



Deliverable 4: Comparison of EO versus current weather/location predictions

*Partner in charge: Computomics, SilberGeo*

*Date: 02.03.2021*

*research and innovation programme under grant agreement No 824478.*

*This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 824478.*



|  |  |
| --- | --- |
| **Project name** | **Crop Predictions Take Flight** |
| **Acronym** | **CROPTF** |
| **Start date** | 01.09.2020 |
| **End date** | 31.07.2021 |
| **Project Coordinator** | Computomics |
|  |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Document History** | | | | |
| **Version** | **Issue date** | **Stage** | **Changes** | **Contributor** |
| **V2** | **03.03.2021** | **XXX** | **Split in per-deliverable reports** | **Computomics** |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

|  |
| --- |
| ***Disclaimer*** |
| Any dissemination of results reflects only the author’s view and the European Commission is not responsible for any use that may be made of the information it contains |

|  |
| --- |
| ***Copyright message*** |
| **© PARSEC consortium, 2020**  This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgment of previously published material and of the work of others has been made through appropriate citation, quotation or both. Reproduction is authorised provided the source is acknowledged. |

.

Table of Contents

1. Comparison of EO versus current weather/location predictions 4
   1. Test dataset 4
   2. xSeedScore model training 4
   3. Model evaluation 4
   4. Results 5

# Comparison of EO versus current weather/location predictions

In a field test of new plant varieties, each variety is planted in trial plots in different test locations and often also in different years to assess the performance of the variety in different environments. This helps breeders to get a picture of the performance of a specific genetic makeup in different conditions. It also allows controlling for the effects of unforeseen environmental events like extreme weather occurring in one location that biases test results. This is a valuable approach to assess the new varieties performance, however, it is hindered by both availability of environmental (Earth Observation, EO) data, namely weather and soil characteristics, at the specific trial locations during the growing season and by the mechanism to integrate and analyze EO data together with genomic information.

A comprehensive machine learning (ML) model that aims to predict the performance of plant varieties for agronomic traits has to incorporate both sources. The model can then be asked questions about the expected performance of a plant variety in specific target environments and can help breeders select the most promising varieties.

To improve the breeding process and reduce the time-to-market of new crop varieties for Computomic’s clients (breeding companies), we investigated the incorporation of environmental data like soil and weather into our existing ML solution xSeedScore additionally to the genetic data.

## Test dataset

We used a dataset from an existing cereal crop breeding program from a customer to assess the performance of phenotype prediction of models integrating genetic data and different types of environmental data. The dataset contains genetic profiles of several hundred of plant varieties from a breeding population. Each genetically identical variety was tested in several environments, i.e. in a specific year at a specific location. In total, there are 12 locations and two years in the data set. In each environment, crop yield was measured for the respective varieties.

## xSeedScore model training

xSeedScore generates models that take as input the genetic profile of a variety and the environment (year/growing season and location), and predicts the yield for the variety when grown in the respective environment.

As a baseline, we assigned IDs to the different locations and years, and used these IDs as input together with the genetic profiles to train the model. In the other models, we replaced the IDs by environmental data types: weather data from NOAA, static soil profiles from HWSD, and/or soil moisture data from CDS. Please refer to deliverable reports 1-3 for details.

## Model evaluation

We used a cross-validation approach to evaluate the predictive performance of the models. Briefly, the data set is split into subsets of the same size, e.g. into five subsets. In an iterative fashion, one subset is set aside as the test set, while the other data is pooled and used to train an ML model. This model is then used to predict values on the left-out test set. The predictions can then be compared to the true values to assess the model’s performance. This is repeated so that each subset is left out once. We use Pearson’s correlation coefficient to measure how well the predicted values match the measured ones.

We also left out all data for one of the locations to see if using EO data allows us to improve prediction performance for locations that the model has not seen in its training data. Predicting the performance of plants in environments where they have not been tested can help breeders in designing field trials. It can also help answer questions about which of their varieties will potentially perform well in new environments, for example, different geographic regions, or different weather patterns, which will be important in a changing climate.

## Results

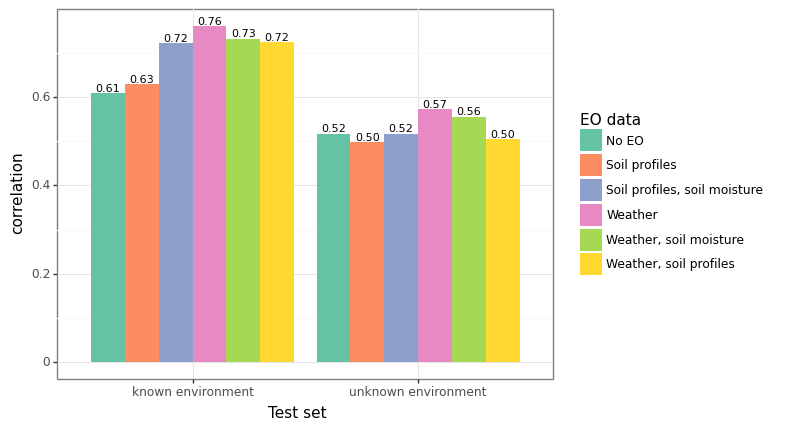


Figure 1: Prediction performance of models using different environmental variables. The y-axis shows Pearson’s correlation between measured phenotype values and predicted values. These are average values over the cross-validation folds. Six different models are compared. All of them use genetic descriptors as input features. The baseline “no EO” model does not use any EO features to describe the growth environments of the plants, only categorial environment IDs. The other models used different combinations of EO descriptors. “Soil profiles” are HWSD soil profiles, “Soil moisture” is Copernicus soil moisture, and “Weather” is NOAA weather. Performance is shown for predictions on two data sets: the cross validation test set, i.e. on plant varieties that the model has not seen before, but the environments they were grown in are included in the training set; and finally, data where the model has not seen the plant varieties nor the environment.

Figures 1 and 2 show results from our experiments. We can see that using EO data significantly improves prediction performance when the environment is known to the model. Using only weather data leads to the largest improvements. The soil profiles of HWSD alone do not improve predictions. One possible reason for this is the low spatial resolution of the data. Soil moisture on its own improves predictions over the baseline, but not when added to weather data. Predictions for environments that were not seen in training were also improved, albeit on a lower level. Here, only weather leads to a modest improvement. Figure 2 shows that the EO data have a significant influence on the predictions of the model, showing their utility and importance.

Our experiments show that using EO data in our predictive models improves their performance for environments where we have training data. This is a valuable result, as it shows that we can help breeders more accurately determine how a plant variety that they have not yet tested in the field would have performed in any of their environments from field trials. This helps them in deciding which varieties to select for the next steps in their breeding program.

The results show that weather data has a significant impact on prediction performance, while soil data from HWSD does not appear to influence the performance. The twelve locations in our data set have only three different HWSD soil profiles, this is probably too little for the model to discriminate between them and identify soil properties that influence phenotype values. It is well possible that for a dataset with more diverse locations the soil descriptors can improve the model performance.

We had hoped that using EO data allows us to improve predictions in locations not seen in the training data available to us. This would allow breeders to check whether planting a specific variety in a new environment that has never been used for field trials in a breeding program is worthwhile. For our dataset, we could show that EO data can help here, even if the improvement in prediction accuracy is modest. It can be assumed that a larger training dataset that covers more and more diverse environments could achieve a model where this task is improved.

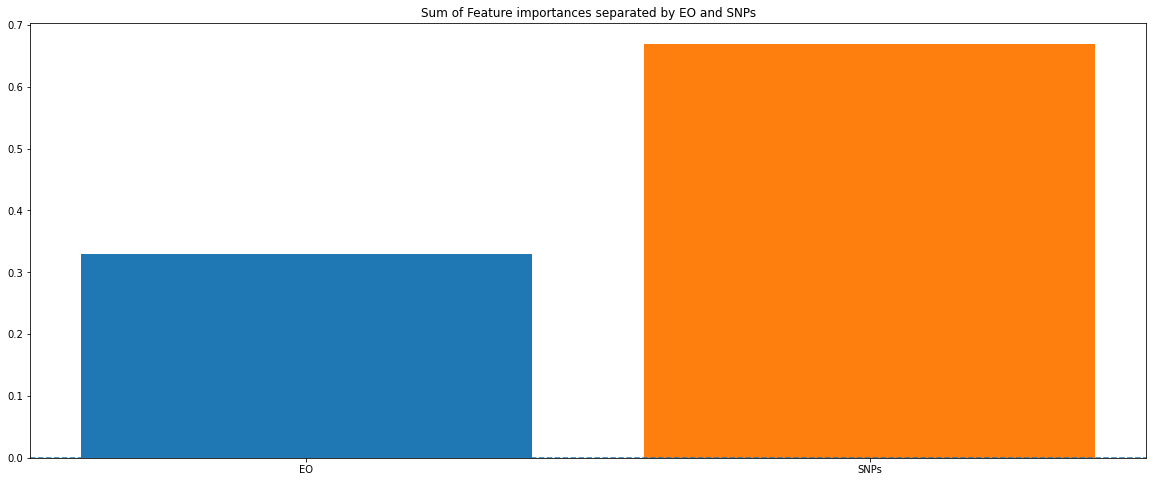


Figure 2: Importance of different data types for predictions. The models we tested have two different types of input features: genetic and EO, which correspond to heritable and non-heritable factors that influence a plant’s phenotype. This figure shows the relative importance of each of the two data types in the “Weather” model, which performed best in our tests. (see figure 2). EO variables are responsible for about one-third of a prediction's results, the other two-thirds come from the genetic component.