European Plant Phenotyping Network

EPPN 2020

D2.6: Integration procedures and software for data from multiple platform and field experiments, with different but overlapping sets of genotypes, across scales of plant organization, traits and management conditions

Fred van Eeuwijk & Emilie Millet (WU)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731013. This publication reflects only the view of the author, and the European Commission cannot be held responsible for any use which may be made of the information contained therein.

Document information

EU Project N°	731013	**Acronym**	EPPN2020
Full Title	European Plant Phenotyping Network 2020		
Project website	www.eppn2020.plant-phenotyping.eu		

Deliverable	**N°**	D2.6	**Title**	Integration procedures and software for data from multiple platform and field experiments, with different but overlapping sets of genotypes, across scales of plant organization, traits and management conditions
Work Package	**N°**	WP2/JRA2	**Title**	Design and analysis of phenotyping experiments across multiple platforms, scales of plant organisation, traits and management conditions

Date of delivery	**Contractual**	31/10/2021 (Month 54)	**Actual**	10/11/2021 (Month 55)
Dissemination level	X	**PU Public, fully open, e.g. web**		
		CO Confidential, restricted under conditions set out in Model Grant Agreement		
		CI Classified, information as referred to in Commission Decision 2001/844/EC.		

Authors (Partner)	WU			
Responsible Author	**Name**	Fred van Eeuwijk	**Email**	fred.vaneeuwijk@wur.nl

Version log			
Issue Date	**Revision N°**	**Author**	**Change**
25/10/2021	0	Fred van Eeuwijk	first version
31/10/2021	1	Emilie Millet	second version
10/11/2021	2	François Tardieu	Reviewed by Coordinator

Executive Summary

D2.6 aimed at making the traits collected in phenotyping platforms useful for plant scientists and breeders interested in the analysis of field datasets. This is, in particular, the prediction and understanding of genotypic main effects (G) for yield in series of field trials, i.e., average genotypic performance, as well as of genotype by environment interactions (G×E), i.e., differential genotypic performance. Phrased in this way, one possible way forward is to enrich existing models for G and G×E by incorporating features related to information obtained on phenotyping platforms. There are two main ways of doing this, factorial regression and multi-trait models. For factorial regression, we include platform phenotyping features, i.e., secondary phenotypes as genotypic covariables in models describing yield in series of field experiments. This means that we try to understand G and G×E as a function of genotypic covariables coming from phenotyping platforms. This approach, which we call factorial regression, was followed, and further developed on field trial data collected in the EU-DROPS project and is reported in Millet et al. (2019). In addition to the incorporation of genotypic covariables into models for yield in series of trials, we can also try to include environmental covariables, which can be both groupings like environmental scenarios and continuous environmental indices. The joint use of genotypic and environmental covariables provides a strong approach for data integration between phenotyping platform data and field trials. The success of this approach depends critically on a good preselection of genotypic and environmental covariables based on physiological insights prior to building a model for yield in the field. Without such a preselection, too many possible models need to be evaluated and the chances of finding a successful model for prediction of yield under G×E are small.

Factorial regression models with genotypic covariables stemming from phenotyping platforms and environmental covariables stemming from environmental characterisations and indices are best embedded in the class of mixed models with appropriate modelling of background genetic and non-genetic variances and correlations. In addition to genotypic covariables representing platform features, factorial regression models for yield in the field will also contain molecular marker information turning these factorial regression models into genomic prediction models with features for G×E. In recent years, several mixed model and empirical Bayes models have been proposed that integrate 1) marker information, 2) secondary phenotyping information, which can include various types of omics data, and 3) environmental information. Many of these proposals are presented as so-called multi-kernel methods. However, in essence these multi-kernel models follow the same structure and logic as the factorial regression models containing genotypic and environmental covariables, except that in multi-kernel methods less attention is given to covariable selection and more attention to the incorporation of high dimensional information. We compared a multi-kernel approach at data integration with a factorial regression approach for the maize data of the EU-DROPS project. Results were published in Millet et al. (2019).

The main alternative to factorial regression / multi-kernel methods are multi-trait models, models in which multiple responses simultaneously are modelled as opposed to the factorial regression models above that usually contain a single response, mostly yield. In a multi-trait model, yield is modelled as a response together with secondary phenotypes. For data integration, it is crucial that marker information can be inserted into the prediction models. A few multi-trait prediction methods have been proposed and these are evaluated in Arouisse et al. (2021) for real and simulated data sets. The same paper also proposes some promising new techniques.

A major work underlying deliverable D2.6 is the broad overview of data integration approaches in the context of the use of platform traits for modelling G×E in yield published by van Eeuwijk et al. (2019).

Table of contents

1. The role of research activities on statistical design and analysis within EPPN2020

EPPN2020 helps the plant community in progressing towards excellence across the whole phenotyping pipeline and addresses in the Joint Research Action 2, the following statistical issues:

- adequate experimental designs for phenotyping experiments
- analysis of single and multiple experiments with adjustment of treatment contrasts for environmental conditions, disturbances, and noise patterns at different spatial and temporal resolutions
- identification of outlying observations from sensor and imaging outputs
- extraction of secondary phenotypic traits on collections of genotypes
- choosing statistical models with an eye on biological interpretation of statistical parameters
- analysis of series of experiments as well as meta-analyses of experiments, with an emphasis on data integration between platform and field experiments

This document is dedicated to the last bullet point in the list above. In full: D2.6 Integration procedures and software for data from multiple platform and field experiments, with different but overlapping sets of genotypes, across scales of plant organization, traits and management conditions. The work on D2.6 is an extension of work on D2.3: Statistical methods and software for analysis of single and multiple platform experiments. D2.3 focussed on relations within and between platform experiments. D2.6 looks at the integration of phenotypic information obtained in field and platform experiments. For D2.6, the main objective is to incorporate secondary phenotypes as collected on one or more phenotyping platforms in models for primary phenotypes as collected in field experiments. In practice, the objective of D2.6 can be translated into the development of statistical models for further analysis of genotypic main effects and genotype by environment interactions for yield as occurring in field experiments by incorporating secondary phenotypes measured on phenotyping platforms into these models.

2. Two approaches underlying data integration

In EPPN2020, phenotyping platforms were used to characterize collections of genotypes. A major research objective was to use this information to understand yield and other traits in the field for the same genotypes. We give here a non-technical description of two major classes of data integration methods that were found useful in EPPN2020. Technical details and further options for data integration are given in van Eeuwijk et al. (2019).

2.1. Factorial regression models

From the perspective of EPPN2020, a plausible integration of platform and field data is to transform platform data into genotypic covariables that can explain (describe) average performance of genotypes across trials as well as differential genotypic performance, i.e., genotype by environment interaction. For a series of field trials, we can define a statistical model as y = E + G + G×E + error, with y the response, E the environmental main effect (~trial mean), G the genotypic main effect (average performance of genotype), G×E the genotype by environment interaction (differences between genotypes that depend on the environment / trial conditions). A simple but powerful data integration strategy is to write G and G×E as functions

of genotypic covariables. A genotypic characterization as measured on the platform is used to explain average differences in yield between genotypes in the field, while the same or another platform characterization is used to describe how environmental conditions change the performance of genotypes with respect to each other. The class of statistical models that inserts genotypic covariables on G and G×E is called factorial regression models. When we insert a genotypic covariable for G, we partition the genetic variation into a part explained by a regression on the genotypic covariable and a residual: G = slope * genotypic covariable + residual G. For the G×E, we estimate a separate parameter or slope for each environment or trial: G×E = environment dependent slope * genotypic covariable + residual G×E. Factorial regression models can also include environmental covariables that can be constructed as environmental groupings and indices. These environmental covariables partition the average differences between trials and the G×E. Groupings of trials can be obtained from statistically clustering trials with similar conditions. The outcome of such a clustering operation are environmental scenarios, environmental types or adaptation zones. These groupings can be based on observed soil and meteorological conditions, but also on outputs of crop growth model simulations that provide synthetic environmental information. In Millet et al. (2019) that was written partly as a case study on data integration between field and phenotyping platforms, various types of genotypic and environmental covariables were used. This case study will be illustrated below.

2.2. Multi-trait models

An area with major activity in data integration between field and platform data is genomic prediction. Genomic prediction aims at predicting phenotypes from marker profiles. A set of genotypes that has both marker and phenotypic data is used to train a prediction model, i.e., parameters are estimated for the prediction model. The estimated prediction model is then applied to a set of genotypes for which only marker data are available, the test set. The hope is that a genomic prediction model allows the evaluation of large numbers of genotypes for one or more phenotypic traits and/or environments at low costs, i.e., only genetic data are required and no phenotypic evaluations. Various criteria have been defined to assess the quality of the predictions by genomic prediction models. Commonly a cross validation strategy is used where a so-called calibration set of genotypes with both markers and phenotypes is split into a training set and a validation set. The prediction model is estimated on the training set and evaluated by predicting the phenotypes for the hold out validation set. Because we have observed phenotypes for the hold out validation set, we can correlate the predicted phenotypes with the observed phenotypes. This correlation is often called prediction accuracy. Of course, the prediction accuracy that was estimated on the validation set is usually an underestimate of the accuracy that will be realized on an independent test set.

In the genomic prediction literature, the paper by Burgueño et al. (2012) is an essential paper describing how marker data can help to predict G×E within a mixed model framework. This framework for predicting yield in multiple environments can be generalized to genomic prediction of multiple traits, i.e., yield and secondary traits. Burgueño et al. distinguish two situations for prediction. Translated to our case, the first situation has all traits, yield and secondary traits, measured on the training set and not at all at the test set. For this prediction scenario, called CV1 (from cross validation), we want to predict responses for new genotypes, and we have marker data and trait observations for the training set, whereas for the test set we only have marker data. An alternative prediction scenario, CV2, has marker data and secondary phenotyping data for both training and test set, while yield is only observed for the training set. The purpose of this prediction scenario is to predict yield for the test set, while the prediction model can use marker and secondary trait information on all genotypes and yield for the genotypes in the training set.

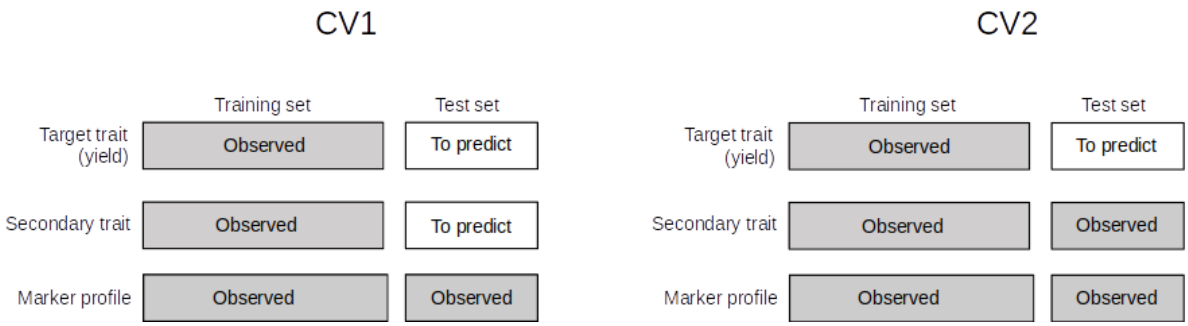

Fig.1 Schematic representation of the two cross-validation schemes.

For the purposes of EPPN2020, we consider CV2 to be of less importance than CV1. Most of the platforms in EPPN2020 aim at thorough physiological characterization of genotypes that are evaluated both at the platform and in the field. It would be interesting when secondary phenotyping on platforms could serve prediction purposes for yield in future fields alongside marker information for new sets of genotypes. This question, starting from a multi-trait prediction model, was investigated in Arouisse et al. (2021), among other methodological questions about the use of secondary phenotype information in genomic prediction. In the next section, we will elaborate a case study to illustrate how secondary phenotyping information can add predictive power to marker information in genomic prediction.

In standard multi-trait mixed models, all traits contribute to the prediction of one or more focal or target responses, for example, yield. One can think of these prediction models as regressions in which the weights or slopes for the predictive traits are functions of genetic correlations between target trait and predictive traits as well as of the heritabilities of target and predictive traits. When the target trait is average performance in the field or sensitivity to drought stress in the field and the predictive traits are all measured at one or more platforms, the multi-trait mixed model is equivalent to the application of so-called index selection (Lynch and Walsh, 1998). A predictive function of platform traits is constructed in such a way that its genetic correlation with the target trait is maximal. One may want to use the predicted values from the selection index as selection criterion in place of the target trait itself. This is attractive when the product of the heritability of the selection index and the squared genetic correlation between index and target trait is larger than the heritability of the target trait, yield in the field. The prediction model underlying the selection index can be improved by also using marker information. An interesting approach to arrive at an index in which the set of secondary phenotypes was tried to be reduced by an automatic variable selection implemented in a penalized regression (Lasso) was described in Lopez-Cruz et al. (2020). In Arouisse et al. (2021), we evaluated this penalized index and compared it with other proposals for combining marker data with secondary phenotypes for prediction.

3. Illustration of data integration approaches

3.1. Factorial regression (Millet et al. 2019)

In Millet et al. (2019), we used a data integration technique for combining field and platform data that is strongly based on factorial regression. The idea here was to use marker and platform data on a maize panel to predict field performance with attention for G×E. Platform data were used to estimate timing of genotype specific phenologies. In maize, the progression of phenological stages closely follows thermal time, with a nearly constant leaf appearance rate. Thermal time based on meristem temperature can be used to calculate leaf stages on platform and in the field, provided that the leaf emission rates are the same. Leaf stages correspond to developmental stages of the ear and can be used to define phenological periods. So, measuring leaf appearance at the platform allows to define the length of phenological stages in the field and to calculate the environmental conditions working at a particular genotype in a field experiment. In that way, environmental covariables can be calculated that are genotype and environment specific. With respect to those covariables, genotypic sensitivities are estimated. For the maize panel, three environmental covariables could be identified that explained a substantial part of the G×E in grain number. The identification of these covariables was facilitated by physiological insights that prioritized certain combinations of covariables for inclusion in the factorial regression model. As yield is grain number × grain weight, and no G×E occurred for grain weight in the maize field experiments studied in Millet et al. (2019), we could describe G×E in yield from G×E in grain number. The average genotypic performance and the genotypic sensitivities that drove G×E for grain number could successfully be predicted from marker profiles in a simple genomic prediction model. So, G×E for yield in the field could be predicted from marker data and environmental covariables by using a factorial regression model. The platform served to estimate the length of genotype specific phenological stages and to calculate environmental characterizations. The quality of the prediction model was evaluated on multiple test sets and compared to a benchmark mixed model that predicts genotypes for specific conditions from random genotype by environment effects that were structured by markers at the genotypic side and environmental similarities calculated from all environmental covariables at the environmental side (Jarquin et al., 2014). The performance of the factorial regression model with genotypic parameters (average performance and sensitivities) predicted from marker profiles was good and surely competitive in comparison to the benchmark model. An impression of the model structure and the predictive performance is given in Figure 4 of Millet et al. (2019) that is included below.

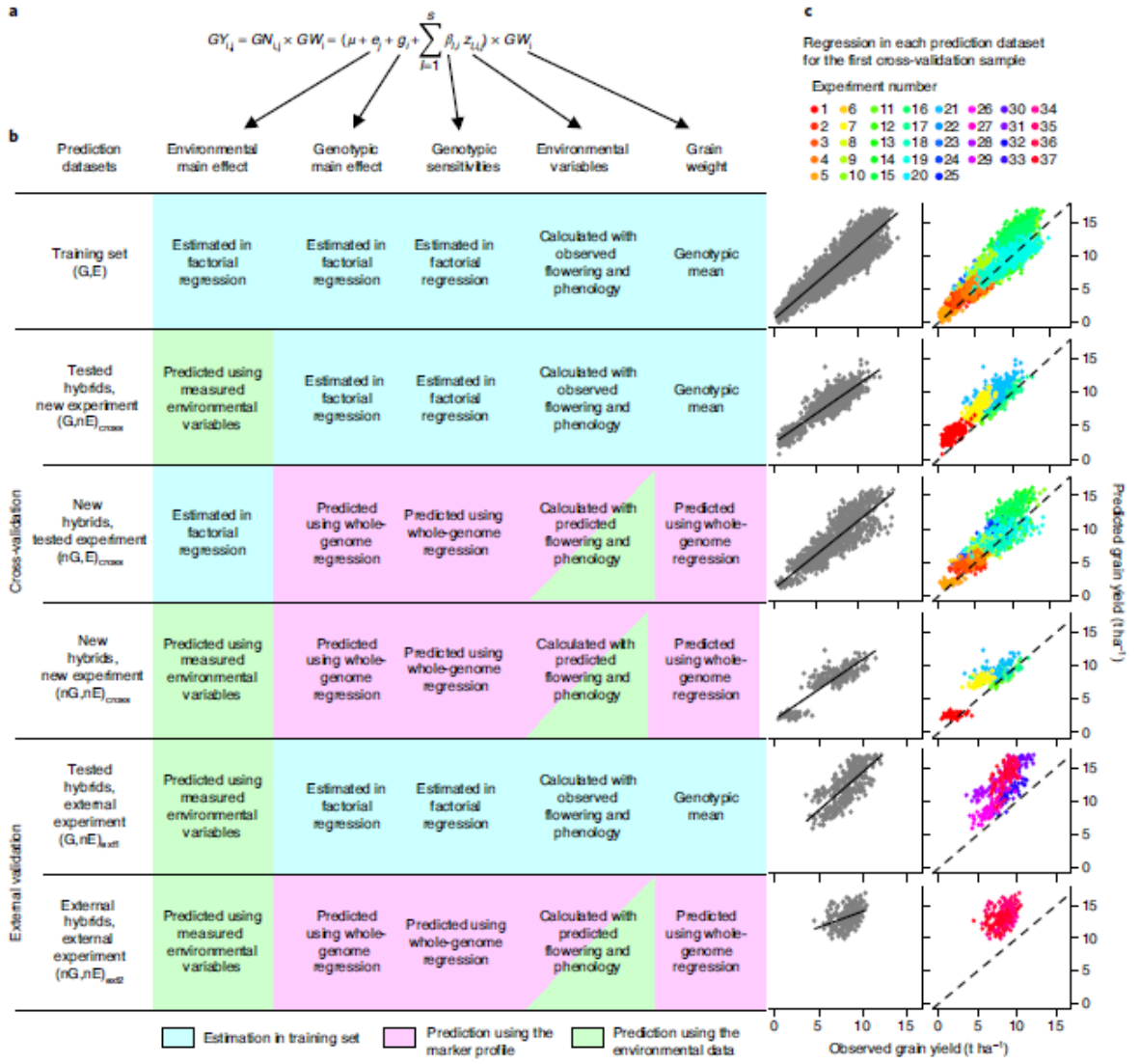

Fig. 4 | Yield prediction: method and results for each dataset. a, Equation for the dissection of grain yield (GY_{ij}). GN_{ij} grain number of hybrid *i* in experiment *j*, GW_i mean grain weight of hybrid *i*. For GN_{ij}, μ, intercept; e_j, environmental main effect in field *j*; g_i, genotypic main effect for hybrid *i*; β_{il}, sensitivity of hybrid *i* to the genotype-specific environmental index z_{lij} (intercepted light, night temperature or soil water potential). **b**, Model parameters, either estimated or predicted depending on whether hybrids and experiments were used as training (G and E, respectively) or test set (nG and nE, respectively). For already tested G and E (light blue cells), parameters were estimated from measured data via the factorial regression for grain number (equation (2)) or averaged across experiments for grain weight (Methods). For nE, the environmental main effects were predicted using measured environmental data (green cells). For nG, genotypic parameters were predicted using a genomic prediction model (magenta cells). **c**, Relationships between predicted and observed grain yield in each dataset for the first cross-validation sample, either cross-validation or external validation sets. In the first column, the overall correlation is presented with a linear regression across hybrids and experiments (black line). The second column shows the same correlations, in which each experiment is represented with a color. Dashed black lines indicate the first bisector. Experimental fields numbered as in Supplementary Table 1.

3.2. Multi-trait prediction (Arouisse et al., 2021)

In Arouisse et al. (2021) we wanted to know whether sets of secondary phenotypes are useful for a better prediction of yield in the field, when the secondary phenotypes are added to a set of markers. In other words, would additional secondary phenotypic predictors add anything to the predictive accuracy of common genomic prediction models like GBLUP? A few solutions have been offered for this problem. The simplest solution is to do a multi-trait GBLUP, but this solution only works for a limited number of secondary traits, while increasingly so high dimensional secondary traits are collected. A popular approach is the use of multi-kernel methods. Multi-kernel methods, or multi-BLUPs, are straightforward extensions of GBLUP, where in addition to a genetic random effect, a second random effect in inserted in the prediction model that reflects the secondary phenotypes. The genetic random effect is structured by marker-based relations between genotypes, while the additional random effect is structured by relationships between genotypes that are calculated from secondary phenotypes. Each structured random effect represents a kernel. An early example of this approach can be found in Riedelsheimer et al. (2012). Another type of solution reduces the dimensionality of the secondary traits by transforming them into a selection index (Lynch and Walsh, 1998) that maximizes the genetic correlation between a linear combination of the secondary traits and the target trait. An interesting twist to the selection index approach was proposed by Lopez Cruz et al. (2020). These authors introduced a penalized version of the selection index that aims to reduce the dimensionality of the set of secondary traits.

Arouisse et al. (2021) evaluated all the above methods for integrating secondary traits into genomic prediction models on various real and simulated data sets. They introduced firstly a dimension reduction of the secondary trait set by applying a penalized regression (Lasso) of target trait on secondary traits or a random forest regression and then inserted the predictions of those regressions in a bivariate GBLUP together with the target trait. The new methods were called LS-BLUP and RF-BLUP, respectively. Secondly, they adapted the multi-kernel model. In place of a random effect structured by secondary traits, they introduced a random effect structured by the genomic predictions of the secondary traits, that is, they performed GBLUP of secondary traits on markers and used the predictions of those GBLUPs to structure the additional random effect. The advantage of the latter approach, called GM-BLUP) is that it can also be used when the secondary traits haven't been measured on the test set.

The results of the application of different prediction models integrating markers and secondary phenotypes on some real data sets are presented in Tables 1 and 2 of Arouisse et al. (2021) and repeated on the next page. The overall conclusion was that none of the methods that added secondary phenotypes to markers did better than GBLUP for the situation that secondary phenotypes were not measured at the test set (CV1). However, LS-BLUP and RF-BLUP did better than GBLUP and also better than other prediction models for the situation in which secondary phenotypes were measured on both training and test set.

TABLE 1 | Prediction accuracy in scenario 1, for various target and secondary traits in Maize and Arabidopsis.

Data sets	Target trait	Secondary phenotypes	GBLUP	GM-BLUP	LS-BLUP	RF-BLUP	RF-BLUP*
1	Number of spreading lesions under fungus stress	Metabolites	0.23	0.22	0.20	0.21	0.21
	Fresh weight of the rosette under Salt_5 stress	Metabolites	0.03	0.00	0.07	0.09	0.09
	Number of spreading lesions under Drought_and_fungus stress	Metabolites	0.19	0.18	0.16	0.16	0.15
	Number of damaged leaves and feeding sites under Caterpillar_3 stress	Metabolites	0.10	0.09	0.06	0.10	0.10
2	Fresh weight	Metabolites	0.30	0.30	0.29	0.30	0.30
3	Flowering time (FT) [4]	Transcripts	0.54	0.55	0.55	0.53	0.55
	Plant height (PH)	Transcripts	0.54	0.55	0.55	0.53	0.51
	Yield	Transcripts + FT+PH	0.53	0.53	0.54	0.52	0.52
	Yield	Transcripts	0.55	0.55	0.55	0.55	0.55

*Acronyms of the methods are as in **Figures 2**, **3**. For RF-BLUP*, we used the randomForest package with the default settings; for RF-BLUP, hyper-parameters were optimized using the caret package (data-sets 1 and 2) or scikit-learn (data-set 3). For data-sets 1 and 2, reported accuracies are averages over 160 test sets (standard errors between 0.006 and 0.007), except for RF-BLUP, where 50 sets were used (SE between 0.010 and 0.014). In dataset 3, 30 test sets were used for all methods (SE between 0.006 and 0.03).*

TABLE 2 | Prediction accuracy in scenario 2, for various target and secondary traits in Maize and Arabidopsis.

Data sets	Target trait	Secondary phenotypes	GBLUP	M-BLUP	Multi-BLUP	GM-BLUP	LS-BLUP	RF-BLUP	RF-BLUP*
1	Number of spreading lesions under fungus stress	Metabolites	0.23	−0.04	0.21	0.22	0.31	0.28	0.28
	Fresh weight of the rosette under Salt_5 stress	Metabolites	0.03	0.09	0.08	0.07	0.23	0.20	0.19
	Number of spreading lesions under Drought_and_fungus stress	Metabolites	0.19	−0.02	0.16	0.17	0.27	0.25	0.23
	Number of damaged leaves and feeding sites under Caterpillar_3 stress	Metabolites	0.10	0.05	0.06	0.07	0.14	0.12	0.11
2	Fresh weight	Metabolites	0.30	0.00	0.29	0.30	0.32	0.30	0.28
3	Flowering time (FT) [4]	Transcripts	0.55	0.54	0.55	0.55	0.66	0.65	0.54
	Plant height (PH)	Transcripts	0.54	0.53	0.54	0.55	0.66	0.64	0.53
	Yield	Transcripts + FT+PH	0.53	0.49	0.50	0.52	0.72	0.71	0.49
	Yield	Transcripts	0.55	0.52	0.53	0.54	0.64	0.65	0.51

*Acronyms of the methods are as in **Figures 2**, **3**. For RF-BLUP*, we used the randomForest package with the default settings; for RF-BLUP, hyper-parameters were optimized using the caret package (data-sets 1 and 2) or scikit-learn (data-set 3). For data-sets 1 and 2, reported accuracies are averages over 160 test sets (standard errors between 0.006 and 0.012), except for RF-BLUP, where 50 sets were used (SE between 0.010 and 0.014). In dataset 3, 30 test sets were used for all methods (SE between 0.006 and 0.03).*

4. Conclusion

Work on data integration within the EPPN2020 Joint Research Action 2 has produced three papers proposing, describing and comparing different statistical procedures. The paper by van Eeuwijk et al. (2019) gives a broad overview of how secondary phenotyping information can be used to improve both models that aim at a better understanding of G×E in the field as well as models that aim principally at a better prediction of G×E. The paper by Millet et al. (2019) contains a good example of how a factorial regression model together with platform information can provide insight and predictive power for G×E in the field. In contrast, the paper by Arouisse et al. (2021) emphasizes the potential improvement of prediction accuracy by using secondary phenotypes alongside markers in generalizations of common genomic prediction models. Software for all the methods is available as additional material to the papers. It concerns programs and procedures in R and Python, with sometimes a reliance on ASReml.

5. References

Arouisse, Bader, Tom PJM Theeuwen, Fred A. Van Eeuwijk, and Willem Kruijer. "Improving genomic prediction using high-dimensional secondary phenotypes." Frontiers in Genetics 12 (2021).

Burgueño, Juan, Gustavo de los Campos, Kent Weigel, and José Crossa. "Genomic prediction of breeding values when modeling genotype× environment interaction using pedigree and dense molecular markers." Crop Science 52, no. 2 (2012): 707-719.

Jarquín, Diego, José Crossa, Xavier Lacaze, Philippe Du Cheyron, Joëlle Daucourt, Josiane Lorgeou, François Piraux et al. "A reaction norm model for genomic selection using high-dimensional genomic and environmental data." Theoretical and applied genetics 127, no. 3 (2014): 595-607.

Lopez-Cruz, Marco, Eric Olson, Gabriel Rovere, Jose Crossa, Susanne Dreisigacker, Suchismita Mondal, Ravi Singh, and Gustavo de Los Campos. "Regularized selection indices for breeding value prediction using hyper-spectral image data." Scientific reports 10, no. 1 (2020): 1-12.

Lynch, Michael, and Bruce Walsh. "Genetics and analysis of quantitative traits." (1998)

Millet, Emilie J., Willem Kruijer, Aude Coupel-Ledru, Santiago Alvarez Prado, Llorenç Cabrera-Bosquet, Sébastien Lacube, Alain Charcosset, Claude Welcker, Fred van Eeuwijk, and François Tardieu. "Genomic prediction of maize yield across European environmental conditions." Nature genetics 51, no. 6 (2019): 952-956.

Van Eeuwijk, Fred A., Daniela Bustos-Korts, Emilie J. Millet, Martin P. Boer, Willem Kruijer, Addie Thompson, Marcos Malosetti et al. "Modelling strategies for assessing and increasing the effectiveness of new phenotyping techniques in plant breeding." Plant science 282 (2019): 23-39.

Riedelsheimer, Christian, Angelika Czedik-Eysenberg, Christoph Grieder, Jan Lisec, Frank Technow, Ronan Sulpice, Thomas Altmann, Mark Stitt, Lothar Willmitzer, and Albrecht E. Melchinger. "Genomic and metabolic prediction of complex heterotic traits in hybrid maize." Nature genetics 44, no. 2 (2012): 217-220.