

ORIGINAL RESEARCH ARTICLE

Crop Breeding & Genetics

Cross-validation of stagewise mixed-model analysis of Swedish variety trials with winter wheat and spring barley

Harimurti Buntaran^{1,2,3}  | Hans-Peter Piepho¹ | Paul Schmidt¹  |
Jesper Rydén² | Magnus Halling³ | Johannes Forkman^{2,3} 

¹ Biostatistics Unit, Institute of Crop Science, Univ. of Hohenheim, Fruwirthstraße 23, Stuttgart 70599, Germany

² Department of Energy and Technology, Swedish Univ. of Agricultural Sciences, Box 7032, Uppsala 750 07, Sweden

³ Department of Crop Production Ecology, Swedish Univ. of Agricultural Sciences, Box 7043, Uppsala 750 07, Sweden

Correspondence

Harimurti Buntaran, Biostatistics Unit, Institute of Crop Science, University of Hohenheim, Fruwirthstraße 23, 70599 Stuttgart, Germany.

Email: harimurti.buntaran@uni-hohenheim.de

Assigned to Associate Editor Alexander Lipka.

Funding information

Stiftelsen lantbruksforskning - Swedish farmers' foundation for agricultural research, Grant/Award Number: O-17-20-963

Abstract

In cultivar testing, linear mixed models have been used routinely to analyze multi-environment trials. A single-stage analysis is considered as the gold standard, whereas two-stage analysis produces similar results when a fully efficient weighting method is used, namely when the full variance–covariance matrix of the estimated means from Stage 1 is forwarded to Stage 2. However, in practice, this may be hard to do and a diagonal approximation is often used. We conducted a cross-validation with data from Swedish cultivar trials on winter wheat (*Triticum aestivum* L.) and spring barley (*Hordeum vulgare* L.) to assess the performance of single-stage and two-stage analyses. The fully efficient method and two diagonal approximation methods were used for weighting in the two-stage analyses. In Sweden, cultivar recommendation is delineated by zones (regions), not individual locations. We demonstrate the use of best linear unbiased prediction (BLUP) for cultivar effects per zone, which exploits correlations between zones and thus allows information to be borrowed across zones. Complex variance–covariance structures were applied to allow for heterogeneity of cultivar \times zone variance. The single-stage analysis and the three weighted two-stage analyses all performed similarly. Loss of information caused by a diagonal approximation of the variance–covariance matrix of adjusted means from Stage 1 was negligible. As expected, BLUP outperformed best linear unbiased estimation. Complex

Abbreviations: σ^2_C , variance component estimate of the cultivar; σ^2_{CZ} , variance component estimate of cultivar \times zone; σ^2_{ZCL} , variance component estimate of cultivar \times location; σ^2_{ZL} , variance component estimate of location; 1S, single-stage; 2S, two-stage; AVVAR, average variance of a difference; BLUE, best linear unbiased estimation; BLUP, best linear unbiased prediction; C \times Z, cultivar \times zone; CS, compound symmetry; EBLUE, empirical BLUE; EBLUP, empirical BLUP; F, fixed effects for cultivars; FA, factor-analytic; FA1, factor-analytic order 1; FE, fully efficient; ID, identity; LR, location-specific residual variances; MET, multi-environment trial; MSEP, mean squared error of prediction differences; SW, Smith's weighting; TPE, target population of environments; U, unweighted; UN, unstructured; W, weighted; ZR, zone-specific residual variances.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2020 The Authors. *Crop Science* published by Wiley Periodicals, Inc. on behalf of Crop Science Society of America

variance–covariance structures were dispensable. To our knowledge, this study is the first to use cross-validation for comparing single-stage analyses with stage-wise analyses.

1 | INTRODUCTION

In crop variety testing, decisions on cultivar selection and recommendation are based on predictions of future performance. Cultivars are tested in multiple locations for several years. Such tests are known as multi-environment trials (METs), where an environment refers to a year–location combination. The Swedish official cultivar testing undertakes METs to provide farmers with zone-specific recommendations. The two most important crops in Sweden are winter wheat and spring barley.

The analysis of MET data is usually done by fitting linear mixed models. A single-stage analysis is preferable for theoretical reasons, since the estimation of fixed and random effects is done in a single model from plot-level data, accounting for all relevant effects in a single model (Piepho, Möhring, Schulz-Streeck, & Ogutu, 2012). A potential disadvantage of this analysis, however, is the computational effort required, especially when the numbers of cultivars and environments are large and a complex variance–covariance structure for the cultivar \times environment interaction effects is assumed (Möhring & Piepho, 2009; Welham, Gogel, Smith, Thompson, & Cullis, 2010).

A stagewise analysis splits up the analysis into two (or more) stages. Through this procedure, it is possible to reduce the computational burden substantially (Damesa, Möhring, Worku, & Piepho, 2017; Piepho et al., 2012). In Stage 1, each trial is analyzed separately via best linear unbiased estimation (BLUE) to obtain adjusted cultivar means per trial. Thus, in this stage, the cultivar effects are modeled as fixed. In Stage 2, the adjusted cultivar means from Stage 1 are analyzed jointly via an appropriate mixed model to compute marginal means for cultivars across trials. In Stage 2, the cultivar effects may be modeled as fixed or random, depending on whether cultivars are fixed or random in the corresponding single-stage model. Stagewise analysis facilitates a combined analysis of different trials with different experimental designs in Stage 1 and subsequently allows one to model structures for the heterogeneity of variance between or among trials easily (Piepho & Eckl, 2014). A major issue with conducting analyses in multiple stages is the choice of method to forward the information on precision (SEs and the variance–covariance matrix of the adjusted means) between stages

in order to account for heteroscedasticity as well as covariance among the adjusted means (Damesa et al., 2017; Möhring & Piepho, 2009).

Several papers have addressed the comparison of single-stage and two-stage analyses and even compared different weighting methods (Möhring & Piepho, 2009; Schulz-Streeck, Ogutu, & Piepho, 2013; Welham et al., 2010). Damesa et al. (2017) reported that single-stage and fully efficient (FE) two-stage analyses, where the full variance–covariance matrix of adjusted means in Stage 1 is passed to Stage 2, demonstrated similar results but were not equivalent. They are mathematically equivalent only if identical variance parameter values are used, which is not the case in the practice because the residual maximum likelihood estimates will differ slightly between the two analyses (Damesa et al., 2017; Piepho et al., 2012). By contrast, Gogel, Smith, and Cullis (2018) advocated a move away from two-stage analysis, since the computing power needed to analyze large and complex MET datasets is already available. Their study of wheat MET data confirmed the equivalence of a two-stage factor-analytic (FA) analysis with a known variance–covariance matrix from Stage 1 to a single-stage analysis. An essential distinction between the studies of Damesa et al. (2017) and Gogel et al. (2018), however, is that Damesa et al. (2017) focus on predicting means across zones and across a whole target population of environments (TPE), whereas the study of Gogel et al. (2018) focused on predictions for individual locations. The TPE defines the future growing conditions of the tested cultivars (Comstock, 1977; Cooper & Hammer, 1996; Cooper et al., 2014). Thus a TPE can be delineated on the basis of geography or agro-ecological factors such as soil and meteorological conditions (van Eeuwijk, Bustos-Korts, & Malosetti, 2016).

Currently, the MET data of Swedish official cultivar testing are analyzed via an unweighted two-stage analysis. In Stage 1, the model includes fixed effects for cultivars (coded as F) and random effects for replicates and incomplete blocks. In Stage 2, the analysis is done per zone via a model with fixed effects for cultivars and random effects for locations. The major drawback of the current unweighted approach is that the model is oversimplified in Stage 2. Specifically, the model does not account for either the heteroscedasticity or heterogeneity of covariances of the adjusted means. Furthermore, in Stage 2, the model

does not exploit any covariance per zone, since the analysis is done per zone.

Möhring and Piepho (2009) showed, via simulation, that weighting can improve efficiency but the unweighted method was acceptable if the assumptions of the model were correct (i.e., when the error variances are independent of the genotype \times environment interaction structure). They also mentioned that the relative merit of different methods for weighting did not depend on the evaluation criterion, but on the dataset. Welham et al. (2010) conducted a simulation study and showed that the two-stage unweighted method performed poorly because of a loss of information in estimating estimates of cultivar performance, both overall and within environments. However, similar to Gogel et al. (2018), Welham et al. (2010) focused on predictions for individual sites, whereas the focus in the present study is on means across a wider region or zones or, in other words, zone-based prediction.

The present study investigates the use of several different variance–covariance structures, specifically the factor-analytic (FA), the unstructured (UN), and the compound symmetry (CS) structures for the cultivar \times zone ($C \times Z$) interaction effects in order to better account for potential heterogeneity. In order to determine the best analysis approach (i.e., single-stage vs. two-stage, unweighted or weighted, and the choice of weighting method), combined with variance–covariance structures on the $C \times Z$ interaction effects, an empirical evaluation was needed. This article therefore reports a cross-validation study for evaluation of these different strategies.

2 | MATERIALS AND METHODS

2.1 | Swedish cultivar trial data

The datasets were obtained from official Swedish cultivar tests. Dry matter yield was analyzed. All trials were laid out as α -designs with two replicates. Within each replicate, there were five to seven incomplete blocks. Sweden is divided into three different agricultural zones: South, Middle, and North (Buntaran, Piepho, Hagman, & Forkman, 2019). A zone is represented by a number of locations. We selected five single-year datasets for both crops to be able to perform a leave-one-out cross-validation with a sufficient amount of data. The number of trials and the cultivars of winter wheat and spring barley are reported in Figure 1. The sets of cultivars differed among years. However, within years, the set of cultivars was the same among zones. Each year was analyzed separately.

Core Ideas

- Cross-validation showed that the two-stage weighting strategy performed similarly to the single-stage analysis with location-specific residual variances.
- In comparison to coefficients of correlations, the MSEP provides a clearer distinction between the EBLUP methods and the EBLUE method and a clearer discrimination between the single-stage and the two-stage approach.
- The choice between a single-stage or a two-stage strategy depends on the computational resources due to the loss of information caused by diagonal approximate weighting is negligible.
- The effects of cultivar and the cultivar \times zone is better to be modelled interaction as random to improve the accuracy of zone-based prediction through borrowing information across zones.
- Predictions for zones are more useful and informative for farmers and breeders than predictions for individual locations, since zones cover broader TPEs.

2.2 | Models for single-stage and stagewise analyses

Figure 2 depicts a scheme of a single-stage and a stage-wise analysis. In the single-stage analysis, a zone-based cultivar yield prediction is obtained in a single analysis. In our stagewise approach, the analysis is done in two stages. Thus from this point forward, the stagewise analyses are referred to as two-stage analyses. In Stage 1, the cultivar means were estimated via empirical best linear unbiased estimation (EBLUE). The term “empirical” here refers to the fact that the variance components must be estimated from the data (Forkman, 2013; Harville, 1991; Haslett & Welsh, 2019). In Stage 2, there were two options, unweighted and weighted. The options of weighting methods will be described in the Weighting Methods section.

In the single-stage analysis, $C \times Z$ effects are predicted via empirical best linear unbiased prediction (EBLUP). In the notation introduced by Patterson (1997), the linear mixed model for single-stage analysis is:

$$Y = Z : C + C \cdot Z + Z \cdot L + Z \cdot C \cdot L + Z \cdot L \cdot R \\ + Z \cdot L \cdot R \cdot B, \quad (1)$$

where Y is the response variable (i.e., the yield), C is the cultivar, Z is the zone, L is the location (i.e., the trial),

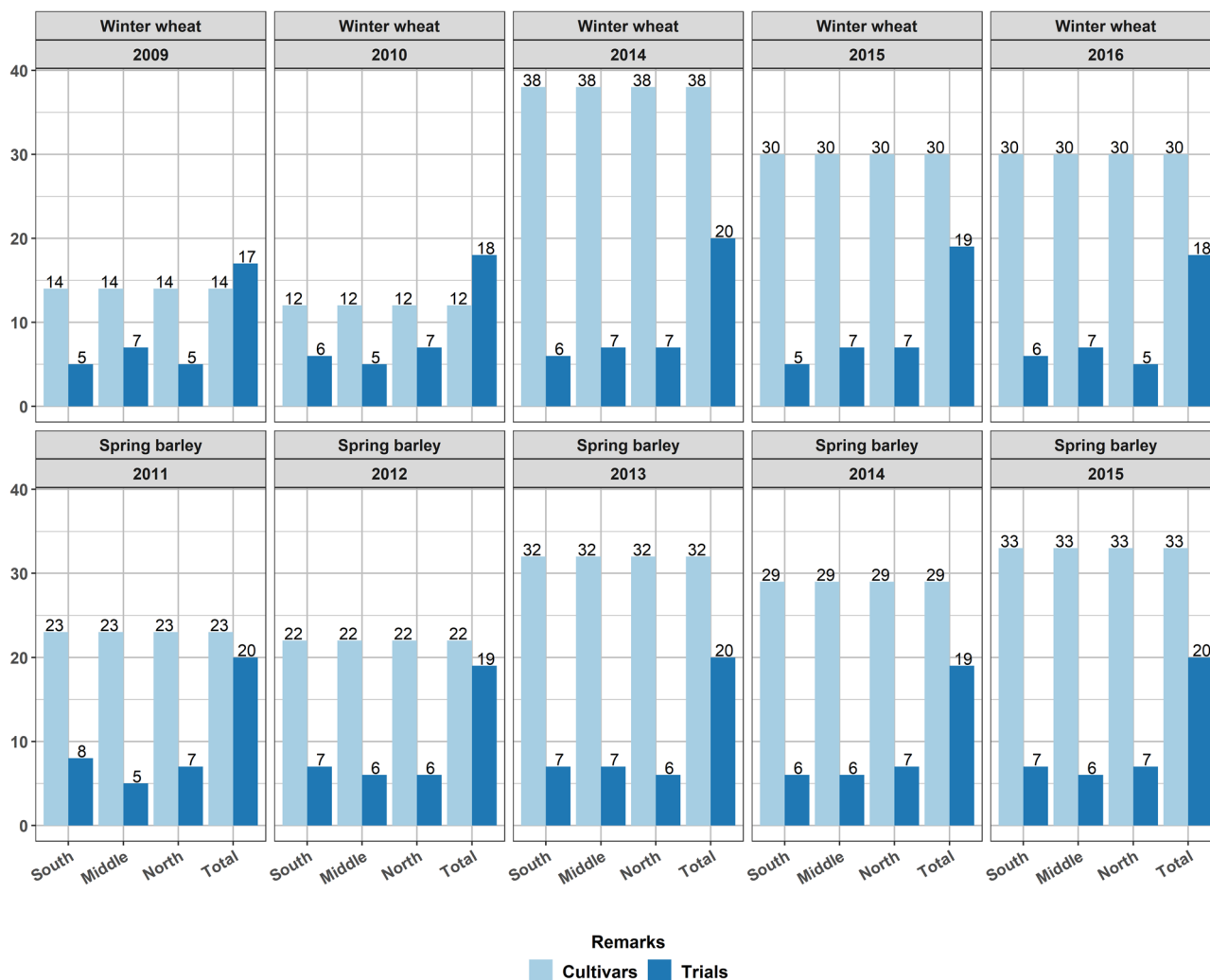


FIGURE 1 Number of winter wheat and spring barley trials by year and agricultural zone in Sweden

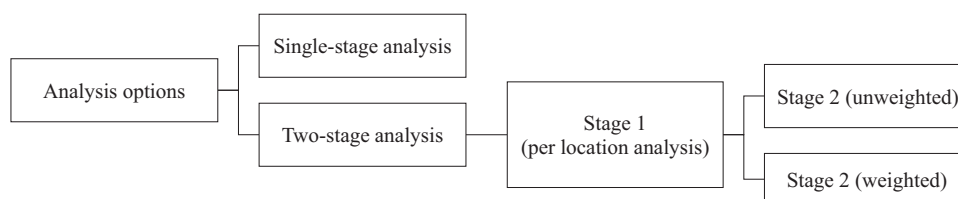


FIGURE 2 Scheme of single-stage and two-stage analyses

R is the replicate within a location, and B is the incomplete block within a replicate. Locations are always nested within zones. The dot operator (•) defines a crossed effect. A crossed effect may refer to an interaction ($C \times Z$ and $Z \times C \times L$) or a nested ($Z \cdot L$, $Z \cdot L \cdot R$, and $Z \cdot L \cdot R \cdot B$) effect. The fixed effects are placed before the colon and the random effects after the colon. With this notation, the grand mean and the error term are implicit.

For the two-stage analysis, in Stage 1, a linear mixed model was used for each location to obtain

the adjusted cultivar means. The cultivar means were estimated via EBLUE through the use of the following model:

$$Y = C : R + R \cdot B. \quad (2)$$

In Stage 2, $C \times Z$ interaction effects can either be modeled as fixed or random. If the $C \times Z$ interaction effect is fixed, then the $C \times Z$ means are estimated via EBLUE. In

TABLE 1 The variance–covariance structures for each term in the models

Term	Variance–covariance structure	Remarks
$\mathbf{Z} \cdot \mathbf{L} \cdot \mathbf{R}^a$	$\mathbf{G}_{\text{ZLR}} = \bigoplus_{j=1}^J \mathbf{G}_{\text{ZLR}(j)}$	Heterogeneous location-specific $\mathbf{G}_{\text{ZLR}(j)}$ is a $K_j \times K_j$ diagonal matrix with diagonal elements $\sigma_{\text{ZLR}(j)}^2$, where K_j is the number of replicates in the j -th location
	$\mathbf{G}_{\text{ZLR}} = \mathbf{I} \sigma_{\text{ZLR}}^2$	Identity
$\mathbf{Z} \cdot \mathbf{L} \cdot \mathbf{R} \cdot \mathbf{B}^b$	$\mathbf{G}_{\text{ZLRB}} = \bigoplus_{j=1}^J \mathbf{G}_{\text{ZLRB}(j)}$	Heterogeneous location-specific $\mathbf{G}_{\text{ZLRB}(j)}$ is an $M_j \times M_j$ diagonal matrix with elements $\sigma_{\text{ZLRB}(j)}^2$, where M_j is the number of blocks in the j -th location.
	$\mathbf{G}_{\text{ZLRB}} = \mathbf{I} \sigma_{\text{ZLRB}}^2$	Identity
$\mathbf{Z} \cdot \mathbf{C} \cdot \mathbf{L}^c$ in Models 1, 3, and 4	$\mathbf{G}_{\text{ZCL}} = \mathbf{I} \sigma_{\text{ZCL}}^2$	Identity
$\mathbf{C}(\mathbf{Z}) = \mathbf{C} + \mathbf{C} \times \mathbf{Z}^d$ in Models 1, 3, and 4	$\mathbf{G}_{\text{C}(\mathbf{Z})} = \bigoplus_{i=1}^I \mathbf{G}_{\text{C}(\mathbf{Z})(i)},$ $\mathbf{G}_{\text{C}(\mathbf{Z})(i)} = \mathbf{I} \sigma_{\text{CZ}}^2 + \mathbf{J} \sigma_{\text{C}}^2$	Compound symmetry I is the number of cultivars.
	$\mathbf{G}_{\text{C}(\mathbf{Z})} = \bigoplus_{i=1}^I \mathbf{G}_{\text{C}(\mathbf{Z})(i)},$ $\mathbf{G}_{\text{C}(\mathbf{Z})(i)} = \sigma_{pp'}^2$	Unstructured p and p' indicate zones.
	$\mathbf{G}_{\text{C}(\mathbf{Z})} = \bigoplus_{i=1}^I \mathbf{G}_{\text{C}(\mathbf{Z})(i)},$ $\mathbf{G}_{\text{C}(\mathbf{Z})(i)} = [\mathbf{A} \mathbf{A}^T + \mathbf{\Psi}]$	Factor-analytic order 1
Residual variance	$\mathbf{R} = \sigma^2 \mathbf{I}$	Identity
	$\mathbf{R} = \bigoplus_{p=1}^P \mathbf{R}_p$	Heterogeneous zone-specific
	$\mathbf{R} = \bigoplus_{j=1}^J \mathbf{R}_j$	Heterogeneous location-specific (LR)

^a $\mathbf{Z} \cdot \mathbf{L} \cdot \mathbf{R}$, replicate in each location within zone.

^b $\mathbf{Z} \cdot \mathbf{L} \cdot \mathbf{R} \cdot \mathbf{B}$, incomplete block within replicate in each location within zone.

^c $\mathbf{Z} \cdot \mathbf{C} \cdot \mathbf{L}$, zone \times cultivar \times location interaction.

^d $\mathbf{C} \times \mathbf{Z}$, cultivar \times zone interaction.

this case, the model is:

$$Y_{\text{adj}} = \mathbf{C} + \mathbf{Z} + \mathbf{C} \cdot \mathbf{Z} : \mathbf{Z} \cdot \mathbf{L} + \mathbf{C} \cdot \mathbf{Z} \cdot \mathbf{L}, \quad (3)$$

where Y_{adj} is the adjusted mean for a cultivar from Stage 1 (i.e., the marginal mean estimated via Model 2). Since the $\mathbf{C} \times \mathbf{Z}$ interaction effects are fixed in Model 3, this model needs reparameterization to estimate the $\mathbf{C} \times \mathbf{Z}$ interaction parameters because the model is not full-rank. If the $\mathbf{C} \times \mathbf{Z}$ interaction effects are random, as in Models 1 and 4, then EBLUP is used for predicting these effects. The model is:

$$Y_{\text{adj}} = \mathbf{Z} : \mathbf{C} + \mathbf{Z} \cdot \mathbf{L} + \mathbf{C} \cdot \mathbf{Z} + \mathbf{C} \cdot \mathbf{Z} \cdot \mathbf{L}. \quad (4)$$

For Models 1 and 4, where the $\mathbf{C} \times \mathbf{Z}$ interaction effects are random, there is no need to reparameterize the $\mathbf{C} \times \mathbf{Z}$ effect estimates because the sum of these estimates is zero. It should be noted that the cultivar \times location ($\mathbf{C} \times \mathbf{Z} \times \mathbf{L}$) interaction effect can be estimated only if weights are used in Stage 2. Otherwise, it is dropped from the model and confounded with the residual error. The weighting options are described in the Weighting Methods section.

2.3 | Cultivar effects: EBLUE vs. EBLUP

Empirical BLUP is described as a shrinkage estimator, since EBLUPs are less spread than EBLUEs (Robinson, 1991). Based on this property, the EBLUPs of high-yielding cultivars tend to be lower than the corresponding EBLUEs, whereas the EBLUPs of low-yielding cultivars tend to be higher than the corresponding EBLUEs. In Stage 1, we assigned the cultivar effects as fixed in order to avoid double shrinkage (Damesa et al., 2017; Piepho et al., 2012).

2.4 | Variance–covariance structures

Table 1 gives an overview of the variance–covariance structures for each factor in the models that were introduced in the previous section. The details of these structures are provided the Appendix.

2.5 | Weighting methods

In Stage 2 of the two-stage analysis, the model may be fitted with precision measures carried forward from Stage 1 (weighting). In Stage 2, the variance–covariance matrix, $\mathbf{\Omega}_j$, of the adjusted genotype means from Stage 1 at the

TABLE 2 The 21 strategies based on the combinations of fitting methods and models for cultivar \times zone ($C \times Z$) effects

Fitting Methods	Structure for the $C \times Z$ effects ^a			Fixed $C \times Z$ effects
	CS	FA1 ^b	UN	
Single-stage				
Single-stage ID residual variance	1S-CS-ID	1S-FA1-ID	1S-UN-ID	–
Single-stage location-specific residual variance	1S-CS-LR	1S-FA1-LR	1S-UN-LR	–
Single-stage All ID random effects and residual variance ^c	1S-AID	–	–	–
Two-stage unweighted ^d				
Two-stage unweighted ID residual variance	2S-CS-U-ID	–	2S-UN-U-ID	–
Two-stage unweighted location-specific residual variance	2S-ID-U-LR	–	2S-UN-U-LR	–
Two-stage unweighted zone-specific residual variance	–	–	–	2S-F-U-ZR ^e
Two-stage weighted				
Two-stage fully efficient	2S-CS-W-FE	–	–	–
Two-stage Smith's weighting location-specific residual variance	2S-CS-W-SW	2S-FA1-W-SW	2S-UN-W-SW	2S-F-SW
Two-stage AVVAR weighting location-specific residual variance	2S-CS-W-AVVAR	2S-FA1-W-AVVAR	2S-UN-W-AVVAR	2S-F-AVVAR

^aAID, all ID random effects and residual variance; AVVAR, average variance of a difference; CS, compound symmetry; F, fixed effect of cultivar; FA1, factor-analytic order 1; FE, fully efficient; ID, identity variance; LR, location-specific residual variance; SW, Smith's weighting; U, unweighted; UN, unstructured variance; W, weighted; ZR, zone-specific residual variance; 1S, single-stage; 2S, two-stage.

^bThe FA1 and UN covariance model excluding the cultivar main random effects to avoid overparameterization.

^cAll ID random effects assumes an ID variance–covariance structure for all random main (cultivar), nested (replicate in each location within zone and incomplete block within replicate in each location within zone), and interaction (cultivar \times zone) effects.

^dSince there is no weighting, the cultivar \times location interaction term cannot be fitted in Stage 2.

^eThe strategy in current practice.

j -th location is used as the weight. However, in application, since the matrix \mathbf{Q}_j is not known, it is replaced by its residual maximum likelihood estimate, \mathbf{Q}_j . The FE weighting carries the full \mathbf{Q}_j matrix from Stage 1 forward to Stage 2 (Damesa et al., 2017).

In practice, storing the full \mathbf{Q}_j matrix is often not easy. For example, it is not always practical to provide facilities or resources that are able to store the full \mathbf{Q}_j matrix for many crops, traits, and seasons. As a more convenient alternative, a diagonal matrix approximation of \mathbf{Q}_j may be used. Smith's weighting (SW) uses the diagonal of the inverse of \mathbf{Q}_j , $D(\mathbf{Q}_j^{-1})$ (Smith, Cullis, & Gilmour, 2001). The diagonal elements are used as weights for the corresponding adjusted means.

The average variance of a difference (AVVAR) is another alternative for weighting via a diagonal approximation of \mathbf{Q}_j . This is computed at each location by taking half of an average variance of a difference (Möhring & Piepho, 2009). Let t_j denote the number of cultivars in the j -th location. The AVVAR weight is computed as $0.5 \times \mathbf{I}_{t(j)} \times (VDIFF)_j$, where $(VDIFF)_j$ is the average squared standard error of a difference at the j -th location:

$$(VDIFF)_j = \frac{t_j \times \text{trace}(\mathbf{Q}_j) - \mathbf{1}'\mathbf{Q}_j\mathbf{1}}{t_j(t_j - 1)/2} \quad (5)$$

2.6 | Strategies

Table 2 summarizes the strategies; in other words, the combination of fitting method (single-stage and two-stage analysis), residual variance model, and assumptions for the $C \times Z$ effects (fixed or random with some covariance structures). Altogether, 21 strategies were compared in the cross-validation study. These were coded as follows:

- The first two characters defines the fitting method: single-stage or two-stage (1S or 2S).
- The second set of characters defines the structure for the $C \times Z$ effects [CS, FA order 1 (FA1), or UN]. For a particular single-stage model, the code AID on the second set of characters refers to the all-identity structure for all random main (cultivar), nested (zone • location • replicate and zone • location • replicate • block), and interaction ($C \times Z$) effects and residual variance. For the fixed $C \times Z$ effect structures, the letter F indicates that the effects of cultivar, zone, and $C \times Z$ are fixed.
- For single-stage models (1S), the third set of characters defines the residual structure: identity (ID) or location-specific residual variances (LR). The LR residual structure refers to plot error estimates of individual location. The ID residual structure represents a naive approach that assumes that all locations have identical plot error

variance. The zone-specific residual variance (ZR) structure was not applied in the single-stage models because this structure does not assess the plot error of each trial. For two-stage models (2S), the third set of characters defines whether the locations were unweighted (U) or weighted (W).

- For two-stage models, the fourth set of characters defines the weighting method (FE, SW, or AVVAR). For two-stage models without weighting, the fourth set of characters defines the residual structure (ID, ZR, or LR).

2.7 | Cross-validation

Many papers use Pearson's product-moment correlation or Spearman's rank correlation for measuring the strength of the relationship between the cultivar estimates of single-stage and two-stage analyses (Damesa et al., 2017; Gogel et al., 2018; Möhring & Piepho, 2009; Piepho et al., 2012). In those papers, the correlation coefficient estimates were often above 0.90, showing that the single-stage and stagewise analyses provide correlated results. In contrast to Pearson's correlation, a cross-validation study can measure the prediction errors of the model via the mean squared error of prediction (MSEP) of difference, which is more desirable for choosing the model that best predicts cultivar performance in MET analysis. The MSEP measures predictive accuracy and is considered to be more informative than the correlation coefficient (Gauch, Hwang, & Fick, 2003).

We conducted a leave-one-out cross-validation for model comparison and selection. We left one location out at a time. That location was used as a validation set; and the remaining locations were used as a training set. For example, if there were 18 locations in a single-year dataset, as was the case for winter wheat in 2016, then there were 18 folds of the cross-validation. We computed an MSEP similar to the MSEP proposed by Piepho (1998) for measuring the prediction accuracy of the models for each single-year dataset. We accumulated the discrepancies between the observed and predicted pairwise differences from the 18 folds of the cross-validation. We then computed the MSEP from this accumulation. Hence, there will be just one value of MSEP from the 18 folds of the cross-validation. The MSEP is a standard statistic for assessing predictive accuracy. Let y and z denote the observed and predicted values, respectively; let I be the total number of cultivars; and let J be the total number of locations. The assessment was based on the discrepancies between the observed ($y_{ij} - y_{i'j}$) and predicted ($z_{ij} - z_{i'j}$) pairwise

differences:

$$MSEP = \frac{\sum_{j=1}^J \sum_{i=1}^I \sum_{i' \neq i}^I [y_{ij} - y_{i'j} - (z_{ij} - z_{i'j})]^2}{JI(I-1)}, \quad (6)$$

where y_{ij} is the observed yield of cultivar i in location j ; $y_{i'j}$ is the observed yield of cultivar i' in location j , where $i \neq i'$; z_{ij} is the predicted yield of cultivar i in location j and $z_{i'j}$ is for the predicted yield of cultivar i' in location j , where $i \neq i'$. The rationale of using pairwise differences is that the main interest in cultivar trials is to predict differences among cultivars rather than individual cultivars' performance (Piepho, 1998).

The best model was the one that produced the smallest MSEP, since it predicted yield differences in the validation set most accurately. We would like to have the most accurate predictions for each agricultural zone as a prediction for the locations within zones. The approaches were ranked on the basis of the average MSEP over the five cross-validation sets, since there were five single-year datasets. We conducted the cross-validation study in R (R Core Team, 2018) and fitted all the models in ASReml-R version 4.1.0.106 (Butler, Cullis, Gilmour, Gogel, & Thompson, 2017) in RStudio (RStudio Team, 2016). The ggplot2 package (Wickham, 2009) was used in R to produce the plots.

3 | RESULTS

3.1 | Cross-validation of the single-stage and two-stage approaches

The MSEP averages are presented in Table 3. For both crops, the single-stage approach with the CS variance-covariance structure of random effects and location-specific residual structure (1S-CS-LR) performed best, since this strategy had the lowest average MSEP. However, the differences between 1S-CS-LR and the three weighted two-stage strategies (2S-CS-W-FE, 2S-CS-W-AVVAR, and 2S-CS-W-SW) were minor for both crops. Thus these four strategies performed very similarly. Furthermore, in Supplemental Figure S1, Supplemental Figure S2, Supplemental Figure S3, and Supplemental Figure S4 depict the scatterplots of the observed differences vs. the predicted differences of the cross-validation results for 1S-CS-LR, 2S-CS-W-FE, 2S-CS-W-AVVAR, and 2S-CS-W-SW. They show that the patterns of these four strategies are very similar. The relevance of a difference may be judged on

TABLE 3 Mean squared error of prediction differences of winter wheat ($n^a = 5$) and spring barley ($n = 5$)

Ranking	Winter wheat		Spring barley	
	Strategy ^b	Mean	Strategy	Mean
		$\text{g}^2 \text{m}^{-4}$		$\text{g}^2 \text{m}^{-4}$
1	<u>1S-CS-LR^c</u>	5,041	<u>1S-CS-LR</u>	1,723
2	<u>2S-CS-W-FE</u>	5,045	<u>2S-CS-W-FE</u>	1,726
3	<u>2S-CS-W-AVVAR</u>	5,049	<u>2S-CS-W-S W</u>	1,727
4	<u>2S-CS-W-S W</u>	5,051	<u>2S-CS-W-AVVAR</u>	1,728
5	1S-UN-LR	5,057	1S-UN-LR	1,728
6	2S-CS-U-ID	5,066	2S-UN-U-LR	1,731
7	2S-UN-W-AVVAR	5,066	2S-CS-U-ID	1,736
8	2S-UN-W-SW	5,072	1S-CS-ID	1,736
9	1S-CS-ID	5,080	1S-UN-ID	1,739
10	2S-FA1-W-AVVAR	5,084	2S-UN-W-SW	1,739
11	1S-UN-ID	5,088	2S-FA1-W-SW	1,740
12	1S-FA1-LR	5,090	2S-UN-U-ID	1,741
13	2S-FA1-W-SW	5,091	2S-UN-W-AVVAR	1,741
14	2S-UN-U-ID	5,091	2S-FA1-W-AVVAR	1,742
15	1S-AID	5,102	2S-CS-U-LR	1,743
16	1S-FA1-ID	5,107	1S-AID	1,758
17	2S-CS-U-LR	5,123	1S-FA1-LR	1,804
18	2S-UN-U-LR	5,210	2S-F-SW	1,838
19	2S-F-AVVAR	5,327	2S-F-AVVAR	1,840
20	2S-F-SW	5,334	2S-F-U-ZR ^d	1,850
21	2S-F-U-ZR ^d	5,389	1S-FA1-ID	1,870

^aNumber of datasets.

^bThe four top-performing strategies have been underlined.

^cAVVAR, average variance of a difference; CS, compound symmetry; F, fixed effect of cultivar; FA1, factor-analytic order 1; FE, fully efficient; ID, identity variance; LR, location-specific residual variance; SW, Smith's weighting; U, unweighted; UN, unstructured variance; W, weighted; ZR, zone-specific residual variance; 1S, single-stage; 2S, two-stage.

^dThe current practice of analysis in Swedish cultivar testing.

the basis of:

$$\frac{\sqrt{(MSEP_1 - MSEP_2)}}{\sqrt{\frac{\sum_{v=1}^5 (\sigma_{C_v}^2 + \sigma_{CZ_v}^2)}{5}}} \times 100\%, \quad (7)$$

where $MSEP_1$ is the MSEP for Strategy 1, $MSEP_2$ is the MSEP for Strategy 2, p refers to the cross-validation set for each year's dataset, $\sigma_{C_v}^2$ is the variance components of the cultivar of the v -th cross-validation set, and $\sigma_{CZ_v}^2$ is the variance components of the $C \times Z$ interactions of the p -th cross-validation set. The denominator is the SD of the genotypic

values, cultivar + $C \times Z$, computed by taking the square root of the average of the variance component estimates of cultivar and $C \times Z$ effects from the five datasets. In winter wheat, the difference in MSEP between the two top-performing strategies (1S-CS-LR and 2S-CS-W-FE) is 4 and the square root of this difference, which is 2, corresponds to 5.33% of the SD of genotype values, which is very small. The value of 5.33% is calculated as 2 divided by the SD of genotypic values, then multiplying the result by 100%.

On the other hand, the current practice (2S-F-U-ZR) was the lowest performing strategy for winter wheat and the second lowest for spring barley. Supplemental Figure S5 and Supplemental Figure S10 depict the scatterplots of the observed differences vs. the predicted differences for 2S-F-U-ZR and shows that the dots are more spread out than in the four best strategies, which means the 2S-F-U-ZR strategy provides the lowest prediction accuracy. When the 1S-CS-LR strategy is compared with the current practice (2S-F-U-ZR), the difference in MSEP corresponds to 49.69%. Hence, the improvement of single-stage analysis with BLUP over two-stage unweighted BLUE was considerable.

Furthermore, none of the weighting methods greatly improved the fixed $C \times Z$ effect strategies (2S-F-AVVAR and 2S-F-SW) much compared with the current approach (2S-F-U-ZR). Complex variance-covariance structures for $C \times Z$ did not improve the predictive model performance. The FA1 structure performed much worse than the ID structure for single-stage strategies as well as for two-stage strategies. No model with the FA structure was among the five top-performing strategies. In spring barley, the 1S-FA1-ID strategy performed the worst, even worse than the current practice. In general, the MSEP of the UN structure was similar to the MSEP of the FA1 structure. The exception was the UN structure in the single-stage strategy with heterogeneous residual location-specific variance (1S-UN-LR), which, for both crops, showed the fifth-best average MSEP.

In both crops, the simple unweighted two-stage strategy, 2S-CS-U-ID, performed better than 1S-AID. This outcome showed that the simple EBLUP two-stage unweighted strategy produced better predictions than the simple single-stage EBLUP strategy. Thus, the use of adjusted means from Stage 1 was more accurate than a single-stage approach that neglects the heterogeneity of variance in replicates and incomplete blocks across locations.

3.2 | Application to winter wheat and spring barley datasets

As examples, in this section, we consider the application of the four top-performing strategies (1S-CS-LR, 2S-CS-W-FE,

TABLE 4 Variance component estimates of the winter wheat 2016 and spring barley 2015 datasets

Variance parameter	Winter wheat 2016				Spring barley 2015			
	Strategy ^a				Strategy ^a			
	1S-CS-LR	2S-CS-W-FE	2S-CS-W-AVVAR	2S-CS-W-SW	1S-CS-LR	2S-CS-W-FE	2S-CS-W-AVVAR	2S-CS-W-SW
σ_C^2	557.2	614.7	596.0	601.7	1,153.3	1,118.3	1,123.2	1,121.3
σ_{ZL}^2	4,8874.4	4,8874.5	4,8937.4	4,9024.8	4,967.5	7,628.6	8,056.9	8,056.3
σ_{CZ}^2	129.7	125.0	129.4	130.3	203.9	217.2	216.1	216.2
σ_{ZCL}^2	794.4	1,479.3	1,399.4	1,472.2	301.2	416.4	422.1	428.1

^a1S-CS-LR, single-stage analysis; 2S-CS-W-FE, two-stage fully efficient analysis; 2S-CS-W-AVVAR, two-stage analysis with average variance of a difference (AVVAR) weights (Möhrling & Piepho, 2009); 2S-CS-W-SW, two-stage analysis with Smith's diagonal weights (Smith et al., 2001); CS, compound symmetry; FE, fully efficient; LR, location-specific residual variance; SW, Smith's weighting; W, weighted; 1S, single-stage; 2S, two-stage; σ_C^2 , variance component estimate of the cultivar; σ_{CZ}^2 , variance component estimate of cultivar \times zone; σ_{ZCL}^2 , variance component estimate of cultivar \times location; σ_{ZL}^2 , variance component estimate of location.

TABLE 5 Estimates of cultivar variance (on the diagonal), correlation (above the diagonal), and covariance (below the diagonal) with CS structure on C \times Z effect in the winter wheat 2016 dataset

Strategy ^a												
Zone	1S-CS-LR			2S-CS-W-FE			2S-CS-W-AVVAR			2S-CS-W-SW		
	EBLUP			EBLUP			EBLUP			EBLUP		
	1	2	3	1	2	3	1	2	3	1	2	3
1	686.88	0.81	0.81	739.72	0.83	0.83	725.44	0.82	0.82	732	0.82	0.82
2	557.18	686.88	0.81	614.71	739.72	0.83	596.02	725.44	0.82	601.7	732	0.82
3	557.18	557.18	686.88	614.71	614.71	739.72	596.02	596.02	725.44	601.7	601.7	732

^a1S-CS-LR, single-stage analysis; 2S-CS-W-FE, two-stage fully efficient analysis; 2S-CS-W-AVVAR, two-stage analysis with average variance of a difference (AVVAR) weights (Möhrling & Piepho, 2009); 2S-CS-W-SW, two-stage analysis with Smith's diagonal weights (Smith et al., 2001); CS, compound symmetry; FE, fully efficient; LR, location-specific residual variance; SW, Smith's weighting; W, weighted; 1S, single-stage; 2S, two-stage; EBLUP, empirical best linear unbiased prediction.

2S-CS-W-AVVAR, and 2S-CS-W-SW) to the winter wheat 2016 and spring barley 2015 datasets. The variance component estimates for effects of cultivar (σ_C^2), location (σ_{ZL}^2), C \times Z interactions (σ_{CZ}^2), and cultivar \times location interactions (σ_{ZCL}^2) are presented in Table 4. Lists of the variance component estimates for both datasets are available in Supplemental Table S1 and Supplemental Table S2.

For winter wheat, there were only small differences among the four approaches in the estimates of cultivar, location, and C \times Z variances (Table 4). The estimate of σ_{ZCL}^2 by the 1S-CS-LR strategy was approximately 55% of the variance estimates for the other three two-stage strategies (2S-CS-W-FE, 2S-CS-W-AVVAR, and 2S-CS-W-SW). In spring barley, σ_{ZL}^2 by the 1S-CS-LR method was approximately 63% of the estimates for the other three two-stage strategies. The variance component estimates for cultivar, C \times Z, and zone \times location effects were similar among these four approaches.

The cultivars' variances, covariances, and correlations are presented in Table 5 and Table 6 for winter wheat and spring barley, respectively. Since there were only three zones, the CS, UN, and FA1 structures were equal regarding the number of variance-covariance estimates. The cultivar variance estimate (the diagonal part of Table 5 and

Table 6) for these four top-performing strategies was relatively similar for both crops. The same goes for the genetic correlation and covariance in both crops. Therefore, there was no notable difference among these four strategies. It should be noted that to exploit genetic variance and correlation information, the C \times Z effect has to be assigned as a random effect, which is not the case with the current practice in Sweden, where the C \times Z effect is fixed. The genetic correlation was around 0.80, meaning that the yield effect of each cultivar is similar among the three zones, though the ranking of the cultivars is likely to be different among zones since the genetic correlations were not close to one.

Figure 3 depicts the zone-pairwise scatterplots of cultivar predictions (EBLUP) and estimates (EBLUE) for each model and crop. Figure 3a shows pairwise cultivar predictions and estimates of C \times Z effects for the South and North zones. Figure 3b presents pairwise cultivar predictions and estimates of C \times Z effects for the Middle and North zones, and Figure 3c shows pairwise cultivar predictions and estimates of C \times Z effects for the South and Middle zones. In general, it can be seen that the EBLUP methods (1S-CS-LR, 2S-CS-W-FE, 2S-CS-W-AVVAR, 2S-CS-W-SW) have narrower ellipses than the EBLUE method (2S-F-U-ZR). Thus with the EBLUP method, the

TABLE 6 Estimates of cultivar variance (on the diagonal), correlation (above the diagonal), and covariance (below the diagonal) with CS structure on C × Z effect structure in spring barley 2015 dataset

Strategy ^a												
Zone	1S-CS-LR			2S-CS-W-FE			2S-CS-W-AVVAR			2S-CS-W-SW		
	EBLUP			EBLUP			EBLUP			EBLUP		
	1	2	3	1	2	3	1	2	3	1	2	3
1	1,357.20	0.85	0.85	1,335.53	0.84	0.84	1,339.40	0.84	0.84	1,337.50	0.84	0.84
2	1,153.30	1,357.20	0.85	1,121.85	1,335.53	0.84	1,123.20	1,339.40	0.84	1,121.30	1,337.50	0.84
3	1,153.30	1,153.30	1,357.20	1,121.85	1,121.85	1,335.53	1,123.20	1,123.20	1,339.40	1,121.30	1,121.30	1,337.50

^a1S-CS-LR, single-stage analysis; 2S-CS-W-FE, two-stage fully efficient analysis; 2S-CS-W-AVVAR, two-stage analysis with average variance of a difference (AVVAR) weights (Möhring & Piepho, 2009); 2S-CS-W-SW, two-stage analysis with Smith’s diagonal weights (Smith et al., 2001); CS, compound symmetry; FE, fully efficient; LR, location-specific residual variance; SW, Smith’s weighting; W, weighted; 1S, single-stage; 2S, two-stage; EBLUP, empirical best linear unbiased prediction.

cultivar rankings were more similar between two zones than with the EBLUE method, as expected.

Table 7 reports the Akaike information criterion and −2 residual log-likelihood of the four top-performing strategies in both crops. The FE, AVVAR, and SW strategies did not differ much for both crops in likelihood and Akaike information criterion. All these four strategies converged without demanding much computational time, as summarized in Table 8. With the current software and computational resources, the single-stage approach only took a couple of seconds (macOS X 10.15.1, Apple Inc., Cupertino, CA; 64-bit operating system, 16 GB RAM). However, the two-stage FE analysis (2S-CS-W-FE) was the most demanding computationally compared with the other three strategies because it needed more time and memory allocation for forwarding the full variance–covariance matrix of the adjusted means from Stage 1 to Stage 2. Note that the variance–covariance matrix of the winter wheat 2016 dataset consisted of 540 columns and 540 rows, resulting in 291,600 entries.

The top 10 rankings and the predictions and estimates of dry matter yield of winter wheat and spring barley in the South zone are presented in Tables 9 and 10, respectively. These tables compare the four top-performing strategies with the current strategy. The full lists of the adjusted cultivar estimates per zone, for both crops, are available in Supplemental Table S3 and Supplemental Table S4.

In winter wheat, the top-performing cultivar was the same, cultivar G 0512LT3, for all approaches. However, moderate shrinkage from 960 to ~928 g m^{−2} was observed with the four strategies that applied EBLUP (1S-CS-LR, 2S-CS-W-FE, 2S-CS-W-AVVAR, and 2S-CS-W-SW) compared with EBLUE (2S-CS-U-ZR), which is the strategy that is currently used in Sweden. Furthermore, the rankings of the EBLUE methods were considerably dissimilar to those of the EBLUP methods. The most obvious difference was the cultivar RGT Reform. This cultivar ranked third in the single-stage analysis with EBLUP methods and fourth by

the two-stage weighting analyses with EBLUP methods, while by the EBLUE method, it ranked 10. On the other hand, cultivar Brons ranked third by the EBLUE method but was ranked lower by the two-stage weighting strategies with EBLUP methods. Moreover, this cultivar was not in the 10 top-performing cultivars according to the single-stage analysis with EBLUP. The cultivar ranking was the same for the two different weighting methods. Nonetheless, the ranking of these two-stage analyses was slightly different from that of the single-stage analysis. Note that according to the single-stage analysis, the cultivar Hereford was among the top 10, but not according to the two-stage analyses.

In spring barley, the top performing cultivar was also the same for the four strategies: cultivar Dragoon. The shrinkage by EBLUP (1S-CS-LR, 2S-CS-W-AVVAR, and 2S-CS-W-SW) was smaller than in winter wheat. Thus the non-genetic variation in spring barley was smaller than that in winter wheat. The rankings of cultivars among the four strategies with EBLUP were very similar, which agrees with Figure 2. However, these EBLUP rankings differed from the EBLUE rankings (2S-F-U-ZR).

Table 11 and Table 12 present, for winter wheat and spring barley, respectively, Pearson’s product–moment correlations and Spearman’s rank correlations for all adjusted cultivar predictions and estimates, using the four top-performing strategies (1S-CS-LR, 2S-CS-W-FE, 2S-CS-W-AVVAR, 2S-CS-W-SW) and the current strategy (2S-F-U-ZR.) Here, the cultivar ranking of the four top-performing strategies can be compared with the cultivar ranking of the current strategy. For winter wheat, both Pearson’s and Spearman’s correlations were high among the four strategies with EBLUP but relatively low between these four strategies and the 2S-F-U-ZR strategy. The correlations between the two-stage analyses were close to one. For spring barley, the correlations among these four strategies were higher than in winter wheat, even for the correlations between the strategies with EBLUP and the 2S-F-U-ZR strategy.

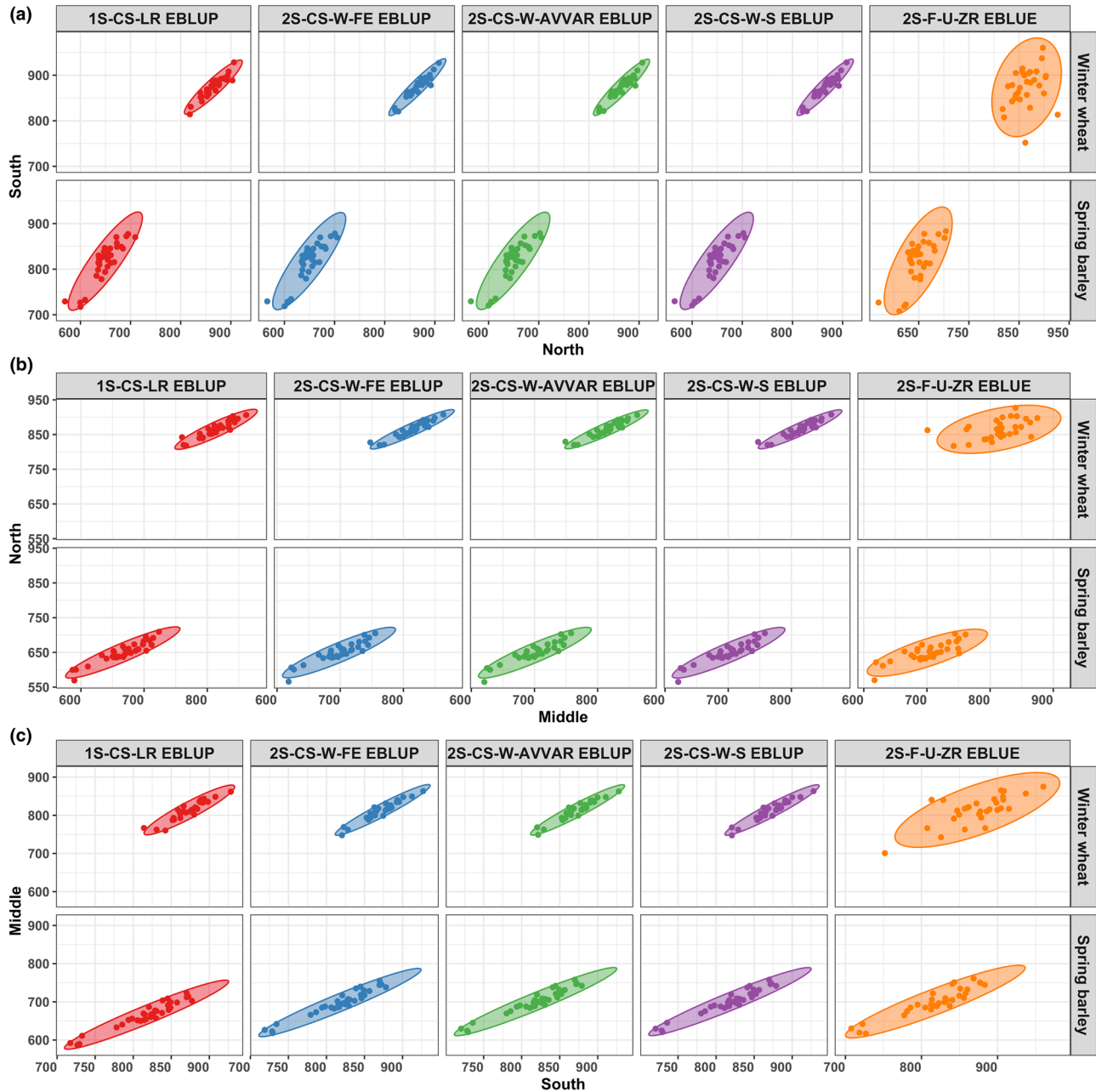


FIGURE 3 Zone-pairwise scatterplots of cultivar estimates of cultivar \times zone interaction effects for the four top-performing strategies and the strategy in current practice [two-stage unweighted zone-specific residual variance (2S-F-U-ZR EBLUE)]. Cultivar predictions and estimates for (a) North and South, (b) Middle and North, and (c) South and Middle

4 | DISCUSSION

The cross-validation study revealed that the two-stage strategies with diagonal weighting, either with SW (2S-CS-W-SW) or with AVVAR weights (2S-CS-W-AVVAR), were very close in performance to the two-stage FE strategies (2S-CS-W-FE) and the single-stage strategy with heterogeneous location-specific residual variance (1S-CS-LR). In addition, these four strategies outperformed the current strategy (2S-F-U-ZR). Hence, the current-

practice strategy should be discontinued for routine analysis. The differences in MSE among 1S-CS-LR, 2S-CS-W-FE, 2S-CS-W-AVVAR, and 2S-CS-W-SW were slight, which confirmed that the loss of information resulting from a two-stage analysis with diagonal weights instead of a single-stage analysis is acceptable (Möhring & Piepho, 2009).

The reason why the single-stage analysis and the two-stage with weighting performed similarly was because of the weighting that was used in the two-stage analysis.

TABLE 7 Akaike information criterion (AIC) and -2 residual log-likelihood ($-2LL$) values for the four top-performing strategies in winter wheat and spring barley. Note that $-2LL$ and AIC can be compared only within two-stage (2S) analyses, but not between single-stage (1S) and 2S analyses

Strategy ^a	Winter wheat		Spring barley	
	$-2LL$	AIC	$-2LL$	AIC
1S-CS-LR	9,586	9,702	10,918	11,045
2S-CS-W-FE	4,918	4,926	5,328	5,336
2S-CS-W-AVVAR	4,910	4,917	5,332	5,341
2S-CS-W-SW	4,916	4,924	5,330	5,338

^a1S-CS-LR, single-stage analysis; 2S-CS-W-FE, two-stage fully efficient analysis; 2S-CS-W-AVVAR, two-stage analysis with average variance of a difference (AVVAR) weights (Möhrling & Piepho, 2009); 2S-CS-W-SW, two-stage analysis with Smith's diagonal weights (Smith et al., 2001); CS, compound symmetry; FE, fully efficient; LR, location-specific residual variance; SW, Smith's weighting; W, weighted.

Again, the performance of the two-stage FE method was very close to the single-stage analysis, since all information from Stage 1 was carried forward to Stage 2. For the diagonal approximate weighting, the performance was similar to that of the single-stage analysis, since the weighting was based on diagonal part of the inverse of the variance–covariance matrix. This inverse of the variance–covariance matrix had small, and hence negligible, off-diagonal elements, whereas the diagonal elements used for weighting in Stage 2 were large by comparison. Thus the use of two-stage weighting is reasonable.

Henderson (1978) compared the single-stage method, the so-called “mixed model”, with the regressed least squares (RLS), and modified regressed least squares methods for predicting breeding values. The latter two methods can be considered as two-stage analyses. Although, in this study, the results of the single-stage and two-stage analysis were very similar, in his study, the results of the “mixed model” and regressed least squares were not. The notable distinctions between our study and Henderson's study are (a) that regressed least squares does not consider weighting, and (b) that the goal in Henderson's study was to predict breeding values, which required pedigree data; however, our analysis does not use pedigree data.

In comparison with the MSEPs in Table 3, when Pearson's and Spearman's correlation coefficients were used exclusively, as presented in Table 10 and Table 11, it is difficult to detect that EBLUP performed better than EBLUE, especially in spring barley. Besides, it is also difficult to see any difference in performance between the single-stage and two-stage approaches. The MSEP provides a clearer distinction between the EBLUP methods and the EBLUE method and a clearer discrimination between the single-stage and the two-stage approach. According to Kobayashi and Salam (2000), correlation is not the best measure for

TABLE 8 Computing time of each strategy in winter wheat and spring barley

Strategy ^a	Computing time	
	Winter wheat	Spring barley
1S-CS-LR	2.98	3.05
2S-CS-W-FE	59.16	97.21
2S-CS-W-AVVAR	8.64	9.15
2S-CS-W-SW	8.80	8.96

^a1S-CS-LR, single-stage analysis; 2S-CS-W-FE, two-stage fully efficient analysis; 2S-CS-W-AVVAR, two-stage analysis with average variance of a difference (AVVAR) weights (Möhrling & Piepho, 2009); 2S-CS-W-SW, two-stage analysis with Smith's diagonal weights (Smith et al., 2001); CS, compound symmetry; FE, fully efficient; LR, location-specific residual variance; SW, Smith's weighting; W, weighted; 1S, single-stage; 2S, two-stage.

model evaluation, since the mean squared deviation is easier to interpret and more useful for direct comparisons between model output and measurement. Thus the MSEP from this cross-validation study was used as additional evidence.

The cross-validation was done by excluding one location at a time for each year dataset rather than excluding a complete year dataset. The reason for not excluding a complete 1-yr dataset and taking the other 4 yr as a validation set was that the training set and the validation set must be in a consecutive scheme because the goal of MET is to predict future cultivar performance. Moreover, cultivars drop out of the system each year, meaning that in the cross-validation, the set of cultivars is thinned out over time. Thus it is not possible to randomly assign any year to be the validation set. In Supplemental Figure S1, Supplemental Figure S2, Supplemental Figure S3, Supplemental Figure S4, and Supplemental Figure S5, which depict the scatterplots of the observed differences vs. the predicted differences of the cross-validation results for 1S-CS-LR, 2S-CS-W-FE, 2S-CS-W-AVVAR, 2S-CS-W-SW, and 2S-F-U-ZR for winter wheat datasets, show a similar pattern among different datasets. The same plots for the spring barley datasets are depicted in Supplemental Figure S6, Supplemental Figure S7, Supplemental Figure S8, Supplemental Figure S9, and Supplemental Figure S10 and shows similar patterns. Therefore, conducting per-year cross-validations will provide similar results to the cross-validation that excludes a complete year.

The MSEP obtained in our cross-validation study clearly showed that the four strategies with random cultivar effects were more accurate than the strategy in current practice, which uses fixed cultivar effects. From a statistical perspective, fitting the effects of cultivars as random is better than fitting them as fixed because the rankings of the estimated cultivars are expected to be close to the rankings of the cultivar effects and hence provide more

TABLE 9 Top 10 ranking and the predictions and estimates of dry matter yield (DMY) of winter wheat cultivars in the South zone

Cultivar	Strategy ^a									
	1S-CS-LR		2S-CS-W-FE		2S-CS-W-AVVAR		2S-CS-W-SW		2S-F-U-ZR ^b	
	EBLUP		EBLUP		EBLUP		EBLUP		EBLUE	
	Ranking	DMY	Ranking	DMY	Ranking	DMY	Ranking	DMY	Ranking	DMY
		g m ⁻²		g m ⁻²		g m ⁻²		g m ⁻²		g m ⁻²
Brons	–	–	–	–	10	883.7	10	884.1	3	915.0
Creator (SJ 8544003)	9	888.4	8	892.3	8	890.3	8	890.4	5	907.5
Effekt (SW 85131)	10	887.1	7	893.2	7	890.8	7	890.9	7	904.7
Ellen (SW 75638)	4	893.0	3	900.9	3	899.6	3	900.1	4	908.2
Etana	2	908.0	2	912.4	2	910.5	2	910.8	2	937.1
Festival (SW 95594)	–	–	9	885.7	9	885.2	9	885.5	8	904.4
G 0512LT3	1	928.3	1	927.1	1	927.6	1	928.1	1	960.1
Hereford	6	891.5	10	884.3	–	–	–	–	–	–
Mariboss	5	891.6	5	894.7	5	893.8	5	893.3	6	907.0
RGT Reform	3	898.4	4	895.4	4	895.2	4	895.0	10	898.4
Rivero (Nord 07098/125)	–	–	–	–	–	–	–	–	9	899.7
Rockefeller (SJ8584007)	7	889.5	6	893.5	6	892.6	6	892.3	–	–
W 237	8	888.7	–	–	–	–	–	–	–	–

^a1S-CS-LR, single-stage analysis; 2S-CS-W-FE, two-stage fully efficient analysis; 2S-CS-W-AVVAR, two-stage analysis with average variance of a difference (AVVAR) weights (Möhrling & Piepho, 2009); 2S-CS-W-SW, two-stage analysis with Smith's diagonal weights (Smith et al., 2001); CS, compound symmetry; F, fixed effect of cultivar; FE, fully efficient; LR, location-specific residual variance; SW, Smith's weighting; U, unweighted; W, weighted; ZR, zone-specific residual variance; 1S, single-stage; 2S, two-stage; EBLUP, empirical best linear unbiased prediction; EBLUE, empirical best linear estimation.

^bCurrent practice.

TABLE 10 Top 10 ranking and the predictions and estimates of dry matter yield (DMY) of the spring barley cultivars in the South zone

Cultivar	Strategy ^a									
	1S-CS-LR		2S-CS-W-FE		2S-CS-W-AVVAR		2S-CS-W-SW		2S-F-U-ZR ^b	
	EBLUP		EBLUP		EBLUP		EBLUP		EBLUE	
	Ranking	DMY	Ranking	DMY	Ranking	DMY	Ranking	DMY	Ranking	DMY
		g m ⁻²		g m ⁻²		g m ⁻²		g m ⁻²		g m ⁻²
Avenger (SC 42591 M4)	10	845.3	10	844.6	10	844.4	10	844.2	–	–
Deveron (LGB 11–8345)	4	870.2	3	869.8	4	869.7	4	869.6	4	868.6
Dragoon	1	877.2	1	878.1	1	879	1	879	1	883.3
Highway (NOS 19339–81)	9	845.7	–	–	9	844.9	9	845.1	10	843
KWS Irina	7	848.9	6	849.9	6	852.2	6	852.3	6	857.9
NOS 19313–83	–	–	9	845.2	–	–	–	–	9	849.7
Odyssey	8	846.6	8	846	8	846.9	8	847.1	7	855.7
RGT Planet	3	870.4	4	869.7	3	871.6	3	871.5	2	877.6
Sanette (SY 409-226)	6	849.1	7	849.2	7	849.2	7	849.4	8	850.9
Scholar	2	872.4	2	872.2	2	872.8	2	872.8	3	876.9
Thermus (SJ 111703)	5	857.6	5	856.4	5	856.6	5	856.6	5	860.3

^a1S-CS-LR, single-stage analysis; 2S-CS-W-FE, two-stage fully efficient analysis; 2S-CS-W-AVVAR, two-stage analysis with average variance of a difference (AVVAR) weights (Möhrling & Piepho, 2009); 2S-CS-W-SW, two-stage analysis with Smith's diagonal weights (Smith et al., 2001); CS, compound symmetry; F, fixed effect of cultivar; FE, fully efficient; LR, location-specific residual variance; SW, Smith's weighting; U, unweighted; W, weighted; ZR, zone-specific residual variance; 1S, single-stage; 2S, 2S, two-stage; EBLUP, empirical best linear unbiased prediction; EBLUE, empirical best linear estimation.

^bCurrent practice.

TABLE 11 Correlations among adjusted cultivar estimates in the winter wheat 2016 dataset with Pearson's product-moment correlation above the diagonal and Spearman's rank correlation below the diagonal

Approach	Strategy ^a				
	1S-CS-LR	2S-CS-W-FE	2S-CS-W-AVVAR	2S-CS-W-SW	2S-F-U-ZR ^b
	EBLUP	EBLUP	EBLUP	EBLUP	EBLUE
1S-CS-LR	1.0000	0.9894	0.9904	0.9895	0.8986
2S-CS-W-FE	0.9866	1.0000	0.9997	0.9997	0.9227
2S-CS-W-AVVAR	0.9881	0.9987	1.0000	0.9999	0.9243
2S-CS-W-SW	0.9872	0.9991	0.9997	1.0000	0.9244
2S-F-U-ZR	0.8889	0.9125	0.9156	0.9144	1.0000

^a1S-CS-LR, single-stage analysis; 2S-CS-W-FE, two-stage fully efficient analysis; 2S-CS-W-AVVAR, two-stage analysis with average variance of a difference (AVVAR) weights (Möhrling & Piepho, 2009); 2S-CS-W-SW, two-stage analysis with Smith's diagonal weights (Smith et al., 2001); CS, compound symmetry; F, fixed effect of cultivar; FE, fully efficient; LR, location-specific residual variance; SW, Smith's weighting; U, unweighted; W, weighted; ZR, zone-specific residual variance; 1S, single-stage; 2S, two-stage; EBLUP, empirical best linear unbiased prediction; EBLUE, empirical best linear estimation.

^bCurrent practice.

TABLE 12 Correlations among adjusted cultivar estimates in the spring barley 2015 dataset with Pearson's product-moment correlation above the diagonal and Spearman's rank correlation below the diagonal

Approach	Strategy ^a				
	1S-CS-LR	2S-CS-W-FE	2S-CS-W-AVVAR	2S-CS-W-SW	2S-F-U-ZR ^b
	EBLUP	EBLUP	EBLUP	EBLUP	EBLUE
1S-CS-LR	1.0000	0.9841	0.9812	0.9812	0.9784
2S-CS-W-FE	0.9721	1.0000	0.9998	0.9998	0.9977
2S-CS-W-AVVAR	0.9704	0.9995	1.0000	1.0000	0.9978
2S-CS-W-SW	0.9704	0.9996	1.0000	1.0000	0.9978
2S-F-U-ZR	0.9644	0.9964	0.9964	0.9964	1.0000

^a1S-CS-LR, single-stage analysis; 2S-CS-W-FE, two-stage fully efficient analysis; 2S-CS-W-AVVAR, two-stage analysis with average variance of a difference (AVVAR) weights (Möhrling & Piepho, 2009); 2S-CS-W-SW, two-stage analysis with Smith's diagonal weights (Smith et al., 2001); 2S-F-U-ZR, current practice method; CS, compound symmetry; F, fixed effect of cultivar; FE, fully efficient; LR, location-specific residual variance; SW, Smith's weighting; U, unweighted; W, weighted; ZR, zone-specific residual variance; 1S, single-stage; 2S, two-stage; EBLUP, empirical best linear unbiased prediction; EBLUE, empirical best linear estimation.

^bCurrent practice.

accurate predictions (Smith, Cullis, & Thompson, 2005). From a biological perspective, the cultivars can be considered as a random sample of the current genetic variability (Curti, de la Vega, Andrade, Bramardi, & Bertero, 2014). Regarding the use of complex variance-covariance structures, the MSEP results revealed that models with complex variance-covariance structures were likely to be overfitted, since the MSEP was larger for these complex models than for models with simpler variance-covariance structures. An incorrect variance-covariance matrix structure still provides unbiased parameter estimates for fixed effects but not for the associated SEs. However, in this study, the fixed effects were fitted for zone only. This effect is not the main interest. Moreover, when the $C \times Z$ interactions effects were fixed, a variance-covariance structure was not applicable. However, it is true that the choosing a suitable variance-covariance matrix is essential for obtaining reliable results via BLUP.

With only a small number of zones (i.e., three zones), there may be no or only little gain in applying FA1 or UN variance-covariance structures. In fact, with three zones, the FA1 structure has the same number of parameters as the UN structure (i.e., six parameters). The variance component estimates for $C \times Z$ effects, as presented in Table 4, were relatively small compared with the other components. When the variance component estimates are small, there may be no need for complex variance-covariance structures. Nevertheless, when the variance component estimates are large, then more parameters with complex variance-covariance might be needed to account for the heterogeneity of variance for $C \times Z$ effects. In a cross-validation study of a similar dataset, comparing the EBLUE and EBLUP methods, it was also revealed that complex variance-covariance structures did not improve the accuracy of yield predictions (Buntaran et al., 2019). To the best of our knowledge, a cross-validation study has not

previously been conducted for comparing single-stage and stagewise analyses of MET data.

We applied the four top-performing strategies and the current strategy to actual datasets. The prediction differences among the 1S-CS-LR, 2S-CS-W-FE, 2S-CS-W-AVVAR, and 2S-CS-W-SW strategies were very small. For that reason, the choice of strategy depends on computational resources (Gogel et al., 2018). The two-stage weighting strategy is preferable when the computational resources are limited. Furthermore, in some cases, data from all the locations or trials may not be available at once. In this case, using two-stage analysis will be more practical, since (a) the readily available trial data can be analyzed instantly and provide individual trial information, and (b) there are savings in time by storing the adjusted means of each trial and the accompanying precision measures, which later can be used for Stage 2 analyses while data from other trials are still being collected. The two-stage weighting strategy may be preferable, especially when each trial has a different experimental design. In this case, fitting a model for single-stage analysis may not be easy. The two-stage weighting method is also preferable if one wants to assess each trial thoroughly because with a vast number of trials, it will be difficult to check each trial thoroughly with a single-stage analysis because of the abundance of variance component estimates produced by the single-stage analysis. However, it should be noted that with the current software used in this study, the two-stage FE analysis needs more memory allocated for conducting the analysis in Stage 2 and obtaining the EBLUPs than the other two-stage analyses (2S-CS-W-AVVAR and 2S-CS-W-SW). Other available software packages might be more suitable to perform two-stage fully efficient analysis but not the single-stage analysis. The software used in this study was designed to perform single-stage analyses. The R code for four strategies (i.e., 1S-CS-LR, 2S-CS-W-FE, 2S-CS-W-AVVAR, and 2S-CS-W-SW) is available in Supplemental File S1.

A single-stage analysis has the theoretical benefit that estimating fixed effects and predicting random effects are done under the assumed single-stage model (Piepho et al., 2012). Moreover, the optimality of the performance of the single-stage analysis has been confirmed via simulation by Welham et al. (2010). Gogel et al. (2018) recommended using one-stage analysis for MET datasets with only a few trials. However, their objective was to obtain accurate predictions for individual locations, whereas in our study, we aimed at accurate predictions for zones. As Damesa et al. (2017) pointed out, it is more informative to obtain predictions per agro-ecological zone or a broader TPE than predictions for individual field trial sites because farmers are interested in cultivars that perform well on average in broad environmental conditions and the next grow-

ing season. The next growing season may be considered as a new environmental condition that no trials have previously been conducted in.

From a breeder's perspective, predicting cultivar performance in a specific trial site is rarely of interest. Official Swedish cultivar trials have the same objective (i.e., to recommend cultivars that perform well for each zone, not for individual trial locations). Thus accurate information regarding which cultivars perform well on average within zones or perform above average across locations is essential for both farmers and breeders.

With the widespread use of linear mixed models for analyzing MET data, the frequent question is whether to model cultivar effects as fixed or random. We recommend modeling cultivar effects as random when the primary goal is to select the best cultivars from the population under study and when the effects and residuals presumably follow a normal distribution. In this case, BLUP outperforms BLUE with regard to agreement between predictions of cultivar rankings and true rankings (McCulloch, Searle, & Neuhaus, 2008; Searle, Casella, & McCulloch, 1992). The shrinkage feature will avoid overoptimistic predictions of the top-performing cultivars and over-pessimistic predictions of poorly performing cultivars. Furthermore, with random $C \times Z$ effects, the accuracy of predictions within zones is improved, since information is borrowed across zones by exploiting the cultivar correlations between zones (Atlin, Baker, McRae, & Lu, 2000; Kleinknecht et al., 2013; Piepho et al., 2016). However, it is important to note that when the genotype correlations between or among zones are small, then the information that can be borrowed across zone will be very little. In this case, BLUP will not be more beneficial than BLUE. However, when the cultivar correlations between or among zones are high, then BLUP will be favorable to BLUE.

Henderson (1963) showed the derivation of BLUPs in the mixed-model equations without assuming a normal distribution. Besides, as pointed out by Lee, Nelder, and Pawitan (2017, p.144), to support the benefits of using BLUP: "With a random effect specification, we gain significant parsimony. In such situations, even if the true model is the fixed effect model, i.e., there is no random sampling involved, the use of random effect estimation has been advocated as shrinkage estimation (James & Stein, 1992)."

The concept of two-stage analysis can be viewed as being similar to Bayesian Updating (Sorensen & Gianola, 2002). The idea of Bayesian Updating is to use the prior distribution from the previous posterior distribution. In this case, the Bayes theorem has "memory" and the inferences can be updated sequentially. In comparison with the two-stage analysis, the result of Stage 1 can be regarded as a posterior distribution that will be used as the prior distribution for Stage 2. Moreover, BLUP is empirical Bayesian when

the distribution of random effects is Gaussian (Robinson, 1991). Thus, Bayesian Updating might be comparable with the “frequentist” BLUP of two-stage analysis. A further study comparing the “frequentist” two-stage analysis with the Bayesian Updating framework would be worthwhile.

The merit of BLUP will only be valid when data are missing at random. If the missing data pattern is not at random, biased variance component estimates may occur (Piepho & Möhring, 2006). Another difficulty with missing data when BLUP is used is that if some varieties are missing at some locations, their predictions will be shrunk towards the overall mean to a larger extent than the other varieties. Thus varieties that are actually good but have been experimented with less will probably not come out as top-performing in the analysis of the study. This feature of BLUP is still beneficial because a prediction based on little information is uncertain (both upwards and downwards). Best linear unbiased prediction therefore protects against poor decisions for cultivar selection. A good cultivar candidate still needs to be tested across a vast number of trials to obtain reliable information of its performance.

Although we recommend the use of BLUP for estimating cultivar effects, we discourage the use of BLUP in Stage 1 of the two-stage approach. The use of BLUP in Stage 1 will cause double shrinkage, since BLUP is also used in Stage 2. If BLUP were to be used in Stage 1, predictions would need to be unshrunk before proceeding to Stage 2 (Smith et al., 2001) but it is not obvious how this should be done or how to perform the weighting in Stage 2. Some progress could potentially be made by taking recourse to the so-called “deregressed proofs” as used in animal breeding (Calus, Vandenplas, ten Napel, & Veerkamp, 2016).

5 | CONCLUSION

This cross-validation study provided insights into the performance of single-stage and two-stage strategies. The two-stage weighting strategy (FE, AVVAR and SW) performed similarly to the single-stage analysis with location-specific residual variances. The choice between a single-stage or a two-stage strategy depends on the computational resources, since the loss of information caused by diagonal approximate weighting is negligible. In our study, with only three zones, complex variance–covariance structures were not necessary, since these caused overfitting. We recommend modeling the effects of cultivar and the $C \times Z$ interaction as random, because this improves the accuracy of zone-based prediction through borrowing information across zones. Predictions for zones are more useful and informative for farmers and breeders than predictions for individual locations, since zones cover broader TPEs.


ACKNOWLEDGEMENTS

This research was funded by Stiftelsen Lantbruksforskning – Swedish Farmers’ Foundation for Agricultural Research (Project No. O-17-20-963).


CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

ORCID

Harimurti Buntaran  <https://orcid.org/0000-0001-8917-3781>

Paul Schmidt  <https://orcid.org/0000-0003-1528-2082>

Johannes Forkman  <https://orcid.org/0000-0002-5796-0710>

REFERENCES

- Atlin, G. N., Baker, R. J., McRae, K. B., & Lu, X. (2000). Selection response in subdivided target regions. *Crop Science*, 40, 7–13. <https://doi.org/10.2135/cropsci2000.4017>
- Buntaran, H., Piepho, H.-P., Hagman, J., & Forkman, J. (2019). A cross-validation of statistical models for zoned-based prediction in cultivar testing. *Crop Science*, 59, 1544–1553. <https://doi.org/10.2135/cropsci2018.10.0642>
- Butler, D. G., Cullis, B. R., Gilmour, A. R., Gogel, B. J., & Thompson, R. (2017). *ASReml-R reference manual, version 4*. Wollongong, NSW: Univ. of Wollongong.
- Calus, M. P. L., Vandenplas, J., ten Napel, J., & Veerkamp, R. F. (2016). Validation of simultaneous deregression of cow and bull breeding values and derivation of appropriate weights. *Journal of Dairy Science*, 99, 6403–6419. <https://doi.org/10.3168/jds.2016-11028>
- Comstock, R. (1977). Quantitative genetics and the design of breeding programme. In E. Pollak, O. Kempthorne, & T. B. Bailey (Eds.), *International Conference on Quantitative Genetics* (pp. 705–718). Ames, IA: Iowa State Univ. Press.
- Cooper, M., & Hammer, G. L. (1996). *Plant adaptation and crop improvement*. Wallingford, UK: CAB International.
- Cooper, M., Messina, C. D., Podlich, D., Totir, L. R., Baumgarten, A., Hausmann, N. J., ... Graham, G. (2014). Predicting the future of plant breeding: Complementing empirical evaluation with genetic prediction. *Crop Pasture Sci*, 65(4), 311–336. <https://doi.org/10.1071/CP14007>
- Curti, R. N., de la Vega, A. J., Andrade, A. J., Bramardi, S. J., & Bertero, H. D. (2014). Multi-environmental evaluation for grain yield and its physiological determinants of quinoa genotypes across north-west Argentina. *Field Crops Research*, 166, 46–57. <https://doi.org/10.1016/j.fcr.2014.06.011>
- Damesa, T. M., Möhring, J., Worku, M., & Piepho, H.-P. (2017). One step at a time: Stage-wise analysis of a series of experiments. *Agronomy Journal*, 109, 845–857. <https://doi.org/10.2134/agronj2016.07.0395>
- Forkman, J. (2013). The use of a reference variety for comparisons in incomplete series of crop variety trials. *Journal of Applied Statistics*, 40, 2681–2698. <https://doi.org/10.1080/02664763.2013.825703>
- Gauch, H. G., Hwang, J. T. G., & Fick, G. W. (2003). Model evaluation by comparison of model-based predictions and measured values. *Agronomy Journal*, 95, 1442–1446. <https://doi.org/10.2134/agronj2003.1442>

- Gogel, B., Smith, A., & Cullis, B. (2018). Comparison of a one- and two-stage mixed model analysis of Australia's National Variety Trial Southern Region wheat data. *Euphytica*, 214, 44. <https://doi.org/10.1007/s10681-018-2116-4>
- Harville, D. A. (1991). [That BLUP is a good thing: The estimation of random effects]: Comment. *Statistical Science*, 6, 35–39.
- Haslett, S., & Welsh, A. H. (2019). EBLUPs: Empirical best linear unbiased predictors. In N. Balakrishnan, T. Colton, B. Everitt, W. Piegorisch, F. Ruggeri, & J. L. Teugels (Eds.), *Wiley StatsRef: Statistics reference online*. <https://doi.org/10.1002/9781118445112.stat08180>
- Henderson, C. R. (1963). Selection index and expected genetic advance. In W. D. Hanson & H. F. Robinson (Eds.), *Statistical genetics and plant breeding* (pp. 141–163). Washington DC: National Academy of Sciences, National Research Council.
- Henderson, C. R. (1978). Undesirable properties of regressed least squares prediction of breeding values. *Journal of Dairy Science*, 61, 114–120. [https://doi.org/10.3168/jds.S0022-0302\(78\)83559-0](https://doi.org/10.3168/jds.S0022-0302(78)83559-0)
- Isik, F., Holland, J., & Maltecca, C. (2017). Multi environmental trials. *Genetic data analysis for plant and animal breeding*. (pp. 227–262). Cham: Springer International Publishing.
- James, W., & Stein, C. (1992). Estimation with quadratic loss. In S. Kotz & N. L. Johnson (Eds.), *Breakthroughs in statistics: Foundations and basic theory*. (pp. 443–460). New York: Springer New York.
- Kleinknecht, K., Möhring, J., Singh, K. P., Zaidi, P. H., Atlin, G. N., & Piepho, H. P. (2013). Comparison of the performance of best linear unbiased estimation and best linear unbiased prediction of genotype effects from zoned Indian maize data. *Crop Science*, 53, 1384–1391. <https://doi.org/10.2135/cropsci2013.02.0073>
- Kobayashi, K., & Salam, M. U. (2000). Comparing simulated and measured values using mean squared deviation and its components. *Agronomy Journal*, 92, 345–352. <https://doi.org/10.2134/agronj2000.922345x>
- Lee, Y., Nelder, J. A., & Pawitan, Y. (2017). *Generalized linear models with random effects: Unified analysis via H-likelihood*. Boca Raton, FL: CRC Press.
- McCulloch, C. E., Searle, S. R., & Neuhaus, J. M. (2008). *Generalized, linear, and mixed models*. Hoboken, NJ: John Wiley & Sons, Inc.
- Möhring, J., & Piepho, H.-P. (2009). Comparison of weighting in two-stage analysis of plant breeding trials. *Crop Science*, 49, 1977–1988. <https://doi.org/10.2135/cropsci2009.02.0083>
- Patterson, H. D. (1997). Analysis of series of variety trials. In R. A. Kempton, P. N. Fox, & M. Cerezo (Eds.), *Statistical methods for plant variety evaluation*. (p. 139–161). Dordrecht: Springer Netherlands.
- Piepho, H.-P. (1998). Empirical best linear unbiased prediction in cultivar trials using factor-analytic variance-covariance structures. *Theoretical and Applied Genetics*, 97, 195–201. <https://doi.org/10.1007/s001220050885>
- Piepho, H.-P., & Eckl, T. (2014). Analysis of series of variety trials with perennial crops. *Grass and Forage Science*, 69, 431–440. <https://doi.org/10.1111/gfs.12054>
- Piepho, H.-P., & Möhring, J. (2006). Selection in cultivar trials—is it ignorable? *Crop Science*, 46, 192–201. <https://doi.org/10.2135/cropsci2005.04-0038>
- Piepho, H.-P., Möhring, J., Schulz-Streeck, T., & Ogutu, J. O. (2012). A stage-wise approach for the analysis of multi-environment trials. *Biometrical Journal*, 54, 844–860. <https://doi.org/10.1002/bimj.201100219>
- Piepho, H.-P., Nazir, M. F., Qamar, M., Rattu, A.-u.-R., Riaz-ud-Din, Hussain, M., ... Imtiaz, M. (2016). Stability analysis for a country-wide series of wheat trials in Pakistan. *Crop Science*, 56, 2465–2475. <https://doi.org/10.2135/cropsci2015.12.0743>
- R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.
- Robinson, G. K. (1991). That BLUP is a good thing: The estimation of random effects. *Statist. Science.*, 6, 15–32. <https://doi.org/10.1214/ss/1177011926>
- RStudio Team. (2016). *RStudio: Integrated development environment for R*. Boston, MA: RStudio, Inc.
- Schulz-Streeck, T., Ogutu, J. O., & Piepho, H.-P. (2013). Comparisons of single-stage and two-stage approaches to genomic selection. *Theoretical and Applied Genetics*, 126, 69–82. <https://doi.org/10.1007/s00122-012-1960-1>
- Searle, S. R., Casella, G., & McCulloch, C. E. (1992). *Variance components*. New York: John Wiley & Sons
- Smith, A., Cullis, B. R., & Gilmour, A. (2001). Applications: The analysis of crop variety evaluation data in Australia. *Australian & New Zealand Journal of Statistics*, 43, 129–145. <https://doi.org/10.1111/1467-842X.00163>
- Smith, A. B., Cullis, B. R., & Thompson, R. (2005). The analysis of crop cultivar breeding and evaluation trials: An overview of current mixed model approaches. *The Journal of Agricultural Science*, 143, 449–462. <https://doi.org/10.1017/S0021859605005587>
- Sorensen, D., & Gianola, D. (2002). Bayesian updating. *Likelihood, Bayesian, and MCMC methods in quantitative genetics*. (p. 249–257). New York, NY: Springer New York.
- van Eeuwijk, F. A., Bustos-Korts, D. V., & Malosetti, M. (2016). What should students in plant breeding know about the statistical aspects of genotype \times environment interactions? *Crop Science*, 56, 2119–2140. <https://doi.org/10.2135/cropsci2015.06.0375>
- Welham, S. J., Gogel, B. J., Smith, A. B., Thompson, R., & Cullis, B. R. (2010). A comparison of analysis methods for late-stage variety evaluation trials. *Australian & New Zealand Journal of Statistics*, 52, 125–149. <https://doi.org/10.1111/j.1467-842X.2010.00570.x>
- Wickham, H. (2009). *ggplot2: Elegant graphics for data analysis*. New York: Springer-Verlag.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Buntaran H, Piepho H-P, Schmidt P, Rydén J, Halling M, Forkman J. Cross-validation of stagewise mixed-model analysis of Swedish variety trials with winter wheat and spring barley. *Crop Science*. 2020;1–20. <https://doi.org/10.1002/csc2.20177>

APPENDIX

(1) Variance–covariance structures for zone • location • replicate and zone • location • replicate • block in the single-stage model, Model 1

(1a) Heterogeneous variance–covariance structures for zone•location•replicate and zone•location•replicate•block

The heterogeneous variance–covariance structure of zone • location • replicate ($Z \cdot L \cdot R$) was $\mathbf{G}_{ZLR} = \bigoplus_{j=1}^J \mathbf{G}_{ZLR(j)}$, where J is the number of locations, $\mathbf{G}_{ZLR(j)}$ is a $K_j \times K_j$ diagonal matrix with the diagonal elements $\sigma_{ZLR(j)}^2$ and K_j is the number of replicates in the j -th location. The variances of this variance–covariance structure are location-specific:

$$\mathbf{G}_{ZLR(j)} = \begin{bmatrix} \sigma_{ZLR(j)}^2 & 0 & \dots & 0 \\ 0 & \sigma_{ZLR(j)}^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_{ZLR(j)}^2 \end{bmatrix}. \quad (\text{A1})$$

The heterogeneous variance–covariance structure for zone • location • replicate • block ($Z \cdot L \cdot R \cdot B$) was $\mathbf{G}_{ZLRB} = \bigoplus_{j=1}^J \mathbf{G}_{ZLRB(j)}$, where $\mathbf{G}_{ZLRB(j)}$ is an $M_j \times M_j$ diagonal matrix with the elements $\sigma_{ZLRB(j)}^2$, and M_j is the number of blocks in the j -th location. The variances of this variance–covariance structure are location-specific:

$$\mathbf{G}_{ZLRB(j)} = \begin{bmatrix} \sigma_{ZLRB(j)}^2 & 0 & \dots & 0 \\ 0 & \sigma_{ZLRB(j)}^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_{ZLRB(j)}^2 \end{bmatrix}. \quad (\text{A2})$$

(1b) Homogeneous variance–covariance structures for $Z \cdot L \cdot R$ and $Z \cdot L \cdot R \cdot B$

The covariance structure for $Z \cdot L \cdot R$ was $\mathbf{G}_{ZLR} = \mathbf{I}\sigma_{ZLR}^2$:

$$\mathbf{G}_{ZLR} = \begin{bmatrix} \sigma_{ZLR}^2 & 0 & \dots & 0 \\ 0 & \sigma_{ZLR}^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_{ZLR}^2 \end{bmatrix} \quad (\text{A3})$$

The covariance structure for $Z \cdot L \cdot R \cdot B$ was $\mathbf{G}_{ZLRB} = \mathbf{I}\sigma_{ZLRB}^2$:

$$\mathbf{G}_{ZLRB} = \begin{bmatrix} \sigma_{ZLRB}^2 & 0 & \dots & 0 \\ 0 & \sigma_{ZLRB}^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_{ZLRB}^2 \end{bmatrix}. \quad (\text{A4})$$

(2) Variance–covariance structure for zone • cultivar • location in Models 1, 3, and 4

For all models, the covariance structure for zone • cultivar • location ($Z \cdot C \cdot L$) was $\mathbf{G}_{ZCL} = \mathbf{I}\sigma_{ZCL}^2$:

$$\mathbf{G}_{ZCL} = \begin{bmatrix} \sigma_{ZCL}^2 & 0 & \dots & 0 \\ 0 & \sigma_{ZCL}^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_{ZCL}^2 \end{bmatrix}. \quad (\text{A5})$$

(3) Variance–covariance structures for cultivar (zone) = cultivar + $C \times Z$ in Models 1, 3, and 4

We considered three different variance–covariance structures for $C \times Z$ effects: (a) ID, (b) CS, (c) UN, and (3) FA. In the ID structure, the cultivar main effect is excluded in the model because cultivars are independent among zones. In the CS structure, the cultivar main effect is included because the cultivars are correlated among zones. In the UN and FA structures, because of the nonzero covariance, the model has to be reparameterized by dropping the cultivar main effect in order to avoid overparameterization. The dimension of the \mathbf{G}_{CZ} covariance matrix is the number of cultivars multiplied by the number of zones. For the CS, UN, and FA models, this matrix is block diagonal, $\mathbf{G}_{CZ} = \bigoplus_{i=1}^I \mathbf{G}_{CZ(i)}$, where I is the number of cultivars. The size of the blocks, $\mathbf{G}_{CZ(i)}$, is the number of zones.

(3a) Identity

The identity or identical variance–covariance structure assumes independence and homogeneity, $\mathbf{G}_{CZ} = \mathbf{I}(\sigma_{CZ}^2)$:

$$\mathbf{G}_{CZ} = \begin{bmatrix} \sigma_{CZ}^2 & 0 & \dots & 0 \\ 0 & \sigma_{CZ}^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_{CZ}^2 \end{bmatrix}. \quad (\text{A6})$$

(3b) Compound symmetry

The CS model implies that both the variance and covariance are homogeneous: $\mathbf{G}_{CZ} = \mathbf{J}\sigma_C^2 + \mathbf{I}\sigma_{CZ}^2$. The structure of the CS variance–covariance matrix is:

$$\mathbf{G}_{CZ(i)} = \begin{bmatrix} \sigma_C^2 + \sigma_{CZ}^2 & \sigma_C^2 & \dots & \sigma_C^2 \\ \sigma_C^2 & \sigma_C^2 + \sigma_{CZ}^2 & \dots & \sigma_C^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_C^2 & \sigma_C^2 & \dots & \sigma_C^2 + \sigma_{CZ}^2 \end{bmatrix}. \quad (\text{A7})$$

The correlation between zones is $\frac{\sigma_C^2}{\sigma_C^2 + \sigma_{CZ}^2}$.

(3c) Unstructured

The unstructured variance–covariance structure allows both heterogeneous covariance and variance. Thus each zone has a unique cultivar variance and each pair of zones has a unique covariance. The number of parameters needed for this variance–covariance structure is $\frac{p(p+1)}{2}$, where p is the number of zones. In this study, six parameters were needed for three zones.

$$\mathbf{G}_{CZ(i)} = \begin{bmatrix} \sigma_{Z_1}^2 & \sigma_{Z_{12}}^2 & \dots & \sigma_{Z_{1p}}^2 \\ \sigma_{Z_{21}}^2 & \sigma_{Z_2}^2 & \dots & \sigma_{Z_{2p}}^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{Z_{p1}}^2 & \sigma_{Z_{p2}}^2 & \dots & \sigma_{Z_p}^2 \end{bmatrix}. \quad (\text{A8})$$

(3d) Factor-analytic Order 1

Factor analytic structures are often more useful than the UN structure for taking heterogeneity into account in complex genotype \times environment models. These structures have fewer parameters than the UN structure (Isik, Holland, & Maltecca, 2017). We used a FA structure with a single multiplicative term (FA1). According to this structure, $\mathbf{G}_{CZ(i)}$ is modeled as $\mathbf{\Lambda}\mathbf{\Lambda}^T + \mathbf{\Psi}$, where $\mathbf{\Lambda}$ is a vector of dimension $1 \times p$, which consists of factors loading λ_1 to λ_p , and $\mathbf{\Psi}$ is $p \times p$ diagonal matrix, which consists of zone-specific cultivar variances ψ_1^2 to ψ_p^2 . For the FA1 model with i unrelated cultivars tested in p zones, this is:

$$\mathbf{\Lambda} = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_p \end{bmatrix}, \mathbf{\Psi} = \begin{bmatrix} \psi_1^2 & 0 & \dots & 0 \\ 0 & \psi_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \psi_p^2 \end{bmatrix}. \quad (\text{A9})$$

Hence, the variance–covariance structure for $\mathbf{G}_{CZ(i)}$ is:

$$\mathbf{G}_{CZ} = [\mathbf{\Lambda}\mathbf{\Lambda}^T + \mathbf{\Psi}] = \begin{bmatrix} \lambda_1^2 + \psi_1^2 & \lambda_1\lambda_2 & \dots & \lambda_1\lambda_p \\ \lambda_2\lambda_1 & \lambda_2^2 + \psi_2^2 & \dots & \lambda_2\lambda_p \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_p\lambda_1 & \lambda_p\lambda_2 & \dots & \lambda_p^2 + \psi_p^2 \end{bmatrix}. \quad (\text{A10})$$

The off-diagonal elements of these blocks, $\mathbf{\Lambda}\mathbf{\Lambda}^T + \mathbf{\Psi}$, are the products of the parameters λ_p and $\lambda_{p'}$, which refer to the p -th and p' -th zone, respectively. Thus the nested effects of the same cultivar in different zones are correlated, whereas the interaction effects from different cultivars are uncorrelated.

(4) Variance–covariance structure for residual variance

Three different variance–covariance structures for the residual variance were used in this study: (a) the ID residual structure, (b) the heterogeneous residual structure with zone-specific variance (ZR), and (c) the heterogeneous residual structure with location-specific variance (LR).

(4a) Identity structure

The identity residual structure assumes homoscedasticity with $\mathbf{R} = \sigma^2 \mathbf{I}$. The matrix form is as follows:

$$\mathbf{R} = \begin{bmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{bmatrix}. \quad (\text{A11})$$

This structure was used only in the single-stage analysis.

(4b) Heterogeneous residual zone-specific structure

The heterogeneous ZR variance–covariance structure is $\mathbf{R} = \oplus_{p=1}^P \mathbf{R}_p$, where \mathbf{R}_p is a diagonal matrix, for the p -th zone, with the diagonal element $\sigma_{\epsilon(p)}^2$. This structure can be implemented in a single-stage analysis and in a two-stage analysis. The two-stage analysis is the current

Swedish practice (EBLUE C \times Z).

$$\mathbf{R}_p = \begin{bmatrix} \sigma_{\varepsilon(p)}^2 & 0 & \dots & 0 \\ 0 & \sigma_{\varepsilon(p)}^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_{\varepsilon(p)}^2 \end{bmatrix}. \quad (\text{A11})$$

(4c) Heterogeneous residual location-specific structure (LR)

The heterogeneous LR variance–covariance structure, $\mathbf{R} = \oplus_{j=1}^J \mathbf{R}_j$, where \mathbf{R}_j is a diagonal matrix for the j -th location with the diagonal element $\sigma_{\varepsilon(j)}^2$, is as follows:

$$\mathbf{R}_j = \begin{bmatrix} \sigma_{\varepsilon(j)}^2 & 0 & \dots & 0 \\ 0 & \sigma_{\varepsilon(j)}^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_{\varepsilon(j)}^2 \end{bmatrix}. \quad (\text{A12})$$

This structure can be implemented in a single-stage analysis and in a two-stage analysis.