# 62. Enhancing potato yield estimation using vegetative indices, SAR imagery, and terrain data

I. Pastor[1,2], A. Aizpurua[1], A. Carrasco[3], J. Legorburu[3], J. Castro[1] and A. Uribeetxebarria[1,*]
[1]*NEIKER-Basque Institute of Agricultural Research and Development, Berreaga Kalea, 48.160 Derio, Bizkaia, Spain; [*]auribeetxebarria@neiker.eus2*
[2]*UPV/EHU Barrio Sarriena, s/n., 48940 Leioa, Bizkaia, Spain*
[3]*UDAPA S. Coop., C/ Paduleta n°1, Pol. Ind. de Júndiz 01015 Vitoria-Gasteiz, Álava, Spain*

**Abstract**

Having an accurate potato yield estimation is of the upmost importance to help improve the agronomic management of plots. The aim of this study was to analyse the relationship between synthetic aperture radar (SAR) images, vegetation indices (VI) derived from the Copernicus network, a high-resolution digital elevation model (DEM) and orthophotos produced with a camera mounted in a plane with the yield of a potato plot. Multiple regression technique, was used to estimate yield, resulting in an equation with an error of 17% and of $R^2$ of 0.59. It is possible to obtain fairly accurate estimates of potato yields with freely available satellite and other information.

**Keywords:** DEM, NDVI, orthophoto, Sentinel-1, Sentinel-2

**Introduction**

Variations in potato (*Solanum tuberosum* L.) production in the EU impact many stakeholders in the agri-food sector. This is the reason why since 1993, in-season crop yield forecasts for major crops have been made by EU member. In the case of potato crops from the Basque Country, a significant portion is managed by the UDAPA cooperative. In this case, having an early and accurate estimate of the yield could be useful from a logistical and planning point of view. With the objective to maintain the quality of the final product, a significant portion of the potato supply, sourced from various locations, is received and stored within a short time frame. Therefore, potato harvesting and storage represent one of the most stressful and labour-intensive periods. In this sense, having a reliable harvest estimate is an invaluable tool for the cooperative.

The use of satellite-derived information for yield estimation in agriculture has become increasingly common in recent years. In some cases, work is done at regional level using average yield/plot data such as the studies carried out by Newton *et al.* (2018) in Bangladesh; Gomez *et al.* (2019) in Spain; Salvador *et al.* (2020) in Mexico; or Vannoppen and Gobin (2022) in Belgium. However, as Mhango *et al.* (2021) point out in their paper, intra-plot level works are less common. The underlying reason for the limited presence of this type of work is the lower adoption of yield monitors with potato crop compared to cereals. In terms of plot-level studies, those carried out in Saudi Arabia by Al-Gaadi *et al.* (2016) and Mhango *et al.* (2021) in the UK are stand out. In the first case, differences in yield from 21 to 40 t/ha were detected. Therefore, it is highlighted that intra-plot analysis can provide relevant information on yield variability at plot level that can be considered to adjust agronomic management practices. In this sense, this approach could be utilized to optimize fertilization practices, adjusting them more effectively to crop requirements. These productive differences at the plot level are highly conditioned by soil variability, which can sometimes be inferred from images of the soil surface.

Multispectral optical satellites are the most commonly used for yield estimation (Muruganantham *et al.* 2022), however, other information sources like synthetic aperture radar (SAR) satellite information, or terrain information are increasingly being used, especially for crop identification (d' Andrimont *et al.*, 2021), for crop phenology monitoring (Khabbazan *et al.*, 2019) or cereal yield prediction (Uribeetxebarria *et al.* 2023) with very promising results. Therefore, the inclusion of SAR satellite variables, in addition to those originating from multispectral satellites, could perhaps improve the potato yield estimation algorithms. On the other hand, Al-Gaadi *et al.* (2016) assure that including environmental and soil factors in productivity analysis can help improve yield productivity. The objective of this study was to assess the utility of integrating vegetation indices derived from Sentinel-2 imagery, backscatter information from Sentinel-1, and terrain-specific data to estimate intra-plot yield variability in a potato crop field using multivariate techniques.

## Materials and methods

The study plot was in the municipality of Elburgo (Álava/Araba) which had an area of 5.2 ha (Figure 1). On 20 May 2020, seed potatoes of the Lucinda variety were planted with a planting frame of 75 x 18 cm, which resulted in a planting rate of 74 000 plants/ha. Harvesting took place on 2 September 2020. Regarding the soil, the plot had two very distinct soil types. The first, located in the north, has a lighter colour on the surface, while the darker coloured soil was found in the rest of the plot. Due to the soil evaluation work carried out in the municipality of Vitoria-Gasteiz (Unamunzaga *et al.*, 2018) adjacent to Elburgo, which belongs to the same geological formation, information is available as to which soils correspond to the different tonalities that can be seen in the orthophotos. In the case of the Elburgo plot, two test pits were also examined, one in each soil surface colour, which confirmed the initial ideas. When describing and analysing the soil pits, it was possible to verify that the area that showed light surface colours in the orthophoto corresponded to shallower soils (45–50 cm depth) with Cretaceous lithological material, and the area that appeared dark in the orthophoto showed soils of greater depth (more than 200 cm), and therefore, greater water storage capacity, composed of material from the Quaternary period. The plot yield was obtained by systematic sampling carried out the day before harvest (1 September 2020). Aware that the spatial pattern of crop vigour in this plot tends to remain consistent across seasons in cereals and correlates with yield, the authors concluded that an effective approach to capturing the spatial variability of yield within the field and optimize sampling effort was to use half the range of the previous year's soil adjusted vegetation index (SAVI) semivariogram. Following this methodology, the sampling distance was set at 43 m. In total, 45 points were sampled (Figure 1). Each sampling point was georeferenced using a precision GNSS (Topcom GRS1) that receives RTK corrections from the Bizkaia network, which allows working with positioning errors of less than 3 cm. In each sampling point, all the potatoes of a linear meter were collected, to later measure the diameter of the potatoes and obtain the sum of the potatoes weight. The linear meter was taken towards the west of the point marked by the GNSS. Images from the Sentinel-1 radar satellite (S1) and Sentinel-2 optical satellite (S2) were used in this work. The images were downloaded from the Copernicus Browser platform (https://browser.dataspace.copernicus.eu/). Based on the information from S-1, the backscatter of vertical (VV) and horizontal (VH) orientation were calculated. In the case of S2, all cloud-free images of the study area from 3 May 2020 to 31 August 2020 were used, thus covering the entire phase of crop development. A total of 12 images were used. For S1, the images coincident with S2 were chosen. Seven VIs previously used for potato yield estimation were selected.

In regard to the processing of the images, it is important to highlight that the size of each raster image was reduced to the exact size of the study plot by using a vector mask of the plot area. Finally, using the "Buffer" tool, the raster image size for each VI was reduced by 10 m to ensure that all pixels remained within the parcel and did not capture information from outside its boundaries.

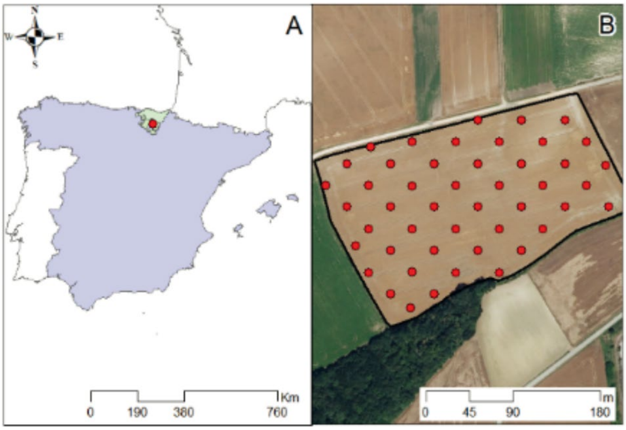


Figure 1. (A) Location of the Elburgo potato plot. (B) Location of potato sampling points within the plot.

Table 1. Vegetation indices calculated in this study with their formula according to the band names.

| Abbreviation | Full name | Equation |
|---|---|---|
| NDVI | Normalized Difference Vegetation Index | (NIR–RED)/(NIR+RED) |
| SAVI | Soil Adjusted Vegetation Index | ((IRC–R)/(IRC+R+L)) (1+L) |
| GNDVI | Green Normalized Difference Vegetation Index | (NIR–GREEN)/(NIR+GREEN) |
| IRECI | Inverted Red-Edge Chlorophyll Index | (NIR–R)/(RE1/RE2) |
| MCARI | Modified Chlorophyll Absorption in Reflectance Index | ((NIR–RED)–0.2*(VNIR–Green))*(NIR/RED) |
| WDVI | Weighted Difference Vegetation Index | IR – a*RED |
| PSRI | Plant Senescence Reflectance Index | (RED–BLUE)/NIR |

In addition to the VIs described above, to perform the multiple regression, the backscatter variables obtained from S1 in both polarizations (VH and VV), and those corresponding to the topographic information (elevation, slope) were added. Information from the three bands (RGB) of the 2009 orthophoto, which was taken when the plot had bare soil, was also included. Finally, a new band derived from the fusion of the individual bands of the 2009 orthophoto was added. Very briefly, to perform the multiple linear regression, it was first necessary to determine the goodness of fit of the model to the data. Subsequently, the model was refined using the "backward" technique until the model had and average variance influence factor (VIF) of less than 10. Since more variables were available than the number of samples, a pre-selection of variables was made to select those to be used in the multiple linear regression model. The selection was made considering the coefficient of determination and those from different sources that could provide different information to the model.

## Results and discussion

Practically all VIs reached the maximum value between 20 and 27 July, about 60 days after planting (Figure 2). Generally, from this point onwards, the values began to decline. The only exception was PSRI (Merzlyak *et al.*, 1999) which reached its maximum value in the last sampling (Figure 2). This is logical as it is sensitive to the retention or accumulation of carotenoids in aging leaves and ripening fruit, leading to behaviour distinct from that of the other indices. Thus, in the initial stages of the

crop there is little tissue in the senescence phase, but in the last sampling on August 31 the high value of the index shows that there was more mature or deteriorating tissue, which is consistent with the final stage of the crop. The rest of VI offered very low values in this final phase. Among the indexes studied, IRECI (Frampton *et al.*, 2013) stands out because it presented a very marked pattern of rise and fall unlike the rest that maintained high and quite similar values during a quite large period of time, drawing a kind of plateau around the maximum value. This VI has been shown to be linearly related to canopy chlorophyll content at a near 1:1 ratio while still performing well for leaf area index (LAI) up to and beyond the common saturation point (Frampton *et al.*, 2013).
In the case of NDVI the maximum mean value was 0.76 (Figure 2). It was similar to the obtained by Vannoppen and Gobin (2022) in Belgium which for late potato would be between 0.6 and 0.7 and for early potato between 0.7 and 0.8. But they would be higher values than those obtained by Newton *et al.* (2018) for plots located in Bangladesh where the maximum mean value was 0.57. However, they are lower than those obtained by Mhango *et al.* (2021) which ranged from 0.81 to 0.94 for UK plots.
Table 2 details the $R^2$ of each VIs at each time point with yield. This relationship is closer as the cycle progresses and the time of highest relationship corresponds to the dates July 27 and August 6 for all vegetation indices (68 and 78 days after planting). In the study conducted in Saudi Arabia by Al-Gaadi *et al.* (2016), data were collected from three plots (pivot) of 30 ha each. In each of the pivot, yield was measured at 45 points. When relating the NDVI value derived from S2 of about 55 - 61 days after planting to yield the $R^2$ ranged from 0.23 to 0.49 depending on the plot. The values associated with the relationship obtained with the SAVI index were similar, as the $R^2$ value ranged from 0.22 to 0.50. In this case, for NDVI and SAVI readings taken 63 days after planting, the corresponding $R^2$ was 0.29 and 0.32, respectively. Therefore, the coefficient of determination obtained is within the range of Al-Gaadi *et al.* (2016). However, the highest coefficient of determination ($R^2$=0.33) corresponds to the MCARI index on August 6 (Table 2).
Subsequently, 42 variables were selected to obtain a model based on multiple linear regression. In addition to the VIs, the backscatter variables obtained from S1 from the same dates as the VIs obtained from S2 were also included. As in previous work (Khabbazan *et al.*, 2019) the S1 backscatter information was used to monitor potato crop development, it was thought that it could also contribute to yield estimation. In addition to the information derived from the satellites, variables such as

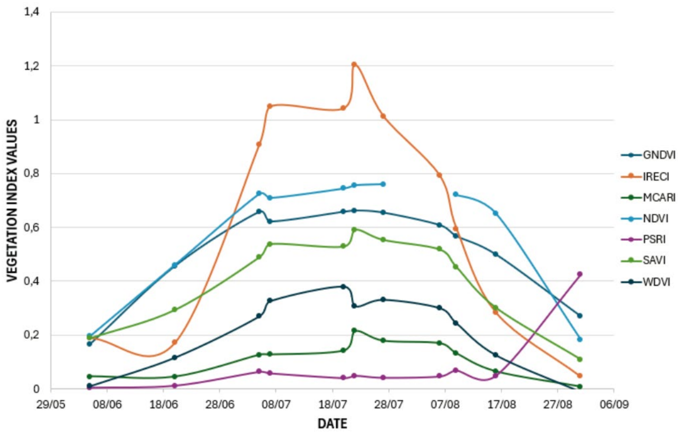


Figure 2. Evolution of the seven vegetation index (VI) values (GNDVI, IRECI, MCARI, NDVI, PSRI, SAVI, WDVI), obtained from Sentinel-2 satellite during the potato crop (Lucinda variety) cycle in the Elburgo, Alava (Spain) plot during the 2020 season. The sowing date was 20 May and the harvest took place on 2 September.

Table 2. Coefficients of determination ($R^2$) between the vegetation indices obtained from Sentinel-2 and Sentinel-1 (VH, VV) and the potato yield in the plot of Elburgo (Alava) at different dates duting the 2020 season.

| VI | 20 June | 5 July | 7 July | 20 July | 22 July | 27 July | 6 August | 9 August | 19 August | 31 August |
|---|---|---|---|---|---|---|---|---|---|---|
| NDVI | 0.18 | 0.24 | 0.24 | 0.23 | 0.22 | 0.29 | 0.28 | 0.25 | 0.01 | 0.11 |
| GNDVI | 0.17 | 0.21 | 0.20 | 0.17 | 0.16 | 0.21 | 0.24 | 0.20 | 0.12 | 0.09 |
| IRECI | 0.23 | 0.19 | 0.21 | 0.21 | 0.17 | 0.21 | 0.18 | 0.15 | 0.07 | 0.00 |
| MCARI | 0.16 | 0.22 | 0.21 | 0.21 | 0.22 | 0.28 | 0.33 | 0.32 | 0.07 | 0.00 |
| PSRI | 0.22 | 0.21 | 0.16 | 0.16 | 0.22 | 0.25 | 0.21 | 0.00 | 0.00 | 0.08 |
| SAVI | 0.20 | 0.26 | 0.26 | 0.26 | 0.24 | 0.32 | 0.26 | 0.24 | 0.09 | 0.03 |
| WDVI | 0.19 | 0.26 | 0.26 | 0.24 | 0.27 | 0.32 | 0.25 | 0.22 | 0.09 | 0.13 |
| VH | 0.06 | 0.00 | 0.00 | | | | 0.07 | 0.00 | 0.00 | 0.00 |
| VV | 0.12 | 0.00 | 0.01 | | | | 0.03 | 0.03 | 0.00 | 0.12 |

slope and elevation of each sampling point were added in addition to the RGB bands of the 2009 orthophoto and another variable derived from the fusion of these.

The final model (Eq. 1) was obtained by eliminating 28 variables that contributed the least to the initial model using the backward technique. This approach ensured that collinearity was controlled (VIF < 10) and that the model remained statistically significant ($p$=0.004). The final model included 14 variables. Multiple tests were conducted over model residuals to confirm that the model did not violate any of the assumptions required for MLR. Normality was verified using the Shapiro–Wilk test ($p$=0.56), and homoscedasticity was confirmed using the Breusch–Pagan test ($p$=0.43), providing evidence that the homoscedasticity assumption was met. The Durbin–Watson test ($p$=0.09) was performed to confirm the absence of autocorrelation among the variables. Spatial autocorrelation was assessed using Moran's I and Lagrange Multiplier tests. While Moran's I test returned a $p$-value of 0.86, the Lagrange Multiplier test yielded $p$-values of 0.20 for spatial error and 0.18 for spatial lag, ruling out the presence of spatial autocorrelation. Generally, in all cases, the variables selected in the model correspond to a date just after or coinciding with the date on which the maximum value of the VI was reached, with WDVI on July 7 being the only exception. Some of the variables correspond to a period in which the value of the VI is still high, as is the case of NDVI in which the selected variable corresponds to 22 July. In the case of the MCARI index, it corresponds to 27 July 27 where the value was maximum. In other cases, the selected variable is in clear decline as is the case of WDVI on 6 August.

$$\begin{aligned}&\text{Yield (kg linear meter}^{-1}\text{)} = -1.87+(6.60*\text{GNDVI August 31})+((-3.49)*\text{IRECI August}\\&\text{06})+((-8.21)*\text{MCARI July 27})+(8.20*\text{MCARI August 09})+(29.29*\text{PSRI July 22})+\\&((-11.36)*\text{PSRI August 31})+(8.65*\text{NDVI July 22})+(2.98*\text{WDVI July 07})+(24.24*\text{WDVI}\\&\text{August 06})+(0.26*\text{VV\_S1\_June 20})+((-3.43)*\text{VV\_S1\_August 09}+\\&(\text{VV\_S1\_August 31}*(-0.12))+(\text{b1\_orto\_24}*0.001)+(\text{slope}*(-0.10)))\end{aligned} \tag{1}$$

Where the formulas for calculating VI (GNDVI, IRECI, MCARI, PSRI, NDVI, and WDVI) are explained in Table 1; VV_S1 corresponds to vertical polarization of S-1 backscatter; b1 is the red band of the orthophoto, and slope is expressed in degrees (°).

The variables from S1 selected were VV S1 June20, VV S1 August 09, VV S1 August 31. The final model retains soil colour and slope variables. As for the influence of the colour of the orthophoto on the yield estimation equation, the different soil properties, especially in terms of water storage capacity of the soils described in each colour (see materials and methods) would explain the influence on yield of this variable. It should be noted that although potatoes are irrigated, the amount of water applied does not usually meet the water requirements that would lead to maximum production. Although the plot is quite flat, the most productive area corresponds to the flattest area which is located to the south and the soil colour is darker.
The $R^2$ of the model was 0.59. The root mean square error (RMSE) was 0.59 kg potato linear meter$^{-1}$, which, considering that the average weight per linear meter was 3.34 kg of potato, represents a deviation of 17% from the real weight. Several authors have made models based on different algorithms in which they have included VI, spectral bands of satellite origin or both, obtaining similar $R^2$ values. In the case of Mhango *et al.* (2021), several yield data were obtained from one linear meter in five plots in United Kingdom (UK), and a model was developed in which yield was estimated for the five plots from NDVI at the onset of tuberization, two wavelength reflectance values ($\lambda_{559}$ and $\lambda_{703}$), and stem density. In this case the $R^2$ was 0.65. In the work of Al-Gaadi *et al.* (2016) in which intra-plot yield data were also available, a model using different NDVI variables was performed and $R^2$ between 0.47 and 0.65 were achieved depending on the plot. On the other hand, there are other types of models using average plot yield data, such as those employed by Gómez *et al.* (2019) in which they reported $R^2$ between 0.89 and 0.93 using S2 images in Spain. In Belgium, Vannoppen and Gobin (2022) improved their yield estimation model using the trapezoidal rule, and adding data on monthly precipitation, maximum temperature and soil water depletion in the root zone. For this study, they used yield data from 685 plots for 3 years and the $R^2$ obtained was 0.56, very similar to that obtained in this study ($R^2$=0.59).

## Conclusions

A model using various S2 satellite-derived VIs, S1 satellite-derived backscatter images, underlying soil properties and DEM-derived properties was developed to estimate the yield of a plot. The final model, which integrated the above variables, showed a significant relationship with yield ($R^2$=0.59), comparable to previous studies in other regions. The accuracy of the model, with an RMSE of 17% with respect to mean yield, suggests that VI, especially those close to the time of peak value, are key to yield estimation. Furthermore, the incorporation of terrain characteristics such as slope and soil surface colour, in conjunction with S1 data, suggests a potential enhancement in the model's robustness. These results confirm the potential of integrating data from multiple sources to predict potato yield in farming systems.

## Acknowledgment

This work is the result of the collaboration between NEIKER and the UDAPA cooperative within the framework of the AgritechZeha and NITRALDA projects, financed by the Basque Government, Department of Economic Development, Sustainability and Environment.

## References

Al-Gaadi, K.A., Hassaballa, A.A., Tola, E., Kayad, A.G., Madugundu, R., Alblewi, B., & Assiri, F. (2016). Prediction of potato crop yield using precision agriculture techniques. PLoS ONE, 11(9), e0162219.

D'Andrimont, R., Verghegghen, A., Lemoine, A., Kempeneers, P., Meroni, M., & van der Velde, M. (2021). From parcel to continental scale – a first European crop type map based on Sentinel-1 and LUCAS Copernicus in-situ observations. Remote Sensing of Environment, 266, 112708.

Frampton, W.J., Dash, J., Watmough, G., & Milton, E.J. (2013). Evaluating the capabilities of Sentinel-2 for quantitative estimation of biophysical variables in vegetation. ISPRS Journal of Photogrammetry and Remote Sensing, 82, 83–92.

Gómez, D., Salvador, P., Sanz, J., & Casanova, J.L. (2019). Potato yield prediction using machine learning techniques and Sentinel 2 data. Remote Sensing, 11(15), 1745.

Khabbazan, S., Vermunt, P., Steele-Dunne, S., Arntz, L.-R., Marinetti, C., & van der Valk, D. (2019) Crop monitoring using Sentinel-1 data: a case study from The Netherlands. Remote Sensing, 11(16), 1887.

Merzlyak, M.N., Gitelson, A.A., Chivkunova, O.B., & Rakitin, V.Y. (1999). Non-destructive optical detection of pigment changes during leaf senescence and fruit ripening. Physiologia Plantarum, 106, 135–141.

Mhango , J.K., Harris, W.E., & Monaghan, J.M. (2021). Relationships between the Spatio-temporal variation in reflectance data from the Sentinel-2 satellite and potato (*Solanum tuberosum* L.) yield and stem density. Remote Sensing, 13(21), 4371.

Muruganantham, P., Wibowo, S., Grandhi, S., Samrat, N.H., & Islam, N. (2022). A Systematic literature review on crop yield prediction with deep learning and remote sensing. Remote Sensing, 14(9), 1990. https://doi.org/10.3390/rs14091990

Newton, I.H., Islam A.F.M.T., Islam A.K.M.S., Islam G.M.T., Tashin A., & Razzaque, S. (2018). Yield prediction model for potato using Landsat time series images driven vegetation indices. Remote Sensing in Earth Systems Sciences, 1, 29–38.

Salvador, P., Gómez, D., Sanz, J., & Casanova, J.L. (2020). Estimation of potato yield using satellite data at a municipal level: a machine learning approach. International Journal of Geo-Information, 9, 343.

Unamunzaga, O., Aizpurua, A., Artetxe, A., Besga, G., Castroviejo, L., Blanco, F., de la Llera, I., Ramos, L., & Astola, G. (2018). Asistencia técnica para la caracterización agrológica del suelo rústico del municipio de Vitoria-Gasteiz (Technical assistance for the agrological characterisation of rural land in the municipality of Vitoria-Gasteiz). https://www.vitoria-gasteiz.org/docs/j34/catalogo/01/83/clasificacionagrologica18memoria.pdf (accessed January 2025)

Uribeetxebarria, A., Castellón, A., & Aizpurua, A. (2023). Optimizing Wheat Yield Prediction Integrating Data from Sentinel-1 and Sentinel-2 with CatBoost Algorithm. Remote Sensing, 15(6), 1640. https://doi.org/10.3390/rs15061640

Vannoppen, A., & Gobin, A. (2022). Estimating yield from NDVI, weather data, and soil water depletion for sugar beet and potato in Northern Belgium. Water, 14(8), 1188.