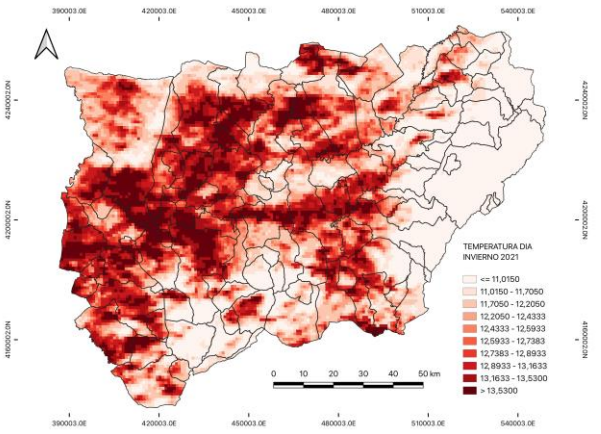

Figure 3. Map of temperatures in the province of Jaen calculated from ECMWF data.

TABLE II. PREDICTION MODELS

Predictor variables	Weather stations			Satellite imagery		
Algorithms	SVM		NN	SVM		NN
	Gaussian	Lineal		Gaussian	Lineal	
Only Rainfall	23.56%	24.29%	25.67%	24.62%	25.70%	27.07%
Rainfall + Temperature	23.63%	24.20%	25.42%	24.58%	25.48%	26.26%

Table II shows the mean absolute errors of the predictive models using different combinations of variables. These first tests confirm that, in general, data from weather stations can be substituted by satellite data in those areas far away from a station.

IV. CONCLUSIONS AND FUTURE WORKS

This study presents a workflow methodology describing the steps followed in the analysis of the predictive calculation of olive crop yield at an early stage. The novelty of the work is how early, within the agricultural year, it makes the crop prediction. Just before the investment of resources begins, when there is still no visible or measurable sign of pollen or the beginning of fruit. The correct selection of the predictor variables and the quality of these variables are fundamental. The integration of satellite imagery into the model improves crop yield prediction. This is due to the fact that a better diagnosis of the state of the satellite imagery can overcome this drawback weather next to the area studied contributes to a good early prediction of its production. In this sense, satellite images have been fundamental in order to have sufficient temporality covering all the municipalities in the province of Jaen. The methodology developed is applicable to future works at different spatial scales, even at local or farm detail level. The short-term future development of this work includes the generation of an intelligent system in which the variables obtained from the satellite images are extracted automatically as well as the downloading of the meteorological values from web services. Thus, a non-expert user will be able to use the system by simply inserting the harvest values of the farm or the area under study. The short-term future development of this work includes the generation of an intelligent system in which the variables obtained from the satellite images are extracted automatically as well as the downloading of the meteorological values from web services. Thus, a non-expert user will be able to use the system by simply inserting the harvest values of the farm or the area under study.

ACKNOWLEDGMENT

This research has been partially funded through the research projects PREDIC I-GOPO-JA-20-0006, cofinanced with the European Union FEDER, and the Junta de Andalucia funds, also the projects PID2021-126339OB-I00 and TED2021132120B-I00, funded by the Spanish Ministry of Science and Innovation.

REFERENCES

[1] C. Silveira, A. Almeida, and A. C. Ribeiro, “How Can a Changing Climate Influence the Productivity of Traditional Olive Orchards? Regression Analysis Applied to a Local Case Study in Portugal,” *Climate*, vol. 11, no. 6, p. 123, Jun. 2023, [Online]. Available: https://www.mdpi.com/2225-1154/11/6/123

[2] P. Deiana et al., “Effect of pedoclimatic variables on analytical and organoleptic characteristics in olive fruit and virgin olive oil,” *European Journal of Agronomy*, vol. 148, 2023.

[3] A. Awan and A. Rab, “Influence of agro-climatic conditions on fruit yield and oil content of olive cultivars,” *Pakistan Journal of Agricultural Sciences*, vol. 51, no. 3, pp. 625–632, 2014.

[4] S. Lavee and M. Wodner, “The effect of yield, harvest time and fruit size on the oil content in fruits of irrigated olive trees (*Olea europaea*), cvs. Barnea and Manzanillo,” *Scientia Horticulturae*, vol. 99, no. 3, pp. 267–277, Feb. 2004. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0304423803001006

[5] S. Archana and P. S. Kumar, “A Survey on Deep Learning Based Crop Yield Prediction,” *Nature environment and pollution technology*, vol. 22, no. 2, pp. 579–592, 2023.

[6] R. Beulah, “A Survey on Different Data Mining Techniques for Crop Yield Prediction,” *International Journal of Computer Sciences and Engineering*, vol. 7, no. 1, pp. 738–744, Jan. 2019.

[7] S. Iniyan, V. A. Varma, and C. T. Naidu, “Crop yield prediction using machine learning techniques,” *Advances in Engineering Software*, vol. 175, p. 103326, 2023.

[8] P. Saini and B. Nagpal, “Empirical Survey Analysis For Crop Yield Prediction & Identification Of Factors Affecting Yield Gaps,” *Journal of Pharmaceutical Negative Results*, vol. 13, pp. 1318–1329, 2022.

[9] X. Xu et al., “Design of an integrated climatic assessment indicator (ICAI) for wheat production: A case study in Jiangsu Province, China,” *Ecological Indicators*, vol. 101, pp. 943–953, Jun. 2019.

[10] P. Filippi et al., “An approach to forecast grain crop yield using multi-layered, multi-farm data sets and machine learning,” *Precision Agriculture*, vol. 20, no. 5, pp. 1015–1029, Oct. 2019 [Online]. Available: https://doi.org/10.1007/s11119-018-09628-4

[11] C. Galán, L. Vázquez, H. García Mozo, and E. Domínguez, “Forecasting olive (Olea europaea) crop yield based on pollen emission,” *Field Crops Research*, vol. 86, no. 1, pp. 43–51, Feb. 2004.

[12] M. Fornaciari, L. Pieroni, F. Orlandi, and B. Romano, “A new approach to consider the pollen variable in forecasting yield models,” *Economic Botany*, vol. 56, no. 1, p. 66, Jan. 2002.

[13] J. Oteros et al., “Better prediction of Mediterranean olive production using´ pollen-based models,” *Agronomy for Sustainable Development*, vol. 34, no. 3, pp. 685–694, 2014.

[14] F. Aguilera and L. Ruiz-Valenzuela, “A new aerobiological indicator to optimize the prediction of the olive crop yield in intensive farming areas of southern Spain,” *Agricultural and Forest Meteorology*, vol. 271, pp. 207–213, Jun. 2019.

[15] L. Ortenzi et al., “Early Estimation of Olive Production from Light Drone Orthophoto, through Canopy Radius,” *DRONES*, vol. 5, no. 4, Dec. 2021.

[16] J. J. Cubillas, M. I. Ramos, J. M. Jurado, and F. R. Feito, “A Machine Learning Model for Early Prediction of Crop Yield, Nested in a Web Application in the Cloud: A Case Study in an Olive Grove in Southern Spain,” *Agriculture*, vol. 12, no. 9, p. 1345, Aug. 2022.

[17] L. Achmakh et al., “Airborne pollen of Olea europaea L. in Tetouan (NW Morocco): heat requirements and forecasts,” *Aerobiologia*, vol. 31, no. 2, pp. 191 – 199, 2015.

[18] P. D. Grunwald, J. I. Myung, and M. A. Pitt, “*Advances in Minimum Description Length: Theory and Applications*, ser. Neural Information Processing series, M. I. Jordan and T. G. Dietterich, Eds. Cambridge, MA, USA: A Bradford Book, Feb. 2005.