Time-Series Of Remotely Sensed Data To Estimate Wheat Grain Yield At Field Level

EARSeL Manchester 2024
Abstract
Corresponding Author: nutini.f@irea.cnr.it

Francesco Nutini[1], Mattia Scarpantoni[1], Federico Filipponi[2], Mirco Boschetti[1]

[1] Institute for Electromagnetic Sensing of the Environment, National Research Council of Italy, Milano, Italy

[2] Institute of Environmental Geology and Geoengineering, National Research Council of Italy, Roma, Italy

Keywords: LAI time-series, agriculture, yield estimation, phenology, Sentinel-2

Challenge

Trying to foretell crop yield has been a key issue since the first farming systems were set, it's even more compelling now because of climate change (increasing uncertainty in the system) and a growing population (forcing more accurate and timely estimates).

Data acquired by satellite earth observation systems give us a unique opportunity to observe crop growth at adequate spatial resolution with a quasi-weekly temporal frequency. Indeed, remote sensing imagery has been extensively used in data-driven approaches to estimate yield employing a variety of approaches: parametric regressions, machine learning algorithms, and statistical models. Past scientific literature highlights some important advice for this study such as the significant exploitation of time-series of bio-physical parameters rather than vegetation indices.

In this context, the main goal is to estimate the yield of winter cereals at field level exploiting a multi-year dataset of ground data, meteorological variables, and satellite derived time series of Leaf Area Index (LAI) with a non-parametric regressive approach.

Methodology

The area of interest (AOI) of this study is a big farm (~3800 ha) in Ferrara province (Italy) where a data set of 183 ground measurements of winter crop yield (spelt, soft and durum wheat and barley) were collected at the end of 4 cropping seasons (2020-2023). For the same period, Sentinel-2 data (S2) were downloaded and LAI maps were computed using a Gaussian Process Regression (GPR) algorithm already calibrated using a multi-site, crop, and year data set. Yield data represent the independent variable, while LAI data is the most important dependent variable considered. Other variables exploited for yield estimation are: i) daily rainfall and temperature estimated by Copernicus Climate Data ERA5; ii) water stress index (Shortwave Infrared Water Stress Index - SIWSI) derived from S2; and iii) soil organic carbon data provided by a farm's soil map.

LAI time-series extracted from the location of the sampling sites were exploited to automatically compute phenological metrics and the occurrence of key crop development stages (Figure1a). These dates were exploited to compute regressor from LAI time-series as well as ERA5 and SIWSI data (e.g. cumulated rainfall before flowering, LAI at peak of season). Moreover, other regressors were computed on fixed periods according to cropping calendar, in order to check the actual contribution given by remotely sensed phenological metrics. These regressors are exploited in a data-driven machine learning approach using multiple random forest (RF) trials. Results are analysed in term of regressors contribution and importance in yield estimation and the best model is applied and compared with a yield map.

Expected Results

Four phenological metrics were computed getting key dates of vegetation cycle (see Figure1a). Occurrences of LAI peaks, corresponding to booting stage according to available field observations, were detected between April and mid-May, coherently with these phenological stages in the AOI. More than 20 regressors for LAI, meteorological, and SIWSI were derived for each sample thanks to these dates.

A preliminary test with linear regression shows that the most correlated regressors with final yield are LAI value at peak of season and seasonal time integrated LAI, coherently with past similar work. Generally speaking, all top correlated regressors are derived from LAI data; still, correlation is weak (maximum r^2 0.39).

First tests conducted with RF algorithm show that better results can be obtained using all LAI regressors (r^2 0.60, RMSE 1.50). Best results are obtained if LAI-derived indicators are exploited together with meteorological ones (r^2 0.69, RMSE 1.32). In this case, regressors derived from meteorological data (such as cumulated rainfall during vegetative phase – A to C points in Figure1a) are among the most important information in modeling yield. No significant improvement is given when exploiting soil information (coherently with literature), crop type (i.e. model is not crop-dependant) and water stress as derived by SIWSI. Regressors obtained on fixed windows on calendar dates slightly outperform the ones from phenological windows. This finding is probably motivated by two reasons: limited variability in season length due to geographically restricted area and uncertainty in start of season date retrieval from early stages LAI estimation.

Outlook For The Future

Next activities will be first focused on validating LAI estimation at the very beginning of cropping season. Field campaigns are being performed for this scope. Moreover, a comparison with LAI calculated on Biophysical processor toolbox in SNAP will be performed and phenological metrics will be computed on this new dataset.

RF trials will be performed using the new phenological dates and adding new meteorological variables from ERA5 (e.g. evapotranspiration). On top of that, accuracy metrics were obtained in cross-validation so far, but a splitting of field data in calibration and validation datasets should be done. It is well known that data acquired from same field/season can present a high degree of autocorrelation. The new "Best model" will be applied on the whole farm, as shown by a preliminary test in Figure1b, and field level estimates and spatial patterns analysed and discussed according to farm data.

Observing yield variability in their spatial and temporal dimension is a prior activity to understand phenomena and to lead thoughtful management of cropping systems. Once calibrated, our view is to apply the model to the whole Po valley (Italy) to spot which districts faced yield loss during a severe drought event that impacted north of Italy in 2022.

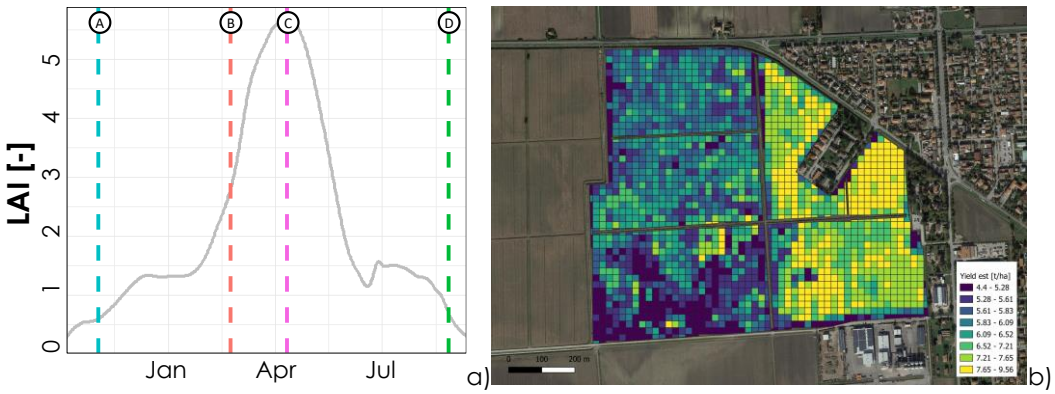

Figure1 (a) example of LAI time series and computed phenological metrics: start (A) and end of season (D), LAI peak (C), a mid-point between start and peak (B). (b) Preliminary test of estimated yield map using the calibrated random forest model.