Yield estimation using machine learning from satellite imagery

David de la Fuente[1], Elena Rivilla[2], Ana Tena[2], João Vitorino[1], Eva Navascués[2], and Antonio Tabasco[1]

[1]GMV, Remote Sensing and Geospatial Analytics Division, 28760 Tres Cantos, Madrid, Spain
[2]Pago de Carraovejas, R&D Department, 47300 Peñafiel, Valladolid, Spain

Abstract. Accurate and early yield estimation (from pea size) allows 1.- Make decisions at field level: green harvesting, irrigation management. 2.- Advance or organise the purchase of grapes from suppliers. 3.- Forecast the volume of wine produced in the campaign that has not yet begun. 4.- Define the quality of the vintage: regular and detailed monitoring of whether, or not, the heterogeneity of the leaf surface, photosynthetic activity or soil moisture observed in the vineyards is as expected at this time, compared with historical values. 5.- Precise control of each vine in production, knowing which vines are no longer productive or should be grubbed up. The Sentinel-2 satellite has generated a time series of images spanning more than six years, which is a great help in analysing the state of permanent crops such as vineyards, where grapes are produced every year. The weekly comparison of what is happening in the current season with what has happened in the previous six seasons is information that is in line with agricultural practices: Winegrowers make the mental exercise of comparing how the vines are developing today with how they developed in previous seasons, with the aim of repeating the years of good yields. In addition, several commercial satellites can now capture images of 50 centimetres pixel resolution or even better, making it possible to check the health of each vine every year. Since 2020, GMV and Pago de Carraovejas have been working together to develop a yield estimation service based on field information and satellite images that feed machine learning algorithms. This paper describes the path followed from the beginning and the steps taken, summarising as follows: 1. - Machine learning algorithm trained with cluster counting and satellite data. 2. - Adjustment of the number of vines in production in each vineyard using very high-resolution imagery. 3. - Machine learning algorithm trained on real production from past campaigns and historical Sentinel-2 time series. The results obtained by comparing the actual grape intake in the winery with the yield estimation range from 91% accuracy in 2020 to 95% accuracy in 2022.

1 Introduction

Knowing in advance and accurately vineyard yield from the grape's pea size stage is a piece of very valuable information for winegrowers and winemakers: they can control the quality of the vintage according to whether heterogeneity in leaf area, photosynthetic activity or humidity is as expected, knowing which areas are no longer productive, better watering and green pruning management, forecast the volume of wine to be produced and organise in advance the purchase of supplies if necessary.

Traditional methods or direct methods of yield estimation are based on theoretical equations with explanatory variables such as the number of grape bunches, the number of berries per bunch and the average berry weight [1]. They are static estimates based on a multi-stage process, highly dependent on adequate human resources, from counting bunches by visiting control parcels once to managing extensive historical databases of bunch weights and past yields. Bunches number and grape weights vary yearly according to the climatic conditions, the general health status of the vineyards and the agronomic practices such as grubbing up of unproductive vines [2].

Indirect methods are leading alternatives to traditional methods. The number of vines in the vineyard decreases with the age of the vineyard, as the vines can become sick and die. Aerial imagery can update variables such as the fault factor due to the dead vines' grubbing-up process [3]. Likewise, aerial imagery can account for spatial variability within each vineyard, based on well-known vegetation indices such as the Normalized Difference Vegetation Index (NDVI) and the Leaf Area Index (LAI) indices. Both indices are widely used with satellite imagery as well [4]. Since 2017, the Sentinel-2 mission stands out, which has meant a leap in satellite imagery with monitoring every five days at 10-20 m spatial resolution, making it possible to monitor the vegetation health status at any instant in time over large areas. Yield estimation with satellite imagery has focused on the NDVI and LAI indices, but few studies have introduced

indices such as the fraction of Absorbed Photosynthetically Active Radiation (fAPAR), which refers only to the green and alive elements of the canopy or the Disease Water Stress Index (DSWI), which refers to plant water stress [5-6].

In addition, in recent years, artificial intelligence (AI) has been added as a cutting-edge technology in machine learning and data mining for yield estimation [1]. Most of the research is based on local studies based on RGB cameras and computer vision techniques. There are few AI-based studies for yield estimation using satellite imagery plus artificial intelligence algorithms [7]. Here, the authors show the first research to apply deep learning techniques to regional yield estimation using satellite imagery. Sentinel-2 NDVI index time series is the satellite imagery analysed.

This publication presents a methodological approach for yield estimation at the vineyard parcel level using a suitable regressor machine learning algorithm trained with cluster counting from fieldwork reduced to a few vineyards, an aerial image to update grubbed vines and dedicated value-added products generated from time series analysis of Sentinel-2 vegetation indices.

2 Materials and methods

2.1 Study area

The vineyard extends over 200 hectares. 160 hectares are dedicated to the production of top-quality wines. It´s located in the Ribera del Duero Appellation of Origin (VQPRD Vin de Qualité Produit dans une Région Déterminée) region at an average altitude of 850 m, on a slope perpendicular to the Duero River that crosses Peñafiel village from east to west (Fig. 1). The terrain is slightly undulating, flat in the centre of the valley and steep and abrupt as one ascends. The vineyard is governed by the precepts of organic viticulture. Three varieties are grown: mainly Tinto Fino (Tempranillo), together with Cabernet Sauvignon and Merlot. There are different training grapevines adapted to the orography in each case: Cordon Royat, vertical vase and Echalas vase. Yields are limited depending on the cluster-canopy ratio combined with green harvesting. Sensors, soil analysis, meteorological data collection and modelling are fundamental axes for decision-making in the vineyard.

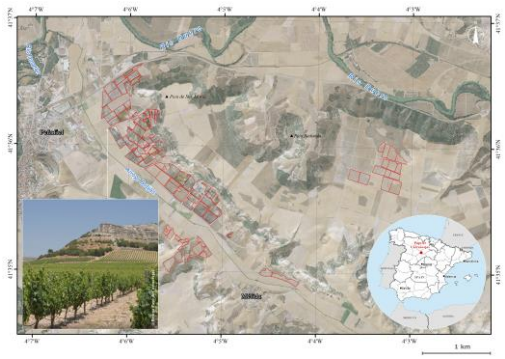

Figure 1. Study area Pago de Carraovejas State winery.

2.2 Manual sampling

Counting bunches requires an investment in staff time and dedication. The process is exhaustive: Bunches are counted every ten rows and every ten vines, excluding the headers. On each vine, larger, intermediate, and smaller bunches are counted. For every certain number of vines, the grapes of a larger, intermediate, and smaller bunch are counted. To calculate the total kilos in a vineyard, the number of vines is multiplied by the average number of bunches classified into larger, intermediate, and smaller bunches, by the average number of berries according to bunch size and by the weight of the berry according to plot and variety history. At least two field visits are made. Once the maturation stage is reached, the weight of the berry from the historical databases is replaced by the real average weight of the bunches. The predictions estimated by this method are very close to reality.

2.3 Aerial imagery

An aerial orthophoto of the Instituto Tecnológico Agrario de Castilla y León (ITACyL) was used here. The flight was carried out in the summer of 2020 at 25 centimeters spatial resolution and spectral resolution in the blue (B), green (G) and red (R) bands. The time of the flight coincides with the veraison stage where the reflectance of the vegetated vines is very different from the reflectance of the bare soil, so it is an optimal imagery for fault factor updating works. Two steps have been carried out.

Firstly, the boundaries of the parcel from the Integrated Administration and Control System (IACS) were refined and adjusted to the extent of the vineyards. This step was made by a photo interpreter. Secondly, an algorithm that finds the best grey thresholds for each aerial band (RGB) was implemented to separate vegetated areas and bare soil. The thresholds vary for each vineyard as the soil conditions. Next, the vegetated areas were increased using an adjusted buffering process in each vineyard according to its planting frame, so that the bared rows between vines lines were included like productive areas. The remaining area is the non-producing area in each vineyard, this being the fault factor. Figure 2 shows the productive area and the fault factor in a vineyard.

Figure 2. Fault factor product generated from aerial imagery.

2.4 Satellite imagery

The complete series of Sentinel-2 satellite images from May 2017 to August 2022 was processed. Three vegetation indices were selected as follows:

1 - The LAI index, which provides canopy information. The relationship between NDVI and LAI is widely known. The NDVI index is indicated in crops where at maximum vigour the bare soil component is missing in the measured reflectance from the satellite, so it comes from vegetation reflectance only. This is not the scenario in vineyards, where bare soils remain in the phase of maximum vigour, so to isolate the bare soil component leaving the part coming from vegetation the reflectance measured from the satellite, the indexes that correct the soil brightness are recommended. The Soil Adjusted Vegetation Index (SAVI) was the selected index. The best relationship between SAVI and LAI in the study area was then established by working on the time series, the aerial orthophoto and local information. The relationship between SAVI and LAI must be adjusted to each crop and meteorological conditions [4].

2 - The fAPAR index, which quantifies the fraction of solar radiation absorbed by living leaves for photosynthetic activity. The used equation was adapted to Sentinel 2 from the algorithm developed for the Copernicus Global Land Service.

3 - The DSWI index, which accounts for the moisture at the soil/vegetation interface. Previous studies have reported a very good correlation with the LAI index [8].

Table 1 Summarises the managed vegetation indices. More than 700 images have been processed to generate 75 fortnightly vegetation indices composites from March to October each year. 17 composites could not be generated because the cloudy conditions in each Sentinel-2 satellite pass (revisit every 5 days) were permanent in all the vineyards under study. Within the 75 generated composites, gaps due to partial cloud coverages were filled by applying an interpolation process. A fortnightly LAI composite time series over one vineyard is shown in (Fig. 3). The handling of the historical series is of great importance, as it allows the comparison of current values with the averages or maximum that could be expected based on what was observed in previous years. An example of this is the recent general heatwave episode in Spain in the summer of 2022, the effect of which is reflected in LAI values that, fortnight by fortnight, are permanently below the expected average.

In addition to the temporal analysis at the parcel level, spatial analyses are also made within the parcels. Within each plot, the status of the LAI, fAPAR and DSWI indices is monitored every 10 metres in comparison to the historical average and maximum values recorded, which in turn are the inputs for a second block of value-added products oriented to support decision-making. An example of this is shown in (Fig. 4), where a zoned vegetation quantity product is generated from the LAI index and updated fortnightly. The information provided is summarised in three possible scenarios:

- Bellow expected: below the historical average.
- Slightly high: between the average and the historical maximum.
- Very high: the vegetation quantity is at an all-time high.

Table 1. Sentinel-2 vegetation indices description.

Acronym	Name	Algorithm	Inputs
SAVI	Soil Adjusted Vegetation Index	*(NIR - RED) / (NIR + RED + L) * (1.0 + L)*	B04, B8A, L
LAI	Leaf Area Index	*(-1 / a2) * Ln ((SAVI - a0) / (-a1))*	SAVI, a0, a1, a2
fAPAR	fraction of Absorbed Photosynthetically Active Radiation	*Function of GREEN, RED, REDEDGE1, REDEDGE2, REDEDGE3, NIR, SWIR1, SWIR2, ViewZen, SunZen, RelAzim*	B03, B04, B05, B06, B07, B8A, B11, B12, ViewZen, SunZen, RelAzim
DSWI	Disease water stress index	*(NIR + B03) / (B11 + B04)*	B03, B04, B8A, B11

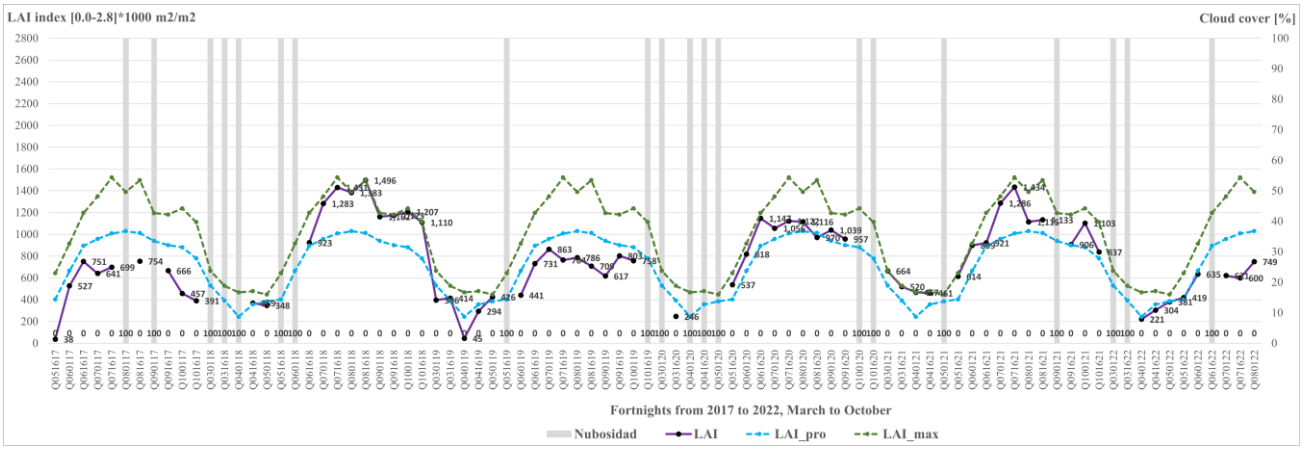

Figure 3. Fortnightly LAI composite time series. The purple line shows the evolution of the plot-averaged LAI, while the blue and green lines are repeated every year because they are the historical average and maximum fortnightly values, respectively.

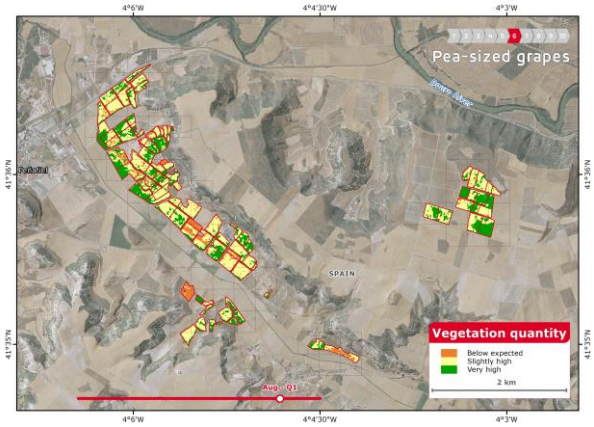

Figure 4. Vegetation quantity product derived from the current and historical fortnightly LAI index.

2.5 Methods

The methodological approach is based on a machine learning (ML) model implemented and refined over the last three years of collaboration between GMV and Pago de Carraovejas. Inputs are manual sampling, aerial orthophoto and satellite imagery. Field visits are carried out on reference parcels and grape cluster counting is performed following the steps described in Sect. 2.2. The expected yield can then be estimated by following theoretical equations used by winegrowers. A general theoretical equation would be the following:

$$\textit{Theoretical yield estimation [kg/ha]} = \textit{Function (n° of vines, n° of bunches per vine, n° of grapes per bunch, berry weight, vineyard area)} \quad (1)$$

In Eq. (1) the number of vines is calculated from the planting frame and the area of the vineyard. Here the fault factor, generated from aerial imagery as described in Sect. 2.3, is used to indicate the current number of vines in each vineyard instead of the theoretical one.

Once the theoretical yield estimation was improved thanks to the updated fault factor, the ML model was implemented using satellite imaging in the current year plus the value-added products generated from the analysis of the historical series. The general equation is as follows:

$$\textit{Satellite yield estimation [kg/ha]} = \textit{Function (Theoretical yield estimation, } LAI_f(c,h), fAPAR_f(c,h), DSWI_f(c,h), \textit{VA-}LAI_f, \textit{VA-}fAPAR_f, \textit{VA-}DSWI_f) \quad (2)$$

Where *f* indicates the fortnightly composite starting with the sprouting of the vineyard, *c* indicates the current year, *h* indicates the historical years and the prefix *VA* indicates the value-added products. The dataset entered in Eq. (2) are yields and satellite-based statistic values at the parcel level.

The choice of the ML model to be implemented started with a previous exploration work based on Auto Machine Learning (AutoML). AutoML trains and tests a wide range of regression ML algorithms starting from a dataset. The result is ranking the regressors by the score achieved by each one and a summary of the needed preprocessing or feature engineering. The results cannot be considered definitive, as the probability that the winning algorithms are overtrained is high, but it is a starting point as it gives clues as to which algorithms are the best along with their hyperparameters. Figure 5 shows the result of AutoML work on a dataset generated according to the indications in Eq. (2).

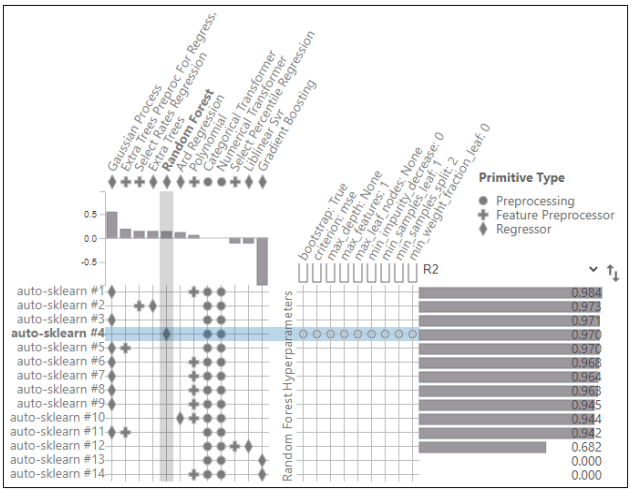

Figure 5. Result of AutoML processing trained with a formed dataset following the Eq. (2) indications.

3 Results and Discussion

3.1 Season 2020

Every year Pago de Carraovejas schedules field visits in the first fortnight of August to conduct cluster counting. The 2020 season was focused on the implementation and calibration of the best ML model: 68 parcels were visited (65 Tempranillo and 3 Cabernet Sauvignon varieties). The vineyards have planting dates ranging from 1989 to 2015, with different planting frames from 2.8*1.1 metres to 3.0*1.5 metres. They range in size from 0.24 to 5.6 hectares. The theoretical yield estimation was made according to Eq. (1).

Aligned with their planning, the fortnightly LAI, fAPAR, DSWI and the corresponding value-added composites were generated up to that fortnight. Statistics were generated at the parcel level such as averages, standard deviations, and percentages. This information, together with the yield estimation, constitutes the dataset needed for the satellite yield estimation following Eq. (2). In the AutoML exercise, the parcels are sorted from the smallest to the largest area, leaving 33% of the sample for validation. This ensures that parcels of any extent are available for both training and validation. The best model results in a predictive score of R2 equal to 0.82 and an RMSE of 1.66 Tons.

The largest deviations in yield estimation were observed in vineyards where vegetation indices indicated high heterogeneity, especially in vineyards where there were patches without vegetation at veraison, which was an indicator of non-productive areas. This was contrasted with the 2020 aerial orthophoto as shown in (Fig. 6). As explained in Sects. 2.3 and 2.5, the fault factor layer was

generated and incorporated into Eqs. (1) and (2). The AutoML exercise was repeated with the sample sorted in the same way and results were achieved in a predictive score of R2 equal to 0.91 and an RMSE of 1.14 Tons. Considering the fault factor had an impact of a 9% improvement in R2 and 0.5 tonnes in RMSE.

Up to that point, an ML model was established which was well-trained with data from manual sampling plus satellite imagery. The final ML model validation was carried out with the real yield once the harvest was finished. The result found was that the yield estimation following only the theoretical equation yielded an average deviation per plot of 2.68 tonnes, while the satellite yield estimation lowered this deviation to only 1.65 tonnes per parcel, i.e., decreasing the deviation by almost 50%.

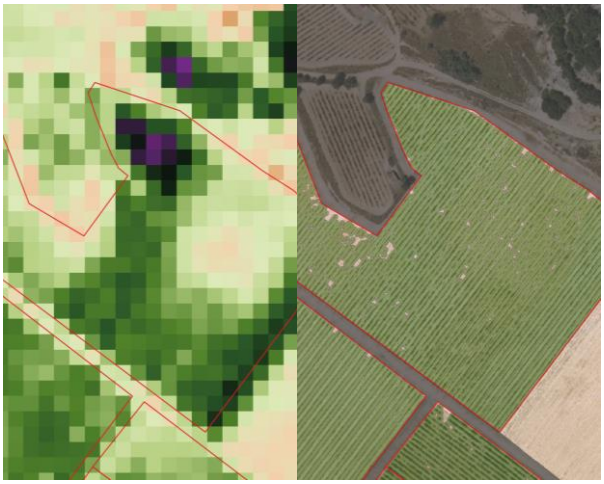

Figure 6. Left, the LAI index for the fortnight of July 2020. Right, the generated fault factor using the aerial imagery from the 2020 flight.

3.2 Season 2021

The objective for 2021 is to test with a small sample of field data that the ML model implemented in the 2020 season works properly. More parcels were added in 2021, bringing up to 89 parcels in total (70 Tempranillo, 13 Cabernet Sauvignon, and 6 Merlot varieties). For the new parcels, the fault factor was generated using the 2020 aerial orthophoto. That is, the fault factor layer was the one corresponding to 2020. Regarding satellite imagery, on one hand, historical LAI, fAPAR and DSWI composites were updated with the 2020 fortnightly composites, on another hand, fortnightly composites were generated until the first fortnight of August of 2021. In that fortnight Pago de Carraovejas visited them for bunch counting except for the Merlot parcels because they are located on terraces. Just 20 vineyards were considered for running the ML model following Eqs. (1) and (2).

The estimation was validated against the real yield. The satellite estimation underestimated the total yield, achieving an overall accuracy of 94.2%, but improved on the 92.3% achieved theoretically. Figure 7 shows the yield estimation at the parcel level and the overall accuracy achieved. According to these results, the ML model trained with satellite images estimated the yield better than only based on field visit plus theoretical equation. The ML model required a small sample of bunch counts, but only in a few reference vineyards, which means fewer visits to the field. Moreover, it estimated the production in the entire property, regardless of its location on terraces and the difficulty of counting bunches in these areas.

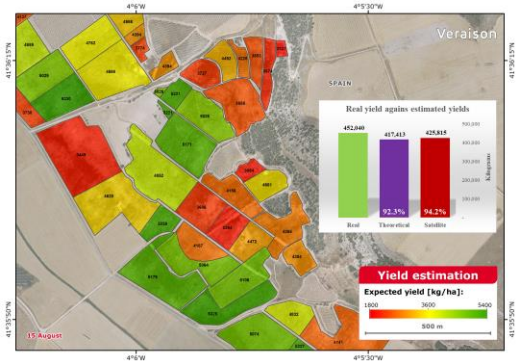

Figure 7. Map with the estimated yield at parcel level (kg/ha), veraison stage, 15th August 2021. The bar chart shows the total yield versus the estimated yields, according to the theoretical equation and ML model with satellite.

3.3 Season 2022

The objective for 2022 is the consolidation of the ML model. The same path described in the 2021 season was repeated in 2022. Just 20 vineyards were considered for running the ML model following Eqs. (1) and (2). The estimation was validated against the real yield and the overall accuracy was 95.1%. However, the yield was underestimated again, being significant in the case of the Merlot variety, where the accuracy achieved was 70%. There are two important factors to consider here:

1 - In the vineyards with Merlot variety, past visits for cluster counting were not conducted, as they are located on terraces, which complicates their implementation. Therefore, the ML model was trained with Tempranillo and Cabernet Sauvignon varieties.
2 - Season 2022 was an anomalous year marked by low precipitation and high temperatures, as reflected in (Fig. 3), where the LAI index was permanently below the expected average value.

To assess the impact of these two factors, a simple analysis was carried out. The yield of Merlot vineyards was divided by the number of vines and the average yield per vine was obtained. The same exercise was carried out with the Tempranillo and Cabernet Sauvignon. The result was that in season 2022, the Merlot variety had an average yield of 3.08 kilos per vine, and the Tempranillo and Cabernet Sauvignon vineyards had a yield of 2.12 kilos per vine. The difference of almost one kilogram per vine in 2022 needs to be analysed in detail in previous years to assess whether it is the expected difference or whether it has been accentuated due to the anomalous climatic year. If the difference is maintained, it must be considered in Eq. (1) that the weight of the berry is different according to variety.

A simulation was carried out by introducing different berry weights in Eq. (1) according to variety and the ML model was re-run. The result was an estimated yield slightly higher than the real yield, reaching up to 98% overall accuracy, such that the underestimation disappeared in global terms, but was maintained in the Merlot variety. Nevertheless, a positive noteworthy aspect is the accuracy achieved in Merlot improved from the previous 70% to 86%. Table 2 summarises the accuracies achieved for each planned target.

Table 2. Planned targets each season and achieved results.

Season	Targets	Results
2020	ML model implementation	Predictive R2: 0.91 Predictive RSME: 1.14 Tns
2021	ML model testing	Real yield: 452 Tns Estimated yield: 425 Tns Overall accuracy: 94.0%
2022	ML model consolidation	Real yield: 568 Tns Estimated yield: 540 Tns Overall accuracy: 95.1%

4 Conclusions

This study presents a methodology for yield estimation at the vineyard parcel level with an overall accuracy of 90-95% compared to the total real vintage for Tempranillo, Cabernet Sauvignon and Merlot varieties. The vineyards under study range in age from 6 to more than 30 years old with surfaces from 0.24 hectares, several training grapevines and different planting frames. The methodological approach is based on a suitable ML regressor selected from AutoML processing. The input data are an aerial orthophoto, fieldwork reduced to a few vineyards for bunch counting and satellite vegetation indices in fortnightly time series. The time for yield estimation has been aligned with the time of the field visit for bunch counting, which is at veraison. The same approach can be adopted if the bunch counting fieldwork is brought forward to the pea-size stage.

The improvement in yield estimation at the parcel level when considering the fault factor in both theoretical and satellite estimation was 9% and 0.5 tonnes in the predictive R2 and RMSE scores, respectively. The frequency of updating the fault factor depends on the dynamism of the vineyards. It may be necessary to update it every year or it may be sufficient every 3 or 5 years. In this study, public area orthophoto from the Spanish National Aerial Observation Plan (PNOA) of 2020 was used and the fault factor was not updated in 2021 and 2022 either. In the case that an annual update is required, it is not possible to carry it out with public aerial orthophoto, but nowadays images from commercial satellites with spatial resolutions of up to 30 centimetres can replace aerial orthophoto. In addition, it is possible to programme the capture at the time of veraison of the vineyards, so that, at the same time as updating the fault factor, vegetation indices of very high resolution can be calculated, since they not only have RGB spectral resolution, but also in other bands of the electromagnetic spectrum such as the red edge or mid-infrared.

Once the ML model is fixed, the cluster counting fieldwork is reduced to a few reference parcels, whereas the yield estimation is made for a much larger number of vineyards than those visited. The extent may be much larger than the study area in this work. However, it is not always possible to do cluster counting fieldwork, either because there is no staff with the time and dedication to do it or because the location of the vineyards, for example on terraces, makes it very difficult to carry out. At present, efforts are being made to develop a predictive model without counting bunches but trained with real yields from past seasons at the parcel level. In addition, not only satellite-based explanatory variables are considered, but also climatic variables from weather stations. Moreover, this model will be trained with years that are in the expected climate average, such as the past anomalous year 2022. The theoretical yield estimation Eq. (1) will not be used and there is no dependence on berry weight per variety as seen in the case of Merlot.

The Sentinel-2 time series began in 2017 with the commissioning of the Sentinel-2A satellite. Today it is the Sentinel-2A/2B satellites that monitor every point of the earth with a 5-day review and its continuity is guaranteed with the scheduled launch of the Sentinel-2C satellite in 2024 plus the Sentinel-2D satellite undergoing tests with a view to its launch at the end of Sentinel-2B's lifetime. This means that the historical fortnightly vegetation index composites used here will continue to be fed from a growing inter-annual sample, picking up the particularities of each event as the machine learning algorithms learn better from each of these events.

References

1. L. Ghiani, A. Sassu, F. Palumbo, L. Mercenaro, F. Gambella, In-Field Auto. Det. of Grape Bunches under a Totally Uncontrolled Environment. *MPDI Sensors* **21**, 3908 (2021)
2. A. Barriguinha, M. de Castro Neto, A. Gil, Vineyard Yield Estimation, Prediction, and Forecasting: A Syst. Lit. Review. *Agronomy* **11**, 1789 (2021)
3. D. Andújar, H. Moreno, J.M. Bengochea-Guevara, A. de Castro, A. Ribeiro, Aerial ima. or on-ground detection? An eco. analysis for vineyard crops, *Computers and Electronics in Agriculture* **157**, 351-358 (2019)
4. S. Bajocco, F. Ginaldi, F. Savian, D. Morelli, M. Scaglione, D. Fanchini, E. Raparelli, S.U.M. Bregaglio, On the use of NDVI to est. LAI in field crops: Impl. a Conv. Equation Library. *MDPI Remote Sens.* **14**, 3554 (2022)
5. C. Magarreiro, C.M. Gouveia, C.M. Barroso, I.F. Trigo, Modelling of Wine Prod. Using Land Surface Temperature and FAPAR—The Case of the Douro Wine Region. *Remote Sensing* **11**(6), 604 (2019)
6. B. Zbigniew, Z. Dariusz, B. Maciej, O. Karolina, O. Adrian, Monitoring For. Bio. and the impact of climate on forest environment using high-resolution

satellite images, *Taylor & Francis, European Journal of Remote Sensing* **51**, 1 (2017)
7. A. Barriguinha, B. Jardim, M. de Castro Neto, A. Gil, Using NDVI, climate data and machine learning to est. yield in the Douro wine region, *International Journal of Applied Earth Observations and Geoinformation* **114**, 103069 (2022)
8. Z. Bochenek, R. Gurdak, F. Niro, M. Bartold, P. Grzybowski, Validation of the LAI bio. product derived from Sentinel-2 and Proba-V images for winter wheat. *Geoinformation Issues* **9**, 1(9), 15-26 (2017)