# Proocedings of the 51st National conference of the Italian Society for Agronomy


Botanical garden – University of Padova

19 - 21 September 2022

# ORAL PRESENTATIONS

# A Machine Learning Modelling Framework For Durum Wheat Yield Forecasting In Central-Southern Italy

Marco Fiorentini[1], Calogero Schillaci[2*], Michele Denora[3], Stefano Zenobi[1], Paola Deligios[4], Roberto Orsini[1], Rodolfo Santilocchi[1], Michele Perniola[3], Luca Montanarella[2] and Luigi Ledda[1]

[1] Dep. D3A, Marche Polytechnic University, IT
[2] European Commission, Joint Research Centre (JRC), Ispra, IT
[3] Department of European and Mediterranean Cultures, Environment, and Cultural Heritage, Univ. Basilicata IT
[4] Department of Agricultural Sciences, Univ. Sassari, IT

**Introduction**

Durum wheat (*Triticum turgidum* subsp. durum Desf.) is the most economically important cereal crop to produce dried pasta. Forecast of crop yield is one of the most critical research areas in crop science, which allows the development of the decision support systems, optimising nitrogen (N) fertilization and food safety Filippi et al., 2019; Shahhosseini et al., 2020; van der Velde et al., 2019). Several authors have focused on the use of on-farm or remote-sensing data sources. Other studies have focused on publicly available datasets but were conducted on one field site or neglected the potential benefit when N fertilizer is commonly applied. Moreover there are two ways to forecast yield which are deterministic based methods (Basso et al., 2011; Valkama et al., 2020) (DSSAT, SALUS, ARMOSA) and on the other hand, stochastic methods based on field and remote sensing offer a new avenue to find a relationship between the biotic and environmental predictors and can be deployed in vast areas. In this work, four independent field experiments on durum wheat in Central-Southern Italy were used to build scalable machine learning (ML) models to predict the durum wheat yield using fertilization N management, pedo-climatic and remote sensing data.

**Materials and Methods**

The Four Italian experimental sites have been considered, two located in the Marche region and two in the Basilicata region. All sites have a different experimental design where durum wheat was grown for several years. To follow the whole crop development, three phenological phases have been defined to perform the UAV multispectral image acquisition. Multispectral images were acquired at crop tillering , stem elongation, and anthesis to compute the Normalized Difference Vegetation Index (NDVI) and Normalized Difference Red Edge index (NDRE). At crop maturityfor each experimental site, test areas were randomly selected and georeferenced by using the Leica Zeno 20 (Fiorentini et al., 2019). The grain yield (t dry matter $ha^{-1}$) has been measured in each test area harvesting 1 m long-row. Moreover before the sowing operation it was measured several soil chemical data such as soil texture, soil organic carbon, potassium, N, and the carbon/nitrogen ratio (C/N), were measured for each field under analysis. The data provided to the different machine learning (ML) algorithms are comprehensive in fertilization management, meteorological data such as temperature and precipitation, physical and chemical properties of the soil and vegetation indexes (VIs). Five different algorithms were compared, (1) Linear Model (LM) as the benchmark algorithm, (2) Support Vector Machine, (3) K Nearest Neighbors, (4) Random Forest and (5) Stochastic Gradient Boosting as (ML) methods.
Model transferability was tested by using a dataset that the models were not used during the training procedure (Chollet, 2021). Moreover, the varImp function of caret R package (Kuhn, 2008) was used to visualize the variable importance to describe which covariate contributed most to the construction of the models. Variable importance refers to how much a given model "uses" that variable to make better predictions.

**Results**

The results show that the grain yield obtained from the four sites are different. GBM and RF obtained significantly lower values of RMSE and MAE and considerably higher values of $R^2$ compared to LM, SVM and KNN models during the calibration process. The results demonstrate that the GBM and RF models performed better than the other models evaluated in this work, with an RMSE of 0.48 t $ha^{-1}$ and 0.47 t $ha^{-1}$, respectively, while the LM, SVM and KNN obtained an RMSE respectively of 0.68, 0.69 and 0.65 t $ha^{-1}$. We have evaluated the models with two different datasets, referred one to the Marche region (Marche 1) and the other one to the Basilicata region (Basilicata 2).The results demonstrate that the GBM model performed better than the other models by obtaining significantly lower values of RMSE and MAE and considerably higher values of $R^2$ compared to other models during the evaluation phase. While regarding the variable importance analysis, it was observed that the N covariate was the most important feature in all the models except for the linear model. Evaluating the feature importance of the ML models, the three most important variables in descending order are

N, precipitation, and temperature. While considering the feature importance of the best model, such as the GBM model, the five most important features are N Management, Precipitation, Temperature, N soil content and NDVI ZS22.

## Conclusions

In this work, we explore the possibility of combining several different source data, such as N management, pedo-climatic and remote sensing data of four Italian experimental sites to train and test five machine-learning algorithms to predict the durum wheat yield.

The GBM was the best ML algorithm, with a 0.58 RMSE of the test set and a lower error metrics variation between calibration and transferability.

It was observed that N, precipitation, and temperature were the features that helped the most to improve the model's accuracy. So they must be considered in future grain yield ML modelling approaches.

The generated model can be scaled for the Marche and Basilicata region, having the soil samples acquired in field and the multispectral satellite images from the Copernicus program and the meteorological data from NASA.

## Literature


Basso B. et al. 2011. A strategic and tactical management approach to select optimal N fertilizer rates for wheat in a spatially variable field. Eur. J. Agron., 35: 215–222

Chollet F. 2021. Deep learning with Python. Simon and Schuster

Filippi P. et al. 2019. An approach to forecast grain crop yield using multi-layered, multi-farm data sets and machine learning. Precis. Agric., 20: 1015–1029

Fiorentini M. et al. 2019. Nitrogen and chlorophyll status determination in durum wheat as influenced by fertilization and soil management: Preliminary results. PLoS One, 14: 1–16.

Fiorentini M. et al. 2021. Remote and Proximal Sensing Applications for Durum Wheat Nutritional Status Detection in Mediterranean Area. Agriculture, 11: 39–57.

Kuhn M. 2008. Building predictive models in R using the caret package. J. Stat. Softw., 28: 1–26.

Shahhosseini M. et al. 2020. Forecasting Corn Yield With Machine Learning Ensembles. Front. Plant Sci., 11: 1120–1136.

Valkama E. et al. 2020. Can conservation agriculture increase soil carbon sequestration? A modelling approach. Geoderma, 369: 114298.

Van Der Velde M. et al. 2019. Use and relevance of European Union crop monitoring and yield forecasts. Agric. Syst. 168: 224–230.