Tropentag 2021, hybrid conference
September 15-17, 2021

Conference on International Research on Food Security, Natural Resource Management and Rural Development
organised by the University of Hohenheim, Germany

Statistical analysis of data from the field phenotyping platform "BreedVision"

Md Anisur Rahaman[a], Hans-Peter Piepho[b] and Hans Peter Maurer[c]

a University of Hohenheim, Institute of Crop Science, Germany.
b University of Hohenheim, Biostatistics unit, Institute of Crop Science, Germany.
c University of Hohenheim, State Plant Breeding institute, Germany.

Abstract

The inclusion of secondary traits and environmental covariates in the prediction model became an important consideration in recent years. Our study was aimed at quantifying the impact of adding secondary traits and covariates in the improvement of the target trait. An experiment was conducted in 2018–19 at Hohenheim for sensor-based non-invasive prediction of yield (biomass and grain) in triticale (× Triticosecale Wittmack) field trials with four trial areas t_1, t_2, t_3 and t_4 and four nitrogen levels N_1= 40 %, N_- 2= 70 %, N_3= 100 %, and N_4= 130 % in each trial area. A trial area contains 25 triticale genotypes in an α-lattice design with ten incomplete blocks within two replicates and the data were recorded for dry matter yield (DMY) and a secondary trait canopy temperature (CT). CT was measured with a sensor machine using the field phenotyping platform "BreedVision" from the Senselgo project during the vegetation period by using hyperspectral cameras and the sensor machine ran twice in each plot for the data recording. Undoutedly, mixed models are effective in handling repeated measures on the same statistical units and are widely used in biological sciences. Therefore, considering repeated measure issues and measurement of two traits in parallel, a mixed linear bivariate model was developed to predict the target trait 'DMY' without measuring it in the field. Radiation intensity and ambient temperature are two covariates also considered in the model. The model with covariates showed reasonable improvement in prediction performance. There was no gain from the model with secondary trait CT (bivariate model), however, we recommend using this model where it is difficult to measure the target trait DMY due to extreme weather conditions and limited seed supply. Thus, our model allows early selections to be made and saving considerable resources in breeding experiments.

Keywords: Mixed modelling, phenotyping, prediction performance, repeated measures

*Corresponding author Email: anis307@gmail.com

Introduction

Sensor-based phenotyping accelerates the breeding experiments in environmentally controlled greenhouse conditions and in field conditions. Measurement of a secondary trait such as canopy temperature in pedigree selection is a major achievement of sensor-based phenotyping platforms (Rutkoski et al., 2016). State Plant Breeding Institute, Universität Hohenheim, in collaboration with the Competence Centre of Applied Agricultural Engineering (COALA), University of Applied Sciences Osnabrück have developed the phenotyping platform "BreedVision" to accelerate and to improve the phenotyping of plants under field conditions. Phenotyping experiments have a complex experimental design with random effects and fixed effects to be included in the model. To analyze vast amount of data coming from field phenotyping platform, linear mixed modeling (LMM) is used due to its flexibility to handle unbalanced data and complex experimental designs (Piepho and Möhring, 2011). Besides, LMM can include both fixed and random effects during model development which makes mixed modeling a most promising statistical tool to analyze complex phenomics data. Previous experiments at the International Wheat and Maize Improvement Center (CIMMYT) found a high correlation between canopy temperature (CT) and Dry matter yield (DMY) (Rutkoski et al., 2016; Sun et al., 2019). These findings motivated us to include this trait in our model to get a better prediction. However, ambient temperature and radiation intensity are two continuous variables that can cause significant variation in the recorded data for canopy temperature. Adding such continuous variables that are commonly called covariates, can help to reduce error and improve the prediction accuracy in a model (Fisher (1937/1947; as cited in Bloom et al., 2007). Our study will aim at finding a suitable univariate model in the first step to analyze sensor-based data from "BreedVision" platforms studying trait DMY and CT separately. Two environmental covariates-radiation intensity and ambient temperature be included only in trait CT for the analysis. Finally, we will derive univariate and bivariate prediction models to i) evaluate the effect of adding covariates into the model and ii) compare the prediction accuracy of univariate and bivariate analysis. An applicable prediction model developed from this study may be useful to predict the target trait 'DMY' without measuring it in the field, thus allowing an early selection to be made and saving considerable resources in phenotyping experiments.

Methods

The experiment was conducted in 2018-19 at Hohenheim with four trial areas with four nitrogen levels N_1= 40%, N_2= 70%, N_3= 100%, and N_4= 130%. Here one nitrogen level was applied in a trial area (e.g., N_1 was only applied in t_1). In a trial area, 25 genotypes were planted in an α-lattice design with ten incomplete blocks within two replicates; in a replicate, 25 genotypes were allocated in five blocks. In all plots, data were recorded for two traits namely, dry matter yield (DMY) and canopy temperature (CT). Additionally, to add more precision to the analysis, two covariates were also recorded as radiation intensity and ambient temperature during trait CT measurements. CT was measured by the sensor machine and the sensor machine ran twice in each plot for data recording. We have 600 observations for the complete dataset, where 200 observations originate from DMY and 400 observations from CT and following models were developed from the recorded data.

We start the analysis with deriving the following univariate model for DMY data-

$$DMY_{ijhnt} = \mu_{nt} + \gamma_{jt} + \tau_{in} + b_{jht} + \varepsilon_{ijhnt}, \quad (1)$$

Here, DMY_{ijhnt} is response of *i*-th genotype in *h*-th block nested within *j*-th replicate, μ is the intercept, γ_j is the effect of *j*-th complete replicate, n,t = subscript for nitrogen levels and trial areas respectively and ε_{ijhnt} = residual plot error associated with DMY_{ijhnt}, $\varepsilon_{ijhnt} \sim N(0, \sigma_e^2)$.

Similarly, a univariate model for CT can be derived in the following equation-

$$CT_{ijkhnt} = \mu_{nt} + \gamma_{jt} + \tau_{in} + b_{jht} + \varepsilon_{ijkhnt}, \quad (2)$$

here ε_{ijkhnt} is the residual error associated with CT_{ijkhnt}, error terms can be stacked into a vector $\boldsymbol{\varepsilon}_{ijkhnt}$ and error vector $\boldsymbol{\varepsilon}_{ijkhnt}$ assumed to be normally distributed with $\boldsymbol{\varepsilon}_{ijkhnt} \sim$ N (0,**R)** and **R** could be defined as follows:

$$\mathbf{R} = \mathrm{I}_4 \otimes \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \mathrm{I}_{50}\sigma_e^2 \quad (3)$$

The experiment has four trial areas, so, CT has 400 observations. In this case, data was sorted by four nitrogen levels or trial areas, blocks within the trial areas, plots within blocks and measuring replicates within the plots. Here, each measuring replicate has 200 observations. To include all four trial areas into the model, it is reasonable to consider a Kronecker product of identity matrix I_4 with the variance-covariance matrix derived in the model developed for single nitrogen level.
Matrix for random effects or **G** matrix could also be derived as follows:

$$\mathbf{G} = \mathrm{I}_4 \otimes (\mathrm{I}_{10} \otimes \mathrm{J}_{10\times 10}\sigma_b^2). \quad (4)$$

To get the G matrix for all four trial areas, also a Kronecker product of identity matrix I_4 was considered with the **G** matrix for the model for single nitrogen level.
In this experiment, ambient temperature and radiation intensity were recorded as covariates. With the covariates, the model (Eq. 2) extends to:

$$CT_{ijkhnt} = \mu_{nt} + \gamma_{jt} + \tau_{in} + \beta_1 x_{1ijkhnt} + \beta_2 x_{2ijkhnt} + b_{jht} + \varepsilon_{ijkhnt}, \quad (5)$$

here β_1 is the change in CT (slope) due to a unit change in ambient temperature ($x_{1ijkhnt}$), β_2 is the change in CT (slope) due to a unit change in radiation intensity ($x_{2ijkhnt}$), $x_{1ijkhnt}$ is the ambient temperature corresponding to CT_{ijkhnt} and $x_{2ijkhnt}$ is the radiation intensity corresponding to CT_{ijkhnt} and all other variables are defined as similar as defined in (Eq. 2).
Data were measured for two traits and these two traits will be analyzed in a model simultaneously. To analyze these two traits, DMY and CT, together, it is required to derive a bivariate model. To extend the model to a bivariate case, it is necessary to add another subscript *b* (= DMY, CT) for traits to the model

$$Y_{ijkhntb} = \mu_{ntb} + \gamma_{jtb} + \tau_{inb} + \beta_1 x_{1ijkhntb} \cdot switch + \beta_2 x_{2ijkhntb} \cdot switch + b_{jhtb} + \varepsilon_{ijkhntb} \quad (6)$$

where *switch*=0 to *switch* off covariates in the trait (DMY) and *switch*=1 to include covariates in the trait CT. The variable *switch* is defined as a quantitative factor. In SAS, it is necessary to remove *switch* from class factors (Piepho et al., 2006). Here $\varepsilon_{ijkhntb}$ is the residual error associated with $Y_{ijkhntb}$ and plot effect is confounded with the error structure.
For checking models' performance in practice, Cross-validation (CV) was used for resampling data to assess the true prediction error of prediction models. We left out 20% of the data points randomly from trait DMY in univariate and bivariate models. Then, we used the remaining univariate DMY data or all remaining bivariate data to predict the data dropped. Subsequently, we calculated the correlation and root mean squared error (RMSE) between the observed and predicted values for the dropped DMY data. The process was repeated 100 and 10000 times using a macro, consequently, the mean correlation and RMSE between observed and predicted DMY values along with their confidence interval was calculated from 100 and 10000 runs then was compared for univariate and bivariate analysis.

Results and Discussion

The Akaike information criterion (AIC) and Bayesian information criterion (BIC) was used to compare between model without covariates (Eq. 2) and model with covariates (Eq. 5). Table 1 represents the results of this comparative analysis.

Table 1: Comparison between the model without covariates (Eq. 2) and model with covariates (Eq. 5)

Model	$-2 \cdot \log L_R$[†]	AIC[‡]	BIC*
Eq. 2	988.5	1194.5	1368.5
Eq. 5	539.2	749.2	926.5

[†]Log-likelihood estimate [‡]Akaike information criterion *Bayesian information criterion

The model without covariates (Eq. 2) has the AIC value of 1194.5 and the BIC value of 1368.5. On the other hand, the model with the covariates (Eq. 5) has an AIC value of 749.2 and the BIC value of 926.5. As smaller AIC and BIC values were observed for the model with covariates, this model can be selected, and this finding will also be used during the bivariate analysis. Rutkoski et al. (2016) reported similar results in case of days to heading as a covariate can improve prediction accuracies in univariate and bivariate models for wheat.

To evaluate prediction accuracies of the model, mean correlation and RMSE of observed and predicted values of DMY from 100 simulations cross-validation were presented in Table 2 to compare the results of univariate (Eq. 1) and bivariate (Eq.6) analysis in 100 simulations.

Table 2: 100 simulations results; calculation of mean correlation and RMSE between observed and predicted values from univariate (Eq.1) and bivariate (Eq. 6) analysis, values for the 95% confidence limits (CLs) of the means are in parenthesis.

Summary statistics	Univariate (Eq. 1)	Bivariate (Eq. 6)
Mean Pearson's Correlation	0.8015 [0.7785-0.8245]	0.8156 [0.7924-0.8387]
Mean RMSE**	0.8598 [0.8501-0.8695]	0.8535 [0.8501-0.8695]

**Root mean square error

For both univariate (0.8015) and bivariate (0.8156) analyses, a strong correlation between computed and predicted value was estimated with confidence limits (CLs) of [0.7785-0.8245] and [0.7924-0.8387], respectively, at the 95% level. The mean RMSE for univariate analysis was 0.8598 and 0.8535 for the bivariate analysis with a CLs of [0.8501-0.8695] and [0.8501-0.8695] respectively at the 95% level. To get more precise results we also calculated results from 10,000 simulations. For 10000 cross-validation simulations, the mean correlation and RMSE between the observed and predicted values were presented in Table 3. A strong correlation between computed and predicted value was estimated for both univariate (0.8599) and bivariate (0.8537) analyses, with CLs of [0.8590-0.8607] and [0.8528-0.8546], respectively at the 95% level.

Table 3: 10000 simulations results; calculation of mean correlation and RMSE between observed and predicted values from univariate (Eq.1) and bivariate (Eq. 6) analysis, values for the 95% confidence limits (CLs) of the means are in parenthesis.

Summary statistics	Univariate (Eq. 1)	Bivariate (Eq. 6)
Mean Pearson's correlation	0.8599 [0.8590-0.8607]	0.8537 [0.8528-0.8546]
Mean RMSE**	0.7949 [0.7929-0.7968]	0.8103 [0.8092-0.8115]

**Root mean squared error

The mean RMSE for univariate analysis was 0.7949 and 0.8103 for the bivariate analysis with a CLs of [0.7929-0.7968] and [0.8092-0.8115] respectively at the 95% level. Here, the univariate

model has a higher correlation and lower RMSE than the bivariate model which suggests that there is no gain from the bivariate model for prediction accuracy.
Our results showed that we have no gain from the bivariate analysis. Sun et al. (2017, 2019) reported contrasting findings in their works with wheat data at CIMMYT where they concluded that inclusion of CT can increase prediction accuracy up to 146% when secondary traits were used in both training and test populations. We used cross-validation by putting the secondary trait only in the training population. Our findings are supported by Rutkoski et al. (2016), as they have drawn a similar conclusion when secondary data were only used in the training population.

Conclusions and Outlook

Based on our results, we can conclude that there is no gain from analysis with secondary trait CT, if we add CT only in the training population. However, CT can be still useful to predict DMY in adverse conditions, when it is impossible to determine the primary trait DMY directly. Besides, CT can be used as an index for early selection of DMY in genomic selection cycles by predicting DMY at early growth stages, which will save a massive number of resources and time in breeding experiments. We also demonstrated that the model with covariates (radiation intensity and ambient temperature) has better performances, this indicates that the inclusion of these two covariates can improve the model performance substantially. From our results and previous literature, we found some loose ends in this research which opens new routes for future researchers. Among them, testing secondary traits across the environment and determining the best growth phase for measuring phenotyping data by sensor-based experiments are important to consider. Also, estimation of heritability to calculate percent prediction accuracy will be helpful to get more information on prediction performance of a model.

References

1. Bloom, H. S., Richburg-Hayes, L. and Black, A. R. 2007. Using covariates to improve precision for studies that randomize schools to evaluate educational interventions. Educational Evaluation and Policy Analysis 29(1):30–59. https://doi.org/10.3102/0162373707299550
2. Piepho, H. P. and Möhring, J. 2011. On estimation of genotypic correlations and their standard errors by multivariate REML using the MIXED procedure of the SAS system. Crop Science 51(6):2449–2454. https://doi.org/10.2135/cropsci2011.02.0088
3. Piepho, H. P., Williams, E. R. and Fleck, M. 2006. A note on the analysis of designed experiments with complex treatment structure. HortScience 41(2):446–452. https://doi.org/10.21273/hortsci.41.2.446
4. Rutkoski, J., Poland, J., Mondal, S., Autrique, E., Pérez, L. G., Crossa, J., Reynolds, M. and Singh, R. 2016. Canopy temperature and vegetation indices from high-throughput phenotyping improve accuracy of pedigree and genomic selection for grain yield in wheat. G3: Genes, Genomes, Genetics 6(9):2799–2808. https://doi.org/10.1534/g3.116.032888
5. Sun, J., Poland, J. A., Mondal, S., Crossa, J., Juliana, P., Singh, R. P., Rutkoski, J. E., Jannink, J. L., Crespo-Herrera, L., Velu, G., Huerta-Espino, J. and Sorrells, M. E. 2019. High-throughput phenotyping platforms enhance genomic selection for wheat grain yield across populations and cycles in early stage. Theoretical and Applied Genetics 132(6):1705–1720. https://doi.org/10.1007/s00122-019-03309-0
6. Sun, J., Rutkoski, J. E., Poland, J. A., Crossa, J., Jannink, J. L. and Sorrells, M. E. 2017. Multitrait, random regression or simple repeatability model in high-throughput phenotyping data improve genomic prediction for wheat grain yield. The Plant Genome 10(2). https://doi.org/10.3835/plantgenome2016.11.0111