

Predictive Modelling of Maize Yield Using Sentinel 2 NDVI

Andrea Soccolini and Marco Vizzari(✉)

Department of Agricultural, Food, and Environmental Sciences, University of Perugia, 06121 Perugia, Italy
{andrea.soccolini,marco.vizzari}@unipg.it

Abstract. Accurate yield prediction is essential for precision agriculture as it enables farmers to optimize their inputs and manage their resources more efficiently, ultimately leading to higher profitability and sustainable farming practices. The aim of this research is to verify whether satellite images can be a valid tool for predicting yield in summer maize (*Zea mays* L.). The study was conducted in 2022 in two adjacent fields in Umbria, Italy. The approach adopted is based on the use of NDVI (Normalized Difference Vegetation Index) data from Sentinel 2 (S2) satellites and yield data from combine harvesters to estimate the grain yield of maize. The NDVI data, derived from S2 images acquired at different growth cycle stages, were used for a Principal Component Analysis (PCA). Subsequently, a multiple linear regression (MLR) model was created. The results show a significant correlation between NDVI and grain yield, suggesting that the central stages of the maize phenological cycle are more indicative for this purpose. Further experiments are necessary to confirm the effectiveness of the proposed approach, using more specific vegetation indices or including soil and climate data in the procedure, thus obtaining more comprehensive responses.

Keywords: precision agriculture · yield mapping · NDVI · PCA · multiple regression linear models

1 Introduction

Rational crop management must consider the spatial and temporal variability of the plots, properly modulating the main agronomic practices based on plant and soil needs [1]. Considering this variability, Precision Agriculture (PA) allows for optimizing the distribution of agronomic inputs, reducing costs and environmental impacts, and maximizing crop production. Various technologies are available in PA to ensure timely crop interventions, simplifying their management. These include Global Navigation Satellite Systems (GNSS), Remote Sensing (RS), Variable Rate Technologies (VRT), and yield mapping, which integrate perfectly with Geographic Information Systems (GIS) and predictive models (see, e.g., [2–4]).

Remote sensing represents a readily employable tool for monitoring vegetation, given the ability of sensors to investigate different regions of the electromagnetic spectrum; the most important ones are the red and green bands in the visible region, the red edge,

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
O. Gervasi et al. (Eds.): ICCSA 2023 Workshops, LNCS 14107, pp. 327–338, 2023.
https://doi.org/10.1007/978-3-031-37114-1_22

and those in the near and mid-infrared. Through algebraic formulas, these bands can be combined to obtain numerous vegetation indices (VI), capable of providing information on vegetation status (see, e.g., [5–8]). Among these, the Normalized Difference Vegetation Index (NDVI) has found the widest diffusion for its simplicity of calculation and interpretation. It represents the normalized difference between red (RED) and near-infrared (NIR) reflectances and has a range of variation between −1 and 1, although in vegetated areas, it is between 0 and 1 [9]. The capacity of this index to estimate some vegetation properties, such as LAI, biomass, and chlorophyll concentration, is widely demonstrated (see, e.g., [10–13]).

The Sentinel-2 satellites (S2) are equipped with a multispectral instrument (MSI) that generates images featuring a spectral resolution of 13 bands, spatial resolutions ranging from 10 to 60 m, a radiometric resolution of 12 bits, and an average temporal resolution of 5 days. They represent a valuable support tool for precision agriculture (see, e.g., [14–17]). The RED (B4) and NIR (B8) S2 bands, used to calculate the NDVI, have a spatial resolution of 10 m.

Yield mapping in PA is carried out by installing GNSS and yield monitoring systems on harvesters. In this way, each yield map can associate the georeferenced grain quantities collected with rectangular areas of different widths. The harvesting surface is expressed as the product of the effective working width of the harvester and the speed of advancement (distance traveled by the harvester/time interval) and is instantly compared to the mass flow of the grain to determine yield [18, 19]. The reliability of a yield map is significantly influenced by the calibration of the numerous sensors mounted on the harvester. Although these operations are sometimes complex and a source of errors [20], the method is effective in providing information on productive variability within plots.

Crop yield prediction is a highly investigated topic, achievable through various methods such as direct field observation during the growing season, crop simulation models, remote sensing, or their combinations [21]. This operation carried out before the end of the growing season, can allow for corrective interventions on crops, improving final yield. However, yield is influenced by multiple crop-specific parameters, climatic conditions, and management choices that complicate the creation of an effective model [22].

Maize is one of the most important cereal crops in the world, providing food, feed, and fuel for millions of people. Accurately forecasting maize yields is, therefore, critical for ensuring food security and economic stability. Remote sensing has been shown to be an effective tool for predicting maize yields, as it allows for the collection of data on various crop parameters, such as plant height, leaf area index, and chlorophyll content [23–25].

In this context, this study aimed to demonstrate the effectiveness of yield prediction in summer corn of a linear model based on the use of Sentinel-2 images.

2 Materials and Methods

2.1 Study Areas

The experimentation was conducted in a single vegetative season (summer 2022) on two adjacent fields belonging to the "Foundation for Agricultural Education in Perugia" (FIA): both plots are located in the Tiber Valley, near Deruta, Umbria, Italy. Field 1 (160 m a.s.l., 42°57′12″ N, 12°22′50″ E) has an area of 21 hectares (Fig. 1, Field 1), while Field 2 (160 m a.s.l., 42°57′08″ N, 12°23′05″ E) has an area of 31 hectares (Fig. 1, Field 2). Both fields are characterized by a texture gradient from east to west, mainly due to the proximity to the Tiber River. The area has a Mediterranean climate, generally alternating between a hot and dry season (May-September) and a cold and rainy season (from October-November to March-April). The experimental crop was summer corn (Zea mays L.) var. P9241, with an FAO maturity class of 300 and a cycle of 95 days, was sown at the end of May. The previous crop was pea (Pisum sativum Asch and Gr.). Both fields were managed using the same cultivation technique, following conservation agriculture principles.

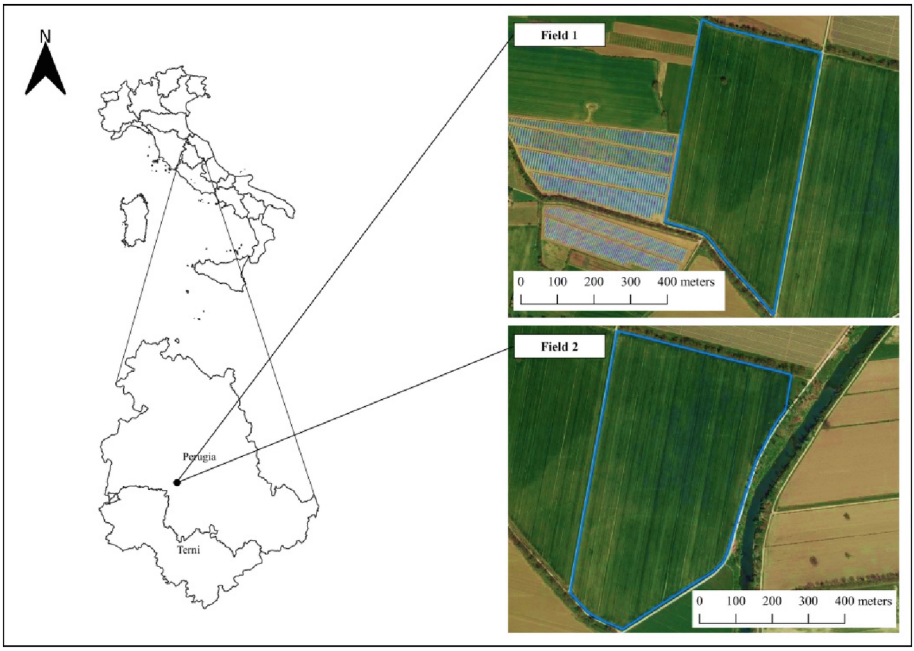

Fig. 1. The geographical location of the study areas.

2.2 NDVI Data

For both plots, 13 Sentinel-2 level 2A images were acquired throughout the entire maize phenological cycle, with a temporal distance between two consecutive images ranging from a minimum of 7 to a maximum of 20 days. The images were downloaded from the Sentinel EO Browser [26]. Using QGIS software version 3.10 64-bit [27], NDVI was calculated from bands 4 and 8 using the "Raster Calculator" tool. For each of the 13 dates, the mean NDVI value for each plot was also calculated within QGIS using the "Statistic for Polygons" tool. A field survey was conducted for each satellite date to observe the crop phenological stage (Table 1).

Table 1. List of 13 images from Sentinel-2 (S2), associated with the relative phenological phase of the corn and the average NDVI value for the two experimental fields (F1 and F2). The days after sowing (DAS) are also reported.

DAS	Date	Phenological phase	F1	F2
0	26 May 2022	Sowing	0.15	0.15
7	02 June 2022	Germination	0.09	0.08
17	12 June 2022	Emergence (VE)	0.12	0.10
24	19 June 2022	4th leaf (V4)	0.13	0.10
32	27 June 2022	6th leaf (V6)	0.31	0.23
39	04 July 2022	9th leaf (V9)	0.44	0.36
54	17 July 2022	Tasseling (Vt)	0.66	0.64
64	27 July 2022	Silking (R1)	0.79	0.77
72	06 August 2022	Blister (R2)	0.80	0.78
89	23 August 2022	Milk (R3)	0.74	0.75
109	12 September 2022	Dough (R4)-Dent (R5)	0.59	0.62
119	22 September 2022	Physiological maturity (R6)	0.41	0.46
132	05 October 2022	Physiological maturity (R6)- Commercial maturity	0.30	0.43

2.3 Yield Data Collection and Pre-processing

For both plots, the harvest was conducted between October 6 and 8, 2022, with the crop in an intermediate stage between physiological and commercial maturity. For the harvest, a Claas Lexion 750 combine harvester was used, equipped with a Topcon YieldTrakk system that processes yield and moisture data through an optical sensor that measures the mass flow of the grain and moisture sensors. Initial calibration was performed in the field on the combine harvester to adjust the effective working width and measure the unit weight of the grain, which was used by the system to convert the measured mass flow (l/s) into Mg. The system produced georeferenced maps in ESRI polygon shapefile format. Only for Field 2, during the harvest, a failure to acquire grain mass and moisture

data were found for about 3 ha of surface area, which explains the missing polygons in its yield map.

A data cleaning technique was applied according to the following steps: (1) all values outside the general mean ± 2.5 * standard deviation interval were eliminated, considered statistically anomalous values [28]; (2) all yield monitoring points were removed within 20 m from the plot boundaries, to remove border effects and yield detection errors at the end of the plot [29]. The data is reported in Mg/ha.

To produce a more easily interpretable yield map, the output shapefile containing geo-referenced yield data was converted into a raster with a spatial resolution of 10 m. This step was carried out using the "Multilevel b-spline interpolation" algorithm provided in the QGIS software. Finally, a circular smoothing filter with a radius of three was applied to further reduce the irregularities presented by the surface.

2.4 Yield Predictive Modeling

The methodology proposed in this experiment involves developing a yield prediction model using Multiple Linear Regression (MLR), an algorithm that falls within the class of Supervised Machine Learning. It represents an extension of the linear regression model, whose objective is to predict the values a dependent variable assumes based on the knowledge of those observed on several independent variables. Specifically, the independent variables are represented by all the vigor maps, seeking to evaluate the ability of this factor to predict maize production and understand the most suitable phases of the cycle for this purpose. Considering the two plots as a single study area, a regular grid of points consisting of 1021 samples was created, overlapping it with all NDVI and yield maps to extract the values of the corresponding pixels. After constructing a value table, a Principal Component Analysis (PCA) was applied to eliminate multicollinearity between variables. The most significant principal components were selected as predictors in multiple linear regression (MLR) models. The model's goodness of fit was evaluated using the analysis of the significance and the coefficient of determination (R^2). Training and a validation set were created using 80% and 20% of the initial dataset for MLR, respectively. The comparison analysis between the sets and the model was based on the values provided by the root-mean-square error (RMSE) and the relative root-mean-square error (RRMSE). All statistical analysis was performed using the R statistical software (version 4.2.2) [30].

3 Results and Discussion

3.1 NDVI Trends and Phenological Cycle

The evolution of NDVI for the two plots during the maize growth cycle is shown in Fig. 2. For both plots, the NDVI trend is consistent with maize cultivation in temperate latitudes: a rapid increase in July until reaching the peak around 0.8 in early August, followed by a rather rapid decrease from late August to early October. NDVI is very sensitive to color change caused by the appearance of vegetation cover, but during the first phase of the growth cycle, in the days following crop emergence, maize seedlings do not provide any significant optical response, justifying values $\leq$0.3 in May–June [31]. The rapid and significant increase in NDVI in July is directly related to the height and leaf biomass development of maize, reaching a peak around 0.8 in early August, corresponding to the first reproductive stages. Due to the saturation phenomenon, NDVI cannot detect significant differences in vigor during this phase. The progressive decrease of NDVI from late August to early October is consistent with maize maturity, in which the crop, in addition to a chromatic change of leaves, undergoes drying, varying the response in terms of RED and NIR reflectance. The NDVI trend of the field shows an initial growth phase and a slightly faster maturation phase than field 2, however, the average values considered for each crop stage do not show significant differences.

A more detailed analysis can be conducted by considering the graphical evolution of NDVI from crop emergence to harvest (Fig. 3). In both cases, greater development is observed along the southern slope, but field 2 shows a delay in vegetative growth almost throughout its entire surface compared to field 1. Referring to the date prior to harvest, despite the previously highlighted very limited deviation in NDVI, the crop scenario appears very diversified: while field 1 shows low vigor over the entire surface, field 2 shows a vegetative gradient from east to west, probably due to the greater humidity in the air coming from the Tiber River, which slowed down plant senescence.

3.2 Yield Mapping

The raw yield map of the two plots is shown in Fig. 4a, while the continuous yield map, obtained with the procedure described earlier and used for both visual purposes and predictive analysis, is displayed in Fig. 4b. By visually comparing these maps with the NDVI maps, a possible relationship between yield and vegetative vigor during the central phases of the growth cycle is noted, especially in areas with low NDVI values. In addition, the interpolation applied to the yield data generated maps that tend to overestimate minimum values and underestimate maximum ones.

3.3 Yield Forecasting

The strong correlation between NDVI and yield, especially during the growing stages of plant development, made it necessary to apply PCA to remove collinearity between the considered variables. The first principal component (PC) confirmed a close relationship with the central stages of the cycle, while the second PC demonstrated an inverse relationship between vigor and yield during the maturation phases. The first four PCs were included in the multiple linear regression model, resulting in a high correlation (R^2 0.84) and an extremely low P-value due to many considered samples. A subsequent step was the creation of training and validation sets to develop and test the model based on the selected predictors. The MLR on the training set showed an RMSE of 0.602 t/ha and an RRMSE of 7.169% (Fig. 5a), while these values slightly decreased when MLR was applied on the validation set (Fig. 5b) (RMSE 0.577 t/ha; RRMSE 6.843%).

The results confirm the strong relationship between intermediate phenological stages and maize grain production and the reliability of simple multiple linear regression models for yield prediction. However, more advanced predictive models could be developed and tested by integrating the crop's climatic, pedological, and biophysical data into the vigor index [23, 32]. Additionally, NDVI could be compared with responses provided by more specific indices referring to specific growth stages (e.g., [24, 33–35]).

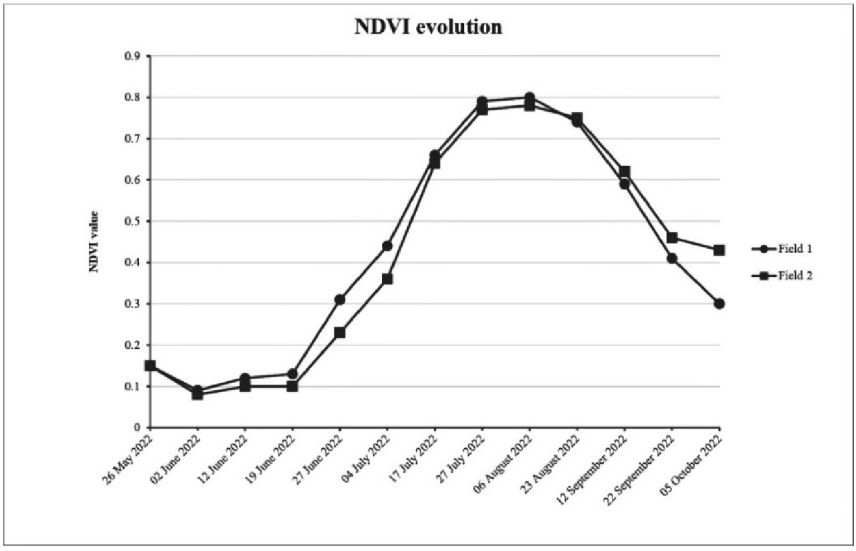

Fig. 2. NDVI's trend of the two experimental fields calculated from Level 2A Sentinel-2 image.

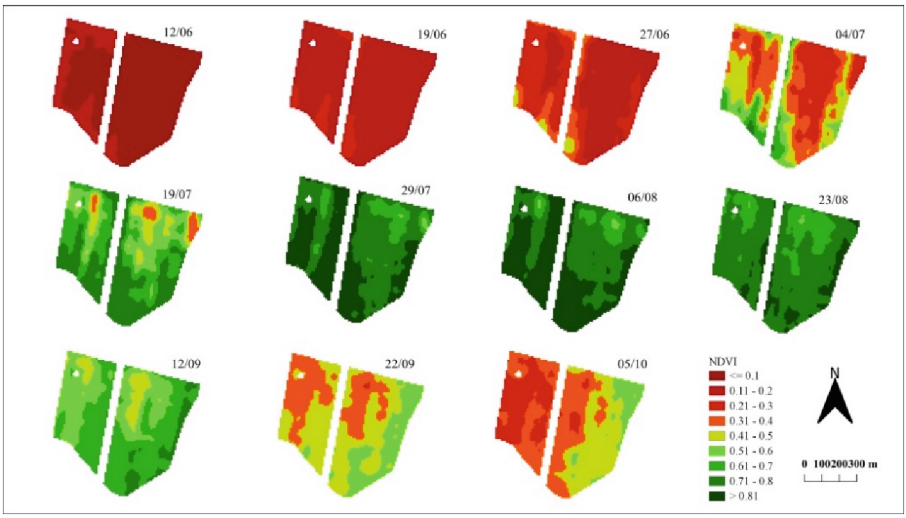

Fig. 3. NDVI's evolution of the two experimental fields, from the crop's emergence to harvest.

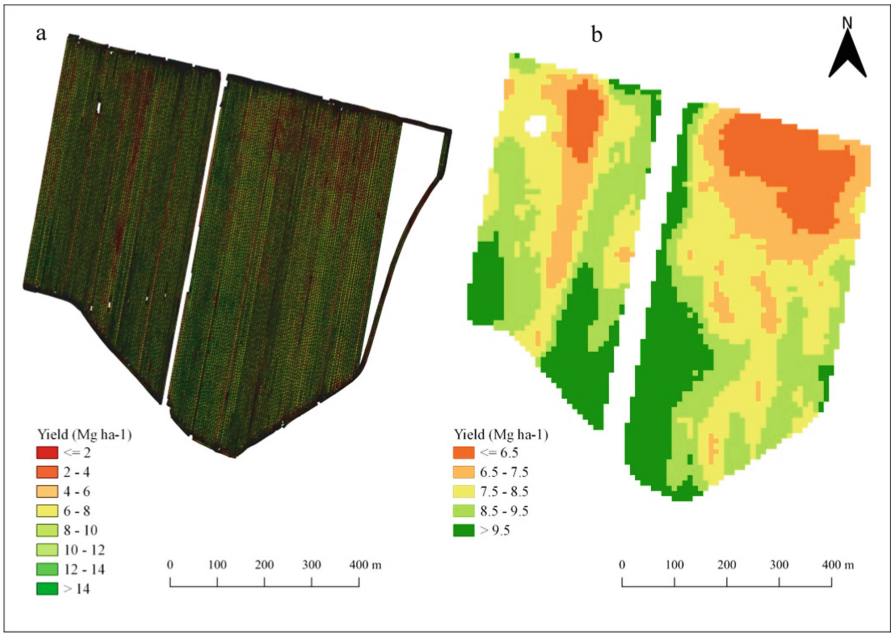

Fig. 4. Yield maps of the two experimental fields, raw (a) and processed data (b).

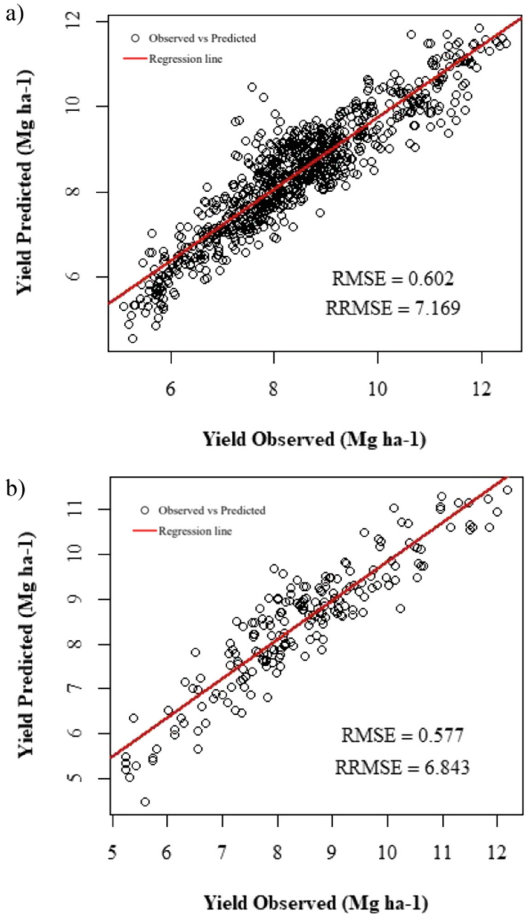

Fig. 5. Comparison graphs between MLR applied on training and validation sets.

4 Conclusions

Our study analyzed the effectiveness of Sentinel 2 images in describing the phenological cycle of maize and their possible use as a predictive model for grain yield. The results confirmed the ability of NDVI to reconstruct the maize growth cycle, showing values consistent with the literature research. Vegetation indices derived from satellite remote sensing can, therefore, successfully identify the spatial variability of plant vigor. However, its causes must necessarily be inferred with a field evaluation. NDVI also proved an effective predictor of maize yield, adapting well to the proposed multiple linear regression model. Specifically, the central phenological stages of maize appeared to be the most representative, in line with bibliographic studies carried out with more complex models for the number and heterogeneity of the variables considered. Further experiments are needed to strengthen the results proposed with this approach, such as including additional parameters in the analysis to obtain more complete responses.

Acknowledgments. The authors wish to thank the farm "Fondazione per l'Istruzione Agraria" (Casalina di Deruta, province of Perugia, Italy) and the farm's Technical Director, Dr. Mauro Brunetti, for its valuable support during all the experimental stages.

References

1. Vecchio, Y., Agnusdei, G.P., Miglietta, P.P., Capitanio, F.: Adoption of precision farming tools: the case of Italian farmers. Int. J. Environ. Res. Public Heal. **17**, 869 (2020). https://doi.org/10.3390/IJERPH17030869
2. Brisco, B., Brown, R.J., Hirose, T., Mc Naim, H., Staenz, K.: Precision agriculture and the role of remote sensing: a review. Can. J. Remote Sens. **24**, 315–327 (2014). https://doi.org/10.1080/07038992.1998.10855254
3. Liaghat, S., Balasundram, S.K.: A review: the role of remote sensing in precision agriculture (2010). http://psasir.upm.edu.my/12390/
4. Yuan, J., Ji, W., Feng, Q., Vrchota, J., Pech, M., Švepešová, I.: Precision agriculture technologies for crop and livestock production in the Czech Republic. Agriculture **12**, 1080 (2022). https://doi.org/10.3390/AGRICULTURE12081080
5. Xue, J., Su, B.: Significant remote sensing vegetation indices: a review of developments and applications. J. Sens. **2017** (2017). https://doi.org/10.1155/2017/1353691
6. Glenn, E.P., Huete, A.R., Nagler, P.L., Nelson, S.G.: Relationship between remotely-sensed vegetation indices, canopy attributes and plant physiological processes: what vegetation indices can and cannot tell us about the landscape. Sensors **8**, 2136–2160 (2008). https://doi.org/10.3390/S8042136
7. Basso, B., Cammarano, D., De Vita, P.: Remotely sensed vegetation indices: theory and applications for crop management. Ital. J. Agrometeorol. **53**, 36–53 (2004)
8. Messina, G., Peña, J.M., Vizzari, M., Modica, G.: A comparison of UAV and satellites multispectral imagery in monitoring onion crop. An application in the 'Cipolla Rossa di Tropea' (Italy). Remote Sens. **12**, 3424 (2020). https://doi.org/10.3390/rs12203424
9. Huang, S., Tang, L., Hupy, J.P., Wang, Y., Shao, G.: A commentary review on the use of normalized difference vegetation index (NDVI) in the era of popular remote sensing. J. For. Res. **32**, 1–6 (2021). https://doi.org/10.1007/S11676-020-01155-1/FIGURES/2
10. Bajocco, S., et al.: On the use of NDVI to estimate LAI in field crops: implementing a conversion equation library. Remote Sens. **14**, 3554 (2022). https://doi.org/10.3390/RS14153554/S1
11. Mihai, H., Florin, S.: Biomass prediction model in maize based on satellite images. AIP Conf. Proc. **1738**, 350009 (2016). https://doi.org/10.1063/1.4952132
12. Gitelson, A.A., Merzlyak, M.N.: Remote sensing of chlorophyll concentration in higher plant leaves. Adv. Sp. Res. **22**, 689–692 (1998). https://doi.org/10.1016/S0273-1177(97)01133-2
13. Benincasa, P., et al.: Reliability of NDVI derived by high resolution satellite and UAV compared to in-field methods for the evaluation of early crop N status and grain yield in wheat. Exp. Agric. **54**, 1–19 (2017). https://doi.org/10.1017/S0014479717000278
14. Segarra, J., Buchaillot, M.L., Araus, J.L., Kefauver, S.C.: Remote sensing for precision agriculture: Sentinel-2 improved features and applications **10**, 641 (2020)
15. Vizzari, M.: PlanetScope, Sentinel-2, and Sentinel-1 data integration for object-based land cover classification in google earth engine. Remote Sens. **14**, 2628 (2022). https://doi.org/10.3390/RS14112628
16. Santaga, F.S., Agnelli, A., Leccese, A., Vizzari, M.: Using Sentinel-2 for simplifying soil sampling and mapping: two case studies in Umbria, Italy. Remote Sens. **13**, 3379 (2021). https://doi.org/10.3390/RS13173379

17. Santaga, F., Benincasa, P., Vizzari, M.: Using Sentinel 2 data to guide nitrogen fertilization in central Italy: comparison between flat, low VRT and high VRT rates application in wheat. In: Gervasi, O., et al. (eds.) ICCSA 2020. LNCS, vol. 12253, pp. 78–89. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58814-4_6
18. Arslan, S., Colvin, T.S.: Grain yield mapping: yield sensing, yield reconstruction, and errors. Precis. Agric. **3**, 135–154 (2002). https://doi.org/10.1023/A:1013819502827
19. Santaga, F.S., Benincasa, P., Toscano, P., Antognelli, S., Ranieri, E., Vizzari, M.: Simplified and advanced Sentinel-2-based precision nitrogen management of wheat. Agronomy **11**, 1156 (2021). https://doi.org/10.3390/agronomy11061156
20. Colvin, T.S., Arslan, S.: A review of yield reconstruction and sources of errors in yield maps. In: Proceedings of the 5th International Conference on Precision Agriculture, Bloomington, Minnesota, USA, 16–19 July 2000, pp. 1–13 (2000)
21. Basso, B., Cammarano, D., Carfagna, E.: Review of crop yield forecasting methods and early warning systems. In: Proceedings of the First Meeting of the Scientific Advisory Committee of the Global Strategy to Improve Agricultural and Rural Statistics (2013)
22. Paudel, D., et al.: Machine learning for large-scale crop yield forecasting. Agric. Syst. **187** (2021). https://doi.org/10.1016/j.agsy.2020.103016
23. Meng, L., Liu, H., Ustin, S.L., Zhang, X.: Predicting maize yield at the plot scale of different fertilizer systems by multi-source data and machine learning methods. Remote Sens. **13**, 3760 (2021). https://doi.org/10.3390/RS13183760
24. Yang, B., Zhu, W., Rezaei, E.E., Li, J., Sun, Z., Zhang, J.: The optimal phenological phase of maize for yield prediction with high-frequency UAV remote sensing. Remote Sens. **14**, 1559 (2022). https://doi.org/10.3390/RS14071559
25. Bonciarelli, U., et al.: Long-term evaluation of productivity, stability and sustainability for cropping systems in Mediterranean rainfed conditions. Eur. J. Agron. **77**, 146–155 (2016). https://doi.org/10.1016/j.eja.2016.02.006
26. Sentinel Hub EO Browser. https://apps.sentinel-hub.com/eo-browser/?zoom=10&lat=41.9&lng=12.5&themeId=DEFAULT-THEME&toTime=2023-04-12T11%3A42%3A50.129Z. Accessed 12 Apr 2023
27. Quantum GIS Development Team: Quantum GIS Geographic Information System (2020). http://qgis.osgeo.org
28. Vizzari, M., Santaga, F., Benincasa, P.: Sentinel 2-based nitrogen VRT fertilization in wheat: comparison between traditional and simple precision practices. Agronomy **9**, 278 (2019). https://doi.org/10.3390/agronomy9060278
29. Vega, A., Córdoba, M., Castro-Franco, M., Balzarini, M.: Protocol for automating error removal from yield maps. Precis. Agric. **20**(5), 1030–1044 (2019). https://doi.org/10.1007/s11119-018-09632-8
30. R Development Core Team: R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (2020)
31. Rolle, M., Tamea, S., Claps, P., Ayari, E., Baghdadi, N., Zribi, M.: Analysis of maize sowing periods and cycle phases using Sentinel 1&2 data synergy. Remote Sens. **14**, 3712 (2022). https://doi.org/10.3390/RS14153712/S1
32. Baez-Gonzalez, A.D., et al.: Large-area maize yield forecasting using leaf area index based yield model. Agron. J. **97**, 418–425 (2005). https://doi.org/10.2134/AGRONJ2005.0418
33. Barzin, R., Pathak, R., Lotfi, H., Varco, J., Bora, G.C.: Use of UAS multispectral imagery at different physiological stages for yield prediction and input resource optimization in corn. Remote Sens. **12**, 2392 (2020). https://doi.org/10.3390/RS12152392

34. Croci, M., Impollonia, G., Meroni, M., Amaducci, S.: Dynamic maize yield predictions using machine learning on multi-source data. Remote Sens. **15**, 100 (2023). https://doi.org/10.3390/RS15010100/S1
35. Kayad, A., Sozzi, M., Gatto, S., Marinello, F., Pirotti, F.: Monitoring within-field variability of corn yield using Sentinel-2 and machine learning techniques. Remote Sens. **11**, 2873 (2019). https://doi.org/10.3390/RS11232873