Published May 9, 2019

RESEARCH

A Cross-Validation of Statistical Models for Zoned-Based Prediction in Cultivar Testing

Harimurti Buntaran,* Hans-Peter Piepho, Jannie Hagman, and Johannes Forkman

ABSTRACT

The principal goals of a plant breeding program are to provide breeders with cultivar information for selection purposes and to provide farmers with high-yielding and stable cultivars. For that reason, multi-environment trials need to be done to predict future cultivar yield, and a robust statistical procedure is needed to provide reliable information on the tested cultivars. In Sweden, the statistical procedure follows the tradition of modeling cultivar effects as fixed. Moreover, the analysis is performed separately by zone and level of fungicide treatment, and so the factorial information regarding cultivar × zone × fungicide combinations is not explored. Thus, the question arose whether the statistical method could be improved to increase accuracy in zone-based cultivar prediction, since the cultivar recommendation is zone based. In this paper, the performance of empirical best linear unbiased estimation (E-BLUE) and empirical best linear unbiased prediction (E-BLUP) are compared using cross-validation for winter wheat (*Triticum aestivum* L.) and spring barley (*Hordeum vulgare* L.), in single-year and multiyear series of trials. Data were obtained from three agricultural zones of Sweden. Several linear mixed models were compared, and model performance was evaluated using the mean squared error of prediction criterion. The E-BLUP method outperformed the E-BLUE method in both crops and series. The prediction accuracy for zone-based yield was improved by using E-BLUP because the random-effects assumption for cultivar × zone interaction allows information to be borrowed across zones. We conclude that E-BLUP should replace the currently used E-BLUE approach to predict zone-based cultivar yield.

H. Buntaran and J. Forkman, Dep. of Energy and Technology, Swedish Univ. of Agricultural Sciences, Box 7032, 750 07 Uppsala, Sweden; H. Buntaran, J. Hagman, and J. Forkman, Dep. of Crop Production Ecology, Swedish Univ. of Agricultural Sciences, Box 7043, 750 07 Uppsala, Sweden; H.-P. Piepho, Biostatistics Unit, Institute of Crop Science, Univ. of Hohenheim, 70599 Stuttgart, Germany. Received 23 Oct. 2018. Accepted 25 Mar. 2019. *Corresponding author (harimurti.buntaran@slu.se). Assigned to Associate Editor Lucia Gutierrez.

Abbreviations: BLUE, best linear unbiased estimation; BLUP, best linear unbiased prediction; CV, cross-validation; DMY, dry matter yield; E-BLUE, empirical best linear unbiased estimation; E-BLUP, empirical best linear unbiased prediction; FA, factor analytic; MET, multi-environment trials; MF, multiyear and fixed effects for cultivar; MR, multiyear and random effects for cultivar; MSEP, mean squared error of prediction differences; SF, single-year and fixed effects for cultivar; SR, single-year and random effects for cultivar.

The aim of multi-environment trials (METs) is to evaluate and test the performance of cultivars in various environmental conditions. The MET results not only provide cultivar information to breeders for selection purposes but also are the basis for advice to farmers in deciding which cultivar is the best or the most suitable concerning their local field conditions. Thus, reliable statistical methods are necessary to give both breeders and farmers accurate information.

In Swedish cultivar trials, the statistical method used for analyzing MET data has not been changed for many years. Moreover, the number of trials has been decreasing in recent years. Hence, there is a demand for improvement in statistical analysis to provide better accuracy for zoned-based cultivar performance assessment and ranking in different environments based on a reduced number of trials. Currently, the analyses are done with an unweighted two-stage analysis (Möhring and Piepho, 2009). At the first stage, each experiment is analyzed using a linear mixed

Published in Crop Sci. 59:1544–1553 (2019).
doi: 10.2135/cropsci2018.10.0642

model with fixed effects of cultivars, fungicide treatments, and cultivar × fungicide treatment interactions, and random effects of replicates and incomplete blocks. Then, at the second stage, another linear mixed model is fitted, by zone and fungicide treatment. In this model, the effects of cultivars are fixed, and the effects of trials within zones are random. The current statistical method does not exploit cultivar × environment interaction explicitly. Furthermore, possible heterogeneity among zones and trials is not investigated or accounted for.

Identification of mega-environments is relevant for assortment of trials that are considered to originate from two or more discrete target populations of environments (van Eeuwijk et al., 2016). Regionalization refers to the subdivision of such a target population into zones, based on geography and agroecological factors. In Sweden, a zone is represented by a number of trials. The trial sites differ from one year to another. Since variation between trials within zones may be large, it is an important question whether a zone-based analysis is indeed necessary. The question is whether the prediction of cultivar performance must be based on a classification of trials into zones, or prediction could be done without such classification. The present article investigates this issue.

The discussion of whether effects of cultivars should be modeled as fixed (and estimated by best linear unbiased estimation [BLUE]) or random (and estimated by best linear unbiased prediction [BLUP]) has been addressed in several previous studies. Smith et al. (2001) argued for modeling cultivar effects as random, because of the "deficiency in the traditional fixed cultivar-effects approach in terms of obtaining reliable predictions of future yield performance." This deficiency was also discussed by Patterson and Silvey (1980), who stated that "differences between trials means for newly recommended cultivars are, on the average, about 27% too large." Best linear unbiased prediction is a shrinkage method, since information about the distribution is used, in essence, to "shrink" the effects towards zero (Stroup, 2012; Galwey, 2014). The magnitude of the shrinkage depends on the "shrinkage factor," and, in a simple model, the shrinkage factor is a function of heritability as described in Galwey (2014, p. 169). Shrinkage thus reduces the spread of the predictions in comparison with fixed-effects estimation (Robinson, 1991). In the case of random effects for cultivar, the assessment of the cultivar mean yield in a particular environment may be viewed as a prediction rather than one of estimation (McCulloch et al., 2008, p. 19). Kleinknecht et al. (2013) conducted simulation studies to compare the estimation of cultivar effects with zoned MET using empirical BLUE (E-BLUE) and empirical BLUP (E-BLUP). The term "empirical" here means that variance components are not known and therefore are estimated from the data at hand. The authors concluded that the use of E-BLUP for routine analysis in cultivar trials is worthwhile, since E-BLUP always performed better in the simulations. Forkman and Piepho (2013) reported the results of a simulation study investigating the performance of E-BLUP in small randomized complete block experiments and concluded that E-BLUP was preferable to E-BLUE.

The optimality of BLUP rests on the validity of underlying assumptions, including that the variance components are known. In real situations, the assumptions are hardly met perfectly, and the variance components must be estimated. The theoretical results of advantages of BLUP as compared with BLUE refer to BLUP, not E-BLUP. Thus, the performance of E-BLUP in practice needs to be explored using simulation or cross-validation (CV).

In this study, we compare the performance of different linear mixed models that use either E-BLUE or E-BLUP, through CV. The models are evaluated using datasets with large as well as small numbers of cultivars. We also specifically investigate models without the effects of zones to examine the necessity of zonation. The overall aim of this paper is to improve the zoned-based prediction accuracy with a reduced number of trials, and so provide support for better cultivar recommendation.

MATERIALS AND METHODS

Single-Year/Within-Year Cross-Validation

Swedish Multi-environment Trial Data

The datasets used for the analyses performed in this study comprised the dry matter yield (DMY) datasets of winter wheat (*Triticum aestivum* L.) and spring barley (*Hordeum vulgare* L.) in Swedish METs. These trials were performed in three different agricultural zones: south (A), middle (D + E), and north (F), as depicted in Fig. 1. The cultivar trials were laid out as split-plot experiments with two replicates. Each experiment included two levels of a fungicide treatment (treated and untreated) as the main-plot factor. Within each fungicide treatment, cultivars were arranged in an α-design with two replications. The number of blocks within a replication is five. Our datasets for single-year CV were complete, meaning that all cultivars were present in all trials within a single year. The structure of the datasets for single-year CV is summarized in Supplemental Tables S1 and S2 for winter wheat and spring barley, respectively.

Linear Mixed Models

The statistical method for analyzing MET data was a two-stage unweighted procedure. In the first stage, each trial was analyzed separately to produce adjusted cultivar means. In this paper, the model is written using the notation introduced by Wilkinson and Rogers (1973) and applied in Patterson (1997) and Piepho et al. (2003). The linear mixed model used in the first-stage is written as

$$C + F + C{\cdot}F : R + R{\cdot}F + R{\cdot}B$$

where C is the cultivar, F is the fungicide treatment, R is the replicate, and B is the incomplete block within a replicate. The fixed effects are specified before the colon and the random

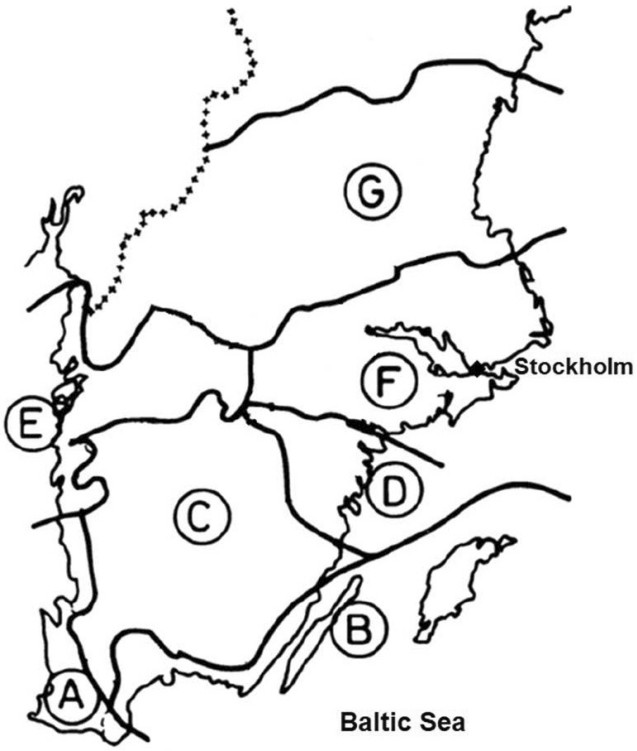

Fig. 1. Swedish agricultural zones.

effects after the colon. The dot between two factors indicates a crossed effect. The response variable (i.e., the yield), the intercept, and the residual error term are implicit.

The adjusted cultivar means from the first stage were used for zone-level analysis in the second stage, without any weighting, since the current Swedish practice is a two-stage unweighted approach. Therefore, there was only a single value of DMY per cultivar and trial, and no residual plot error information was available in the second stage. The CV study was performed using the adjusted means of cultivars from the first stage. In other words, the CV study focused on the improving the model in the second stage. The baseline model, which is currently used in Swedish cultivar testing, can be written as

$$C + Z + F + C{\cdot}Z + C{\cdot}F + Z{\cdot}F + C{\cdot}Z{\cdot}F : L + C{\cdot}L$$

where Z is the zone and L is the trial. For example, C·Z represents the cultivar × zone interaction. Trials are always nested within zones. In this study, there were three zones, and the number of fungicide treatments was two, but generally, there could be J zones and K fungicide treatments. Moreover, there are V cultivars, and T_j trials in the jth zone, $j = 1, 2, \ldots, J$. With this baseline model, the cultivar × zone interaction effect is fixed, and so it is not possible to borrow information across zones. Thus, in an alternative model, we assign the effects of cultivar to be random, and so the effects of the cultivar × zone term will be random to allow "borrowing strength" (Piepho et al., 2016) across zones.

We compared the 20 linear mixed models listed in Table 1. The single-year (S) series models with fixed (F) effects of cultivars are called SF models, and the single-year series models with random (R) effects of cultivars are called SR models. To facilitate readability, the 20 models in Table 1 were categorized into five groups. There were 17 SR models and three SF models. Below, we make some remarks regarding the covariance structures that are listed in the "covariance structure" column of Table 1:

1. The covariance structure for L (trials) was $\mathbf{G}_L = \oplus_{j=1}^{J} \mathbf{G}_{L(j)}$, where $\mathbf{G}_{L(j)}$ is a $T_j \times T_j$ diagonal matrix with diagonal elements $\sigma^2_{L(j)}$. In other words, zone-specific variances were assumed. This structure was applied in all SR models, except the models SR 16 and 17 and the models SF 1 to 3. The models SR 16 and 17 used an identity covariance structure ($\sigma^2\mathbf{I}$) for trials, because these models included no effects of zones.
2. The SR factor analytic (FA) models (SR 8–11) used an FA covariance structure with a single multiplicative term for cultivar × zone interaction effects. This covariance structure is block-diagonal with blocks

$$\begin{bmatrix} \lambda_1^2 & \lambda_1\lambda_2 & \cdots & \lambda_1\lambda_j \\ \lambda_2\lambda_1 & \lambda_2^2 & \cdots & \lambda_2\lambda_j \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_j\lambda_1 & \lambda_j\lambda_2 & \cdots & \lambda_j^2 \end{bmatrix}$$

   The off-diagonal elements of these blocks are products of parameters λ_j and $\lambda_{j'}$, where the j and j' subscripts refers to the jth and j'th zone, respectively. Thus, interaction effects from the same cultivar but from different zones are correlated. Interaction effects from different cultivars are uncorrelated. The diagonal elements are zone-specific interaction variances. The FA structure is used to take into account heterogeneity of variance in the cultivar × zone interaction term.
3. The covariance structure for cultivar effects was the identity structure (i.e., $\mathbf{G}_C = \sigma^2\mathbf{I}$) for all models, except for the models SR 12 and 14. For these models, a heterogeneous covariance structure was used (i.e., $\mathbf{G}_C = \oplus_{j=1}^{J} \mathbf{G}_{C(j)}$, where $\mathbf{G}_{C(j)}$ is a $V \times V$ diagonal matrix with diagonal elements $\sigma^2_{V(j)}$. According to this structure, variances of cultivar effects are zone specific.
4. The heterogeneous residual covariance structure was zone specific: $\mathbf{R} = \oplus_{j=1}^{J} \mathbf{R}_j$, where $\mathbf{R}_j$ is a $VKT_j \times VKT_j$ diagonal matrix with diagonal elements $\sigma^2_{\mathbf{R}(j)}$. The heterogeneous residual covariance structure was applied to models SR 4, 7, 14, and 15 and SF 19 and 20. For all other models, the residual covariance structure was homoscedastic (i.e., $\mathbf{R} = \sigma^2\mathbf{I}$).
5. For the models SF 1 to 3, the covariance structure of L (trials) was $\mathbf{G}_L = \oplus_{k=1}^{K} \oplus_{j=1}^{J} \mathbf{G}_{L(jk)}$, where $\mathbf{T}_{jk}$ is a $T_j \times T_j$ diagonal matrix where all diagonal elements are $\sigma^2_{JK(jk)}$, $j = 1, 2, \ldots, J$; $k = 1, 2, \ldots, K$. In other words, the variance structure for trials was assumed to be zone × fungicide specific.

Cross-Validation

We conducted a CV for model evaluation. The models were evaluated by a twofold CV. In the first fold, the trials were randomized

equally (50/50) within zones to a training dataset A_1 and a validation dataset A_2. In the second fold, A_2 was used as the training dataset, and A_1 was used as the validation dataset. The reason for conducting this type of CV was the decreasing number of trials in recent years. We wanted to train the model with a small number of trials. If the CV was conducted with many folds, then there would be many trials included in the training set, which does not represent the current situation in Swedish cultivar testing. Thus, we preferred a twofold CV. In general, cultivar trials aim to predict differences between tested cultivars rather than means. Piepho (1998) proposed the mean squared error of prediction (MSEP) to assess the accuracy of estimates of differences between cultivars in various environments. In this study, we used a measure similar to Piepho's MSEP based on differences for measuring the prediction accuracy of the models. The assessment was measured based on the discrepancies between observed $(y_{vkt} - y_{v'kt})$ and predicted pairwise differences $(z_{vkt} - z_{v'kt})$:

$$\text{MSEP} = \frac{\sum_{t=1}^{T}\sum_{k=1}^{K}\sum_{v=1}^{V}\sum_{v\neq 1}^{V}\left[y_{vkt} - y_{v'kt} - \left(z_{vkt} - z_{v'kt}\right)\right]^2}{TKV\left(V-1\right)} \quad [1]$$

Table 1. The 20 statistical models used in the single-year series cross-validation.

Model group	Model† name	Fixed terms‡	Random terms§	Covariance structure¶
Basic SR models	SR 1	Z + F + Z·F	C + L + F·L + C·Z + C·L + C·F + **C·L·F** + C·Z·F	$\mathbf{G}_L = \oplus_{j=1}^{J}\mathbf{G}_{L(j)}$
	SR 2	Z + F + Z·F	C + L + **F·L** + C·Z + C·L + C·F + C·Z·F	$\mathbf{G}_C = \sigma^2\mathbf{I}$
	SR 3	Z + F + Z·F	C + L + C·L + C·Z + C·F + C·Z·F	$\mathbf{R} = \sigma^2\mathbf{I}$
	SR 4	Z + F + Z·F	C + L + **C·L** + C·Z + C·F + C·Z·F	$\mathbf{G}_L = \oplus_{j=1}^{J}\mathbf{G}_{L(j)}$ $\mathbf{G}_C = \sigma^2\mathbf{I}$ $\mathbf{R} = \oplus_{j=1}^{J}\mathbf{R}_j$
	SR 5	Z + F + Z·F	C + L + C·Z + C·F + **C·Z·F**	$\mathbf{G}_L = \oplus_{j=1}^{J}\mathbf{G}_{L(j)}$
	SR 6	Z + F + Z·F	C + L + C·Z + C·F	$\mathbf{G}_C = \sigma^2\mathbf{I}$ $\mathbf{R} = \sigma^2\mathbf{I}$
	SR 7	Z + F + Z·F	C + L + C·Z + C·F	$\mathbf{G}_L = \oplus_{j=1}^{J}\mathbf{G}_{L(j)}$ $\mathbf{G}_C = \sigma^2\mathbf{I}$ $\mathbf{R} = \oplus_{j=1}^{J}\mathbf{R}_j$
FA models	SR 8	Z + F + Z·F	L + F·L + C·L + C·F + C·Z + **C·Z·F**	$\mathbf{G}_L = \oplus_{j=1}^{J}\mathbf{G}_{L(j)}$
	SR 9	Z + F + Z·F	L + **F·L** + C·L + C·F + C·Z	$\mathbf{G}_C = \sigma^2\mathbf{I}$
	SR 10	Z + F + Z·F	L + **C·L** + C·F + C·Z	Var(C·Z) = λλ′
	SR 11	Z + F + Z·F	L + C·F + C·Z	$\mathbf{R} = \sigma^2\mathbf{I}$
No random interactions	SR 12	Z + F + Z·F	C + L	$\mathbf{G}_L = \oplus_{j=1}^{J}\mathbf{G}_{L(j)}$ $\mathbf{G}_C = \oplus_{j=1}^{J}\mathbf{C}_j$ $\mathbf{R} = \sigma^2\mathbf{I}$
	SR 13	Z + F + Z·F	C + L	$\mathbf{G}_L = \oplus_{j=1}^{J}\mathbf{G}_{L(j)}$ $\mathbf{G}_C = \sigma^2\mathbf{I}$ $\mathbf{R} = \sigma^2\mathbf{I}$
	SR 14	Z + F + Z·F	C + L	$\mathbf{G}_L = \oplus_{j=1}^{J}\mathbf{G}_{L(j)}$ $\mathbf{G}_C = \oplus_{j=1}^{J}\mathbf{C}_j$ $\mathbf{R} = \oplus_{j=1}^{J}\mathbf{R}_j$
	SR 15	Z + F + Z·F	C + L	$\mathbf{G}_L = \oplus_{j=1}^{J}\mathbf{G}_{L(j)}$ $\mathbf{G}_C = \sigma^2\mathbf{I}$ $\mathbf{R} = \oplus_{j=1}^{J}\mathbf{R}_j$
No zones	SR 16	F	C + L + F·L + C·L + C·F + **C·F·L**	$\mathbf{G}_C = \sigma^2\mathbf{I}$
	SR 17	F	C + L + F·L + C·L + C·F	$\mathbf{R} = \sigma^2\mathbf{I}$
SF models	SF 1	C + F + C·F	Z + L + C·Z + C·L	$\mathbf{G}_L = \oplus_{k=1}^{K}\oplus_{j=1}^{J}\mathbf{G}_{L(jk)}$ $\mathbf{G}_C = \sigma^2\mathbf{I}$ $\mathbf{R} = \sigma^2\mathbf{I}$
	SF 2	C + F + C·F	Z + L + C·Z + C·L	$\mathbf{G}_L = \oplus_{k=1}^{K}\oplus_{j=1}^{J}\mathbf{G}_{L(jk)}$ $\mathbf{G}_C = \sigma^2\mathbf{I}$ $\mathbf{R} = \oplus_{j=1}^{J}\mathbf{R}_j$
	SF 3#	C + Z + F + C·Z + C·F + Z·F + C·Z·F	L + C·L	$\mathbf{G}_L = \oplus_{k=1}^{K}\oplus_{j=1}^{J}\mathbf{G}_{L(jk)}$ $\mathbf{G}_C = \sigma^2\mathbf{I}$ $\mathbf{R} = \oplus_{j=1}^{J}\mathbf{R}_j$

† SR, single-year series, random effects of cultivars; SF, single-year series, fixed effects of cultivars.

‡ C, cultivar; F, fungicide; L, trial; Y, year; Z, zone; FA, factor analytic.

§ Bolding indicates that the term is dropped in the next model.

¶ $\mathbf{G}_L$, matrix structure of trial; $\mathbf{G}_C$, matrix structure of cultivar; $\mathbf{R}$, matrix structure of residual term.

The current-practice model.

where y_{vkt} and z_{vkt} is the observed yield and the predicted yield, respectively, of the cth cultivar in the tth trials, using the kth fungicide treatment, and $T = \sum_{j=1}^{J} T_j$. We ranked the model performance based on the average single-year MSEP for each crop (i.e., the mean of eight MSEPs for winter wheat [based on eight single-year datasets] and the mean of five MSEPs for spring barley [based on five single-year datasets]). The CVs for the FA models and the models with the heterogeneous residuals were performed using PROC MIXED and the other models were fitted using PROC HPMIXED in SAS (SAS Institute, 2013). We used the PROC HPMIXED to reduce the computational time.

Multiyear Series Cross-Validation

Swedish Multi-environment Trial Data

In multiyear series CV, we used the DMY data of winter wheat and spring barley from 2007 until 2016, and these datasets were imbalanced. The number of trials and cultivars in each year for winter wheat and spring barley are summarized in Supplemental Tables S3 and S4, respectively. In this multiyear CV, we aimed to search the best model for the 5-yr series analysis.

Linear Mixed Models

In the 5-yr series CV, year (Y) is an additional factor to be included in the model. A total of 11 models were compared for the 5-yr series CV, as detailed in Table 2. The 10 multiyear (M) series models with random (R) effects of cultivars are called MR models. The multiyear series model with fixed (F) effects of cultivars is called the MF model and is the model used in current practice. The MR 1 model is a basic saturated model. The following three MR models (MR 2–4) were obtained by dropping, one at a time, the single term with the smallest variance. From model MR 5, we omitted the year × fungicide (Y·F) and year × zone × fungicide (Y·Z·F) interactions. In models MR 6 to 8, either the C·Z·F interaction or the C·Z·Y interaction, or both these interactions, were removed. Models MR 9 and 10 are models without effects of zones. The MF model is currently used in practice and was fitted per zone and fungicide treatment.

The covariance structure for L (trials) was $\mathbf{G}_{\mathrm{L}(j)} = \oplus_{j=1}^{J} \mathbf{T}_j$, where $\mathbf{T}_j$ is a $T_j \times T_j$ diagonal matrix where all diagonal elements are $\sigma^2_{\mathrm{L}(j)}$ (i.e., zone-specific variances), except for models MR 9 and 10, since there were no zone effects in those models. For all models, the residual structure was homogeneous ($\sigma^2\mathbf{I}$). Due to convergence problems, it was not possible to fit a residual structure with heterogeneity of variance between zones and years. Note that the residual term here pertains to the highest order interaction and comprises both this interaction and the error that is associated with the genotype–environment mean in the **R** matrix, not just the residual plot error.

Cross-Validation

As we mentioned before, we aimed to search the best model for the 5-yr series analysis. For that reason, we conducted a modified leave-one-out CV to mimic the current Swedish practice of predicting cultivar performance based on results from 5 yr. A set of data from five subsequent years was used as a training set. The following sixth year was used as a validation set, as depicted in Fig. 2. For example, the dataset of yields from 2007 to 2011 was assigned as the training dataset, and the dataset from 2012 was assigned as the validation dataset. Another reason why the CV was done in chronological order was that the set of cultivars in the early years and recent years differ a lot. For example, when the training set comprises recent years and the validation set comprises early years, then there will be very few cultivars in common between both sets. Most of the cultivars that are predicted in the training set would not be available in the validations set, since the validation set comprises early years. Thus, to meet the purpose of this study (i.e., prediction of future yield performance), we conducted the CV in chronological order.

In each CV, we computed MSEP (Eq. [1]). The best model was the one that produced the smallest MSEP, since that model predicted the yield of the following year most accurately. The models were ranked based on the mean of MSEP over the six CV sets. For the multiyear series, the CV was done using PROC HPMIXED in SAS (SAS Institute, 2013).

Table 2. The 11 statistical models used in the multiyear series cross-validation.

Model group	Model name†	Fixed terms‡	Random terms§
Basic MR models	MR 1	Z + F + Z·F	C + L + Y + C·Z + C·Y + C·L + C·F + Y·Z + Y·F + F·L + Y·Z·F + C·Z·F + C·Z·Y + C·F·Y + **C·Z·F·Y**
	MR 2	Z + F + Z·F	C + L + Y + C·Z + C·Y + C·L + C·F + Y·Z + Y·F + F·L + Y·Z·F + C·Z·F + C·Z·Y + **C·F·Y**
	MR 3	Z + F + Z·F	C + L + Y + C·Z + C·Y + C·L + C·F + Y·Z + Y·F + F·L + Y·Z·F + **C·Z·F** + C·Z·Y
	MR 4	Z + F + Z·F	C + L + Y + C·Z + C·Y + C·L + C·F + Y·Z + Y·F + F·L + Y·Z·F + C·Z·Y
BLUP without Y·F & Y·Z·F	MR 5	Z + F + Z·F	C + L + Y + C·Z + C·Y + C·L + C·F + Y·Z + F·L + C·Z·F + C·Z·Y
	MR 6	Z + F + Z·F	C + L + Y + C·Z + C·Y + C·L + C·F + Y·Z + F·L + C·Z·Y
	MR 7	Z + F + Z·F	C + L + Y + C·Z + C·Y + C·L + C·F + Y·Z + F·L + C·Z·F
	MR 8	Z + F + Z·F	C + L + Y + C·Z + C·Y + C·L + C·F + Y·Z + F·L
BLUP no zone	MR 9	F	C + L + Y + F·L + F·Y + C·L + C·F + C·Y + **C·F·Y**
	MR 10	F	C + L + Y + F·L + F·Y + C·L + C·F + C·Y
BLUE	MF¶	C	L + Y + C·Y

† MR, multiyear series, random effects of cultivars; MF, multiyear series, fixed effects of cultivars.

‡ C, cultivar; F, fungicide; L, trial; Y, year; Z, zone.

§ Bolding indicates that the term is dropped in the next model.

¶ The current-practice model.

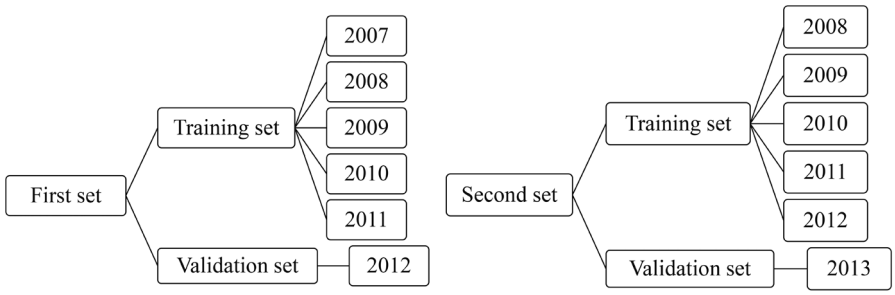

Fig. 2. Illustration of multiyear cross-validation scheme.

RESULTS

Cross-Validation: Random vs. Fixed Effects of Cultivars

The MSEP average of winter wheat and spring barley single-year datasets are listed in Table 3. In general, the E-BLUE (SR) models achieved lower MSEPs than the E-BLUP (SF) models for both crop datasets. The five top-performing models are presented above the dashed horizontal line. For both crops, the SR 5 model performed the best.

The SF models performed poorly for both crop datasets. For winter wheat, the MSEP means of the E-BLUE (SF 1–3) were higher than the MSEP means of most SR models. The current-practice model (SF 3) was the worst performing among the SF models. Only the SR 13 and SR 15 models, which have no interaction terms, performed worse than the current-practice SF model. For spring barley, the current model (SF 3) performed the worst among all investigated models, since it had the largest MSEP.

For the models without zones, particularly SR 16, the results were considerably different between the two crops. The SR 16 model was ranked as the second best model in winter wheat, whereas in spring barley, it was ranked the 16th best model. On the other hand, the SR 17 model was the sixth best model for both crops.

The BLUP with FA structure (SR 8–11) did not perform better than the simpler model (i.e., SR 5). The SR 8 and SR 10 models were the best BLUP with FA structure for winter wheat and spring barley, respectively. The SR model with more interaction terms (SR 1) and the SR model with heterogeneous residual variance (SR 4 and SR 7) were worse-performing than the more parsimonious model SR 5.

Table 4 presents MSEP means of multiyear series in winter wheat and spring barley. The currently used MF model was the most unfavorable model, since for both crops this model showed the largest average MSEP. For winter wheat, the MR 5 model, which does not include the Y·F and Y·Z·F interactions but includes the C·Z·F and C·Z·Y interactions, was the best model in terms of average MSEP.

Table 3. Mean squared error of prediction differences from single-year series cross-validation of winter wheat (*N* = 8) and spring barley (*N* = 5).

	Winter wheat		Spring barley	
Ranking	Model†	Mean	Model	Mean
		$g^2\ m^{-4}$		$g^2\ m^{-4}$
1	SR 5	7017	SR 5	1815
2	SR 16	7032	SR 6	1815
3	SR 2	7037	SR 7	1824
4	SR 8	7041	SR 3	1827
5	SR 6	7046	SR 10	1829
6	SR 17	7054	SR 17	1830
7	SR 9	7083	SR 4	1834
8	SR 1	7085	SR 9	1838
9	SR 3	7103	SR 8	1842
10	SR 10	7118	SR 2	1847
11	SF 1	7148	SR 11	1850
12	SR 7	7172	SR 13	1855
13	SF 2	7197	SF 1	1863
14	SR 4	7235	SR 15	1868
15	SR 14	7271	SF 2	1874
16	SR 11	7278	SR 16	1889
17	SR 12	7291	SR 14	1908
18	SF 3‡	7313	SR 12	1911
19	SR 13	7826	SR 1	1914
20	SR 15	8488	SF 3‡	2053

† SR, single-year series, random effects of cultivars; SF, single-year series, fixed effects of cultivars.

‡ SF 3 is the currently used model in Swedish cultivar testing.

Table 4. Mean squared error of prediction differences from multiyear series cross-validation of winter wheat (*N* =6) and spring barley (*N* =6).

	Winter wheat		Spring barley	
Ranking	Model†	Mean	Model	Mean
		$g^2\ m^{-4}$		$g^2\ m^{-4}$
1	MR 5	7718	MR 7	2092
2	MR 3	7718	MR 3	2094
3	MR 2	7736	MR 5	2094
4	MR 10	7739	MR 2	2094
5	MR 1	7743	MR 1	2095
6	MR 6	7744	MR 8	2096
7	MR 4	7745	MR 4	2097
8	MR 9	7758	MR 6	2097
9	MR 7	7813	MR 10	2124
10	MR 8	7830	MR 9	2125
11	MF‡	8596	MF‡	2320

† MR, multiyear series, random effects of cultivars; MF, multiyear series, fixed effects of cultivars.

‡ MF is the currently used model in Swedish cultivar testing.

For spring barley, the MR 7 model was the best, although it performed less well in winter wheat. The MR 5 model, which was the top performing in winter wheat, was ranked the third best model in spring barley. The MR 3, MR 2, MR 1, and MR 5 models were among the five best-performing models in both winter wheat and spring barley. The MR models without zones (MR 9 and MR 10) did not perform well in the spring barley 5-yr series, whereas in winter wheat, MR 10 was still ranked among the five best models.

Application of the SR 5 and MR 5 Models in Winter Wheat Datasets

We present the application of the SR 5 model in winter wheat dataset 2016 and the MR 5 model in winter wheat dataset 2012 to 2016. The fixed effects significance tests for both examples are given in Table 5, and the variance components are tabulated in Table 6. In the SR 5 model, the fixed effects of fungicide treatment and the interaction of fungicide × zone were significant, whereas in the MR 5 model, only the fixed effects of fungicide treatment were significant. The fixed effects were tested using approximate *F* statistics with Kenward–Roger adjustment for the denominator of degree freedom. For the random effects, the variance components were tested using Wald *Z* tests, considering the sample size was large. It should be noted, however, that this test is not reliable unless the sample size is large (Stroup, 2012). In the SR 5 and MR 5 models, the cultivar × zone (C·Z) variance was relatively small compared with the inter-trial (L) variances. These inter-trial variance components within zones were large, which showed that the trials within zones were heterogeneous. The cultivar × zone (C·Z) variance is used to predict the C·Z means. Therefore, based on these examples, it can be seen that the ranking of predictions of cultivar yield might be similar across zones, since the interaction of cultivar × zone was relatively modest.

Table 7 presents an example of different cultivar ranking between E-BLUE with the SF 3 model and E-BLUP with the SR 5 model in the winter wheat 2016 single-year series dataset. The DMY predictions were smaller using E-BLUP than using E-BLUE in some cultivars (e.g., Etana, G 0512LT3, and Brons). The smaller values using E-BLUP were a consequence of "shrinkage." Means higher than the overall mean are shrunken downward to the overall mean, as can also be seen for some cultivars such as Festival and Rivero, which were not listed among the best 10 cultivars by the E-BLUP method. This shrinkage property avoids otherwise overoptimistic estimates of cultivar performance. On the other hand, the means that are lower than the overall mean are slightly increased (shrunken upward towards the overall mean) using SR models. Thus, the shrinkage property also mitigates too pessimistic predictions of performance for relatively poor cultivars. For example, regarding the best performers, Ohio and RGT Reform were not listed among the best 10 cultivars in the E-BLUE model but were listed among the best 10 cultivars by the E-BLUP method. Table 7 also shows that the ranking of the cultivars is different between

Table 5. Evaluation of fixed effects of the single-year model (SR 5) in winter wheat 2016 and the multiyear model (MR 5) in winter wheat 2012 to 2016.

	Approximate *F* statistics	
Source of variation	SR 5 model	MR 5 model
Zone (Z)	0.09	2.13
Fungicide (F)	115.67**	153.57**
Z·F	4.16*	2.99

*, ** Significant at the 0.05 and 0.01 probability levels, respectively.

Table 6. Evaluation of fixed effects of the single-year model (SR 5) in winter wheat 2016 and the multiyear model (MR 5) in winter wheat 2012 to 2016.

		Variance components	
Source of variation†	Group	SR 5 model	MR 5 model
L	South	22,949	14,756.67**
	Middle	84,136.00*	32,782.51**
	North	21,126	14,324.75**
C		637.06**	2,288.14**
C·Z		273.76**	26.44
C·F		13.08	519.74**
C·Z·F		0	232.64**
C·L			1,587.31**
F·L			2,037.51**
Y			3,044.62
Y·Z			0
C·Y			341.34**
C·Z·Y			227.12**

*, **Significant at the 0.05 and 0.01 probability levels, respectively.

† L, trial; C, cultivar; Z, zone; F, fungicide; Y, year.

Table 7. Example of different cultivar ranking in the winter wheat 2016 from Zone A, fungicide treated. More than half of the cultivars differed in ranking.

	E-BLUE (SF 3)†		E-BLUP (SR 5)‡	
Cultivar	Ranking	DMY	Ranking	DMY
		g m^{-2}		g m^{-2}
Brons	3	915	6	900
Creator	5	913	9	898
Effekt	7	905	3	908
Ellen	4	913	4	906
Etana	2	938	1	928
Festival	6	907	–	–
G 0512LT3	1	963	2	912
Mariboss	9	903	10	893
Ohio	–	–	5	903
RGT Reform	–	–	8	898
Rivero	8	904	–	–
Rockefeller	10	903	7	899

† E-BLUE, empirical best line unbiased estimation; SF, single-year and fixed effects for cultivar; DMY, dry matter yield.

‡ E-BLUP, empirical best line unbiased prediction; SR, single-year and random effects for cultivar.

the E-BLUE and E-BLUP methods. The best cultivar according to the E-BLUE method was cultivar G 0512LT3, whereas using the E-BLUP method, cultivar Etana was the best. The ranking of the other cultivars was also different between the two models. For cultivar recommendation, where a correct ranking of cultivars is essential, the E-BLUP method should be preferred due to its smaller MSEP.

The example of different cultivar ranking between E-BLUE with the MF model and E-BLUP with MR 5 model in the winter wheat dataset from 2012 to 2016 is tabulated in Table 8. Again, we can see a considerable shrinkage in the DMY predictions using E-BLUP in some cultivars (e.g., G0512LT, Lw 06W607-10, RGT Universe, and Torp). The 10 top-performing cultivars also differed a lot between the E-BLUE and E-BLUP methods. For example, G0512LT was the best cultivar according to E-BLUP, whereas RGT Universe was the best cultivar with E-BLUE (ranked sixth in E-BLUP). Therefore, this example shows clearly that the ranking between E-BLUE and E-BLUP differed a lot and that E-BLUP provided more accurate ranking due to shrinkage, as indicated by the lowest MSEP in the MR 5 model. The best variety according to E-BLUE was ranked sixth by the E-BLUP method. Some varieties that were not listed among the 10 best cultivars in the E-BLUE model were listed among the 10 best cultivars by the E-BLUP method (e.g., Hereford, Audi, and Hymack). Again, this example reaffirmed the CV results, suggesting that for cultivar recommendation, where a correct ranking of cultivars is critical, the E-BLUP method should be preferred.

Table 8. Example of different winter wheat cultivar ranking in the multiyear analysis (2012–2016) from Zone A, fungicide treated. More than half of the cultivars differed in ranking.

	E-BLUE (MF)†		E-BLUP (MR 5)‡	
Cultivar	Ranking	DMY	Ranking	DMY
		g m^{-2}		g m^{-2}
Hereford	—	—	7	1047
Audi	—	—	9	1046
Hymack	—	—	8	1047
Sj 6286003	—	—	10	1045
Memory	6	1076	3	1059
SJ 7343505	4	1081	5	1053
Torp	5	1078	4	1054
R 11224	10	1067	—	—
G0512LT	3	1092	1	1060
Lw 08DH642-26	2	1142	2	1059
Lw 06W607-10	1	1143	6	1053
Hacksta	9	1069	—	—
RGT Universe	8	1073	—	—
Maradona	7	1076	—	—

† E-BLUE, empirical best line unbiased estimation; MF, multiyear and fixed effects for cultivar; DMY, dry matter yield.

‡ E-BLUP, empirical best line unbiased prediction; MR, multiyear and random effects for cultivar.

DISCUSSION

In this study, the random-cultivar-effects models outperformed the currently used fixed-cultivar-effects models, in both single-year and multiyear series datasets. The model with random effects of cultivars improved the prediction accuracy for zone-based yield with a few trials, as reflects the current situation in Swedish cultivar trials, because with random effects for cultivar, information is borrowed across zones. In winter wheat, the number of cultivars varied between years, with a larger number of cultivars in recent years (2014–2016) than in earlier years. Still, the E-BLUP/SR models performed better than the E-BLUE/SF models in terms of MSEP. Thus, our study confirmed the conclusions of previous simulation studies that random-cultivar-effects models are preferable for routine zoned-based yield prediction compared with fixed-cultivar-effects models (Piepho and Möhring, 2006; Kleinknecht et al., 2011). Furthermore, it has been reported by Piepho and Möhring (2006) that the main advantage of random-cultivar-effects models over fixed-cultivar-effects models is that mean squared error becomes smaller.

The SR models with FA structure (SR 8–11) performed worse than the simpler model SR 5. The investigated FA covariance structure allows heterogeneous variances and unique pairwise correlations between zones. The FA structure is useful because it allows heterogeneous variance and covariance using fewer parameters than the unstructured covariance structure. However, the restricted maximum likelihood estimation for the FA structure and the model with many interaction terms combined with heterogeneous residual structure were computationally very demanding. For this reason, combinations of FA structures for interaction effects and heterogeneous structures for residual effects were not explored. The application of the FA structure may be more useful when the number of zones is larger than three.

The empirical datasets that we used here were not perfectly normally distributed, which is showed by the residual diagnostics in Supplemental Fig. S1 and S2. However, BLUP per se does not require normality (Searle et al., 1992, p. 270 and 273). The mixed-model equations can be derived from the equations for BLUP without assuming the normal distribution (Satoh, 2018). Therefore, it was expected that our CV results would reveal that random-cultivar-effects models perform better than fixed-cultivar-effects models. In practice, the variance components are unknown and must be estimated. Restricted maximum likelihood estimates may be imprecise in small datasets, so the benefits of using random-cultivar-effects models are uncertain. The simulation study from Forkman and Piepho (2013) reported, however, that imprecise variance component estimates were not a severe problem for the application of E-BLUP in small randomized complete block experiments.

We recommend striving for complete datasets for single-year analysis. Forkman (2013) showed that analyses of incomplete datasets using generalized least squares based on mixed models with random environmental effects can give unexpected estimates. For example, in a series of cultivar trials, the estimated difference between a test cultivar and the reference cultivar may be outside the range of the differences observed within the trials. Also, in a series of multiyear cultivar trials, the estimated difference between a test cultivar and the reference cultivar may not agree with previously reported yearly estimates of differences. In Sweden, it has been a common practice to decide which cultivars should be tested in particular zones, depending on their expected performance in those zones. Specifically, cultivars might not be tested in a zone if they are expected to perform worse in that zone. In this case, the cultivars are not missing at random. If there is doubt that cultivars are missing at random, it might be better to use a model with fixed effects of trials because comparisons among cultivars are then based exclusively on within-trial information and between-trial information is not recovered (Piepho et al., 2012). In the multiyear series, the Swedish practice has been to exclude from the analysis all cultivars that have not been tested in the latest year and at least 2 yr. We recommend that all cultivars should be retained in the analysis. As pointed out by Piepho and Möhring (2006), all cultivars involved in selection decisions should be included in the analysis to avoid selection bias. Piepho and Möhring (2006) also mentioned that removal of data leads to a missing-not-at-random pattern that causes invalid variance component estimates. Moreover, if the missing data pattern is missing-not-at-random, then E-BLUP will systematically be associated with a varying degree of shrinkage, which causes bias. For example, if a cultivar is tested very little, then the shrinkage of all its predicted effects will be large, and thus the prediction will be less accurate.

Based on the single-year CV and multiyear series CV, we did not observe an increase in the prediction accuracy with the higher order interaction effects added or with complex variance-covariance structures compared with models that are more parsimonious and straightforward to fit. Employing the heterogeneous covariance structure for residual effects in the multiyear analyses may be useful to have variance components differing between years or zones. However, in the single-year series CV, the model with heterogeneous variance in the **R** matrix (SR 4) did not perform well compared with the model with homogeneous residual variance. Furthermore, the computation time will be increased and a convergence issue may occur when applying a heterogeneous residual structure. A higher number of interaction terms or a more complex variance-covariance structure may cause overfitting that may decrease the accuracy of predictions. In the single-year CV, the SR 5 model outperformed the other models. It is reasonable to choose this model, since it is a well-performing model that requires less computation time than the more complex models. In the 5-yr series, either the MR 3 or MR 5 model may be chosen, since the differences of MSEP between these models were subtle in both crops. The CV study was preferred to merely using an information criterion like the Akaike information criterion (AIC) for model selection, because a CV study examines whether a model can produce an accurate prediction or not and thus gives a measure (MSEP) of the size of the prediction errors.

In winter wheat, the random-cultivar-effects models without zones performed well for single-year datasets and modestly for 5-yr series datasets, which demonstrates that trials in winter wheat are heterogeneous within zones (as shown in Table 6), as compared with spring barley. A plausible biological reason is that winter wheat is grown in winter weather conditions with large local variation, as compared with spring barley, which is sown in the springtime and therefore grown under less diverse local conditions. In the wintertime, the environmental conditions vary locally, from mild and humid to cold and dry, causing different stress factors to predominate (Olsen et al., 2018).

CONCLUSION

We performed a thorough CV study to assess the performance of random-cultivar-effects and fixed-cultivar-effects models for single and multiyear empirical datasets of winter wheat and spring barley of Swedish cultivar trials. We also presented evidence that borrowing strength across zones from the random effects of cultivar increased the accuracy of zone-based yield prediction. The CV results, based on the MSEP, showed that using more interaction terms, (e.g., F·L, C·L·F, C·F·Y, or C·Z·F·Y) or fitting more complex variance-covariance structures was not necessary. However, it was essential to incorporate zone in the analysis. For these reasons, we concluded that for the routine analysis of single-year series, the SR 5 model (Z + F + Z·F : C + L + C·Z + C·F + C·Z·F) should be used instead of the currently used model. For the multiyear series, we recommend the MR 5 model, Z + F + Z·F : C + L + Y + C·Z + C·Y + C·L + C·F + Y·Z + F·L + C·Z·F + C·Z·Y, for a routine procedure.

Supplemental Material

The manuscript has six supplemental materials (i.e., four tables and two figures). Supplemental Tables S1 and S2 show the number of trials and cultivars in each zone for winter wheat in single-year and multiyear CV, respectively. Supplemental Tables S3 and S4 show the number of trials and cultivars in each zone for spring barley in single-year and multiyear CV, respectively. Supplemental

Fig. S1 and S2 show the residual plot of the SR 5 and MR 5 models in winter wheat datasets.

Conflict of Interest

The authors declare that there is no conflict of interest.

Acknowledgments

This research was funded by Stiftelsen lantbruksforskning-Swedish farmers' foundation for agricultural research (Project nr O-17-20-963).

References

Forkman, J. 2013. The use of a reference variety for comparisons in incomplete series of crop variety trials. J. Appl. Stat. 40:2681–2698. doi:10.1080/02664763.2013.825703

Forkman, J., and H.-P. Piepho. 2013. Performance of empirical BLUP and Bayesian prediction in small randomized complete block experiments. J. Agric. Sci. 151:381–395. doi:10.1017/S0021859612000445

Galwey, N.W. 2014. Introduction to mixed modelling: Beyond regression and analysis of variance. 2nd ed. John Wiley & Sons, Chichester, UK.

Kleinknecht, K., F. Laidig, H.P. Piepho, and J. Möhring. 2011. Best linear unbiased prediction (BLUP): Is it beneficial in official variety performance trials? Biuletyn Oceny Odmian 33:21–33.

Kleinknecht, K., J. Möhring, K.P. Singh, P.H. Zaidi, G.N. Atlin, and H.-P. Piepho. 2013. Comparison of the performance of best linear unbiased estimation and best linear unbiased prediction of genotype effects from zoned Indian maize data. Crop Sci. 53:1384–1391. doi:10.2135/cropsci2013.02.0073

McCulloch, C.E., S.R. Searle, and J.M. Neuhaus. 2008. Generalized, linear and mixed models. 2nd ed. John Wiley & Sons, Hoboken, NJ.

Möhring, J., and H.-P. Piepho. 2009. Comparison of weighting in two-stage analysis of plant breeding trials. Crop Sci. 49:1977–1988. doi:10.2135/cropsci2009.02.0083

Olsen, A.K.B., T. Persson, A. Wit, L. Nkurunziza, E. Sindhøj, and H. Eckersten. 2018. Estimating winter survival of winter wheat by simulations of plant frost tolerance. J. Agron. Crop Sci. 204:62–73. doi:10.1111/jac.12238

Patterson, H.D. 1997. Analysis of series of variety trials. In: R.A. Kempton and P.N. Fox, editors, Statistical methods for plant variety evaluation. Chapman & Hall, London. p. 139–161. doi:10.1007/978-94-009-1503-9_9

Patterson, H.D., and V. Silvey. 1980. Statutory and recommended list trials of crop cultivars in the United Kingdom. J. R. Stat. Soc. Ser. A 143:219–252. doi:10.2307/2982128

Piepho, H.-P. 1998. Empirical best linear unbiased prediction in cultivar trials using factor-analytic variance-covariance structures. Theor. Appl. Genet. 97:195–201. doi:10.1007/s001220050885

Piepho, H.-P., A. Büchse, and K. Emrich. 2003. A hitchhiker's guide to mixed models for randomized experiments. J. Agron. Crop Sci. 189:310–322. doi:10.1046/j.1439-037X.2003.00049.x

Piepho, H.-P., and J. Möhring. 2006. Selection in cultivar trials: Is it ignorable? Crop Sci. 46:192–201. doi:10.2135/cropsci2005.04-0038

Piepho, H.-P., M.F. Nazir, M. Qamar, A.-u.-R. Rattu, Riaz-ud-Din, M. Hussain, et al. 2016. Stability analysis for a countrywide series of wheat trials in Pakistan. Crop Sci. 56:2465–2475. doi:10.2135/cropsci2015.12.0743

Piepho, H.-P., E.R. Williams, and L.V. Madden. 2012. The use of two-way linear mixed models in multitreatment meta-analysis. Biometrics 68:1269–1277. doi:10.1111/j.1541-0420.2012.01786.x

Robinson, G.K. 1991. That BLUP is a good thing: The estimation of random effects. Stat. Sci. 6:15–32. doi:10.1214/ss/1177011926

SAS Institute. 2013. SAS system for Windows 9.4. SAS Inst., Cary, NC.

Satoh, M. 2018. An alternative derivation method of mixed model equations from best linear unbiased prediction (BLUP) and restricted BLUP of breeding values not using maximum likelihood. Anim. Sci. J. 89:876–879. doi:10.1111/asj.13016

Searle, S.R., G. Casella, and C.E. McCulloch. 1992. Variance components. John Wiley & Sons, New York. doi:10.1002/9780470316856

Smith, A., B.R. Cullis, and A. Gilmour. 2001. Applications: The analysis of crop variety evaluation data in Australia. Aust. N. Z. J. Stat. 43:129–145. doi:10.1111/1467-842X.00163

Stroup, W.W. 2012. Generalized linear mixed models: Modern concepts, methods and applications. CRC Press, Boca Raton, FL.

van Eeuwijk, F.A., D.V. Bustos-Korts, and M. Malosetti. 2016. What should students in plant breeding know about the statistical aspects of genotype × environment interactions? Crop Sci. 56:2119–2140. doi:10.2135/cropsci2015.06.0375

Wilkinson, G.N., and C.E. Rogers. 1973. Symbolic description of factorial models for analysis of variance. J. R. Stat. Soc. Ser. C Appl. Stat. 22:392–399. doi:10.2307/2346786