Assessment of statistical analysis of Swedish cultivar testing

A cross-validation study for model selection

Harimurti Buntaran
Faculty of Natural Resources and Agricultural Sciences
Department of Energy and Technology
Uppsala

Licentiate thesis
Swedish University of Agricultural Sciences
Uppsala 2019

Cover: Reaction norm of cross-over genotype×environment interactions
(illustration: Harimurti Buntaran)

Licentiate thesis/Report 107
ISSN 1654-9406
ISBN (print version) 978-91-7760-438-9
ISBN (electronic version) 978-91-7760-439-6
© 2019 Harimurti Buntaran, Uppsala
Print: SLU Service/Repro, Uppsala 2019

Assessment of statistical analysis of Swedish cultivar testing: A cross-validation study for model selection

Abstract

The Swedish official cultivar testing conducts multienvironmental trials (MET) to make recommendations of cultivars that are well adapted to farmers' regional conditions. In the MET, a large number of cultivars are tested in several geographical regions. The tested cultivars perform differently in varying soil types and climates, a phenomenon known as genotype×environment interactions. The MET data structure is often large and highly imbalanced, which causes computational problems when applying some statistical methods. Several issues, such as prediction of crop variety performance and efficient computation of measure of cultivar stability are urgent to be tackled by developing comprehensive and robust statistical methods**.** This study aims to address these issues and provide a gold standard for MET analysis in Swedish official cultivar testing.

In this study, we investigated several linear mixed models by using cross-validation (CV). We proposed to use random cultivar effects, known as best linear unbiased prediction (BLUP) method to replace the current fixed cultivar effects, known as best linear unbiased estimation (BLUE). In theory, BLUP provides more accurate rankings and predictions than BLUE. The current-practice analysis strategy, i.e., two-stage unweighted strategy, was also compared to several strategies such as single-stage strategy and two-stage weighted strategies that comprise some weighting methods. In the CV, mean squared error of differences (MSEP) was used to assess the performance of estimation of cultivar effects by BLUP and BLUE to select a model that provides best prediction accuracy. A new inter-zone stability measure was also proposed to tackle computational burden and provide additional useful information regarding cultivar stability across zones and years.

The MSEP revealed that BLUP outperformed the current-practice method, BLUE, and so improved the accuracy of zone-based prediction. Also, the single-stage and two-stage weighted strategies outperformed the current strategy. The proposed stability measure offered a less computational resource, and provided more flexible stability measure for practical purpose.

Keywords: BLUE, BLUP, cross-validation, genotype×environment interactions, linear mixed models, multienvironment trials, stability, stage-wise analysis

Author's address: Harimurti Buntaran, SLU, Department of Energy and Technology, P.O. Box 7032, 750 07 Uppsala, Sweden
E-mail: harimurti.buntaran@slu.se

Dedication

To Breeders, Farmers, and Applied Statisticians

All models are wrong but some are useful.
George E. P. Box

Contents

List of publications

This thesis is based on the work contained in the following Papers, referred to by Roman numerals in the text:

I Buntaran H.*, Piepho, H.-P., Hagman, J & Forkman J. (2018). Performance of empirical BLUE and empirical BLUP in Swedish crop variety trials. *Biuletyn Oceny Odmian*, 35, pp. 15–17, ISBN 978-83-86224-12-8

II 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 (4), pp. 1544–1553. doi: 10.2135/cropsci2018.10.0642

III Buntaran H.*, Piepho, H.-P., Schmidt, P., Rydén, J., Halling, M., Forkman, J. (2019). One-stage or two-stage? Cross-validation says, “It does not matter, if the two-stage is weighted.” (manuscript).

* Corresponding author.

The contribution of Harimurti Buntaran to the Papers included in this thesis was as follows:

I In collaboration with the co-authors participated in the study planning, performed the analysis, wrote the main part of the manuscript, and was responsible for correspondence with the journal.

II In collaboration with the co-authors participated in the study planning, performed the analysis, wrote the main part of the manuscript, and was responsible for correspondence with the journal.

III In collaboration with the co-authors participated in the study planning, performed the analysis, and wrote the main part of the manuscript.

List of tables

List of figures

Abbreviations

BLUE	Best linear unbiased estimation
BLUP	Best linear unbiased prediction
CPVO	Community plant variety office
CPVR	Community plant variety rights
CS	Compound symmetry
CV	Cross-validation
DUS	Distinctness, uniformity, and stability
EBLUE	Empirical best linear unbiased estimation
EBLUP	Empirical best linear unbiased prediction
FA	Factor analytic
FA1	Factor analytic order 1
FAO	Food and agriculture organisation
GEI	Genotype×environment interactions
GLS	Generalised least squares
ID	Identity
LOO	Leave-one-out
MAR	Missing at random
MET	Multienvironment trials
MF	Multi-year and fixed effects for cultivar
ML	Maximum likelihood
MNAR	Missing-not-at-random
MR	Multi-year and random effects for cultivar
MSE	Mean squared error
MSEP	Mean squared error of prediction differences
MYF	Multi-year and fixed effects for cultivar
MYR	Multi-year and random effects for cultivar
OLS	Ordinary least squares
REML	Residual/Restricted maximum likelihood
SF	Single-year and fixed effects for cultivar
SR	Single-year and random effects for cultivar
SYF	Single-year and fixed effects for cultivar

SYR	Single-year and random effects for cultivar
TPE	Target population of environments
TPG	Target population of genotypes
US	Unstructured
VCOV	Variance-covariance

1 Introduction

Food and Agriculture Organization (FAO) projected that the world's population grows to almost 10 billion by 2050, boosting agricultural demand – in a scenario of modest economic growth – by some 50 percent compared to 2013 (FAO, 2017). Thus, efficient plant breeding and cultivar testing programmes are parts of the solution to meet the growing food demand due to an increasing world population. A critical factor for the success of a plant breeding programme is to select varieties that guarantee high yield and quality in varying environmental conditions because different cultivars perform differently in diverse environments, a phenomenon known as genotype×environment interactions (GEI) (Kang and Gorman, 1989). Therefore, multienvironment trials (MET) are conducted as a crucial part of any plant breeding programme to assess and provide cultivars performance across diverse environmental conditions. In an MET, a large number of cultivars are tested in several geographical regions. A reliable and robust statistical method is required to provide accurate predictions of yield and stability measure of tested cultivars so that the MET results can assist breeders in selecting the best cultivars and providing recommendation for farmers to select well-adapted cultivar to their regional conditions.

1.1 Genotype × environment interactions (GEI)

The main goal of a plant breeding programme is to develop superior cultivars in yield and/or quality across diverse environmental conditions (Malosetti et al., 2013). "Cultivar" is a term defined as a product of plant breeding that is released for access to producers or cultivated variety of a plant (Acquaah, 2012; Fehr, 1987). Breeders face a major challenge to achieve the aim because cultivars are grown across a wide range of environmental conditions. Cultivars are exposed to diverse soil types and fertility levels, temperatures, moisture levels, and

agricultural practices. These variables encountered in crop production can be described collectively as the environment.

When cultivars are compared in different environmental conditions, their performance relative to each other may not be the same. A cultivar may have the highest yield in some environments and another cultivar may outperform in others. This differential response of genotypes across different environments called genotype × environment interactions (GEI) (de Leon et al., 2016). In the GEI concept, the "genotype" term is interchangeably with "variety", "crop", and "cultivar".

The concept of GEI can be depicted as the slope of the line when genotype performance is plotted against an environmental gradient. This concept is also known as the reaction norm: the genotype-specific functional relationship between phenotype and environmental gradients (DeWitt and Scheiner, 2004; van Eeuwijk et al., 2016) as shown in Figure 1. The reaction norm can be illustrated by the combination of GEI and phenotypic plasticity, where phenotypic plasticity is environment-dependent phenotype expression (DeWitt and Scheiner, 2004).

In Figure 1, five scenarios of reaction norm are shown. Figure 1a shows no GEI and no plasticity since there is no different mean of genotype performance across the environments, and the ranking of genotypes are the same across environments. Figure 1b also shows no GEI but plasticity because of the phenotype expression, in this case, yield, changes across the different environment. In Figure 1b, there is no GEI because the genotype and the environment behave additively, and the reaction norms are parallel (no difference ranking and changing mean differences among genotypes). The remaining plots show various situations in which GEI occurs: divergence (Figure 1c), convergence (Figure 1d), and the most crucial one, crossover interaction (Figure 1e). In the case of divergence and convergence, the genotype ranking does not change across environments, but the mean difference between the three genotypes does. In the case of crossover interaction, not only the mean difference between genotypes is changing but also the ranking. Crossover interactions are the most important for breeders as they imply that the selection of the best genotype depends on the specific environment.

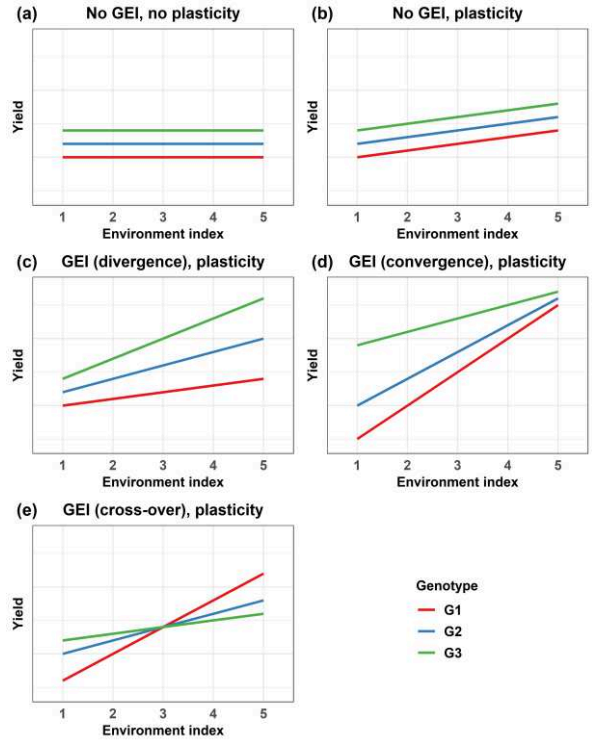

Figure 1. Illustration of GEI for three genotypes in five different environment conditions. No GEI in (a) and (b) versus GEI in (b) until (c). No plasticity in (a) versus plasticity in (b) until (e). The environment index shows the unfavourable environment conditions (1) to favourable environment conditions (5).

1.2 Multienvironment trial (MET)

It is beneficial to introduce the concepts of target population of genotypes (TPG) and target population of environments (TPE) to understand the breeding concepts associated with GEI. The combination of TPG and TPE assists breeders to define the set of genotypes (cultivars/varieties) and environments to obtain valid and precise inference and predictions (van Eeuwijk et al., 2016). The TPG comprises all candidate genotypes to grow the coming years (van Eeuwijk et al., 2016). The TPE contains a group of environments concerning the genotypic performance where new cultivars will be adopted. In other words, TPE describes the future growing conditions of the cultivars in the TPG (Comstock, 1977; Cooper and Hammer, 1996; Cooper et al., 2014). The TPE is also essential to predict GEI since the identification of repeatable GEI is a challenge due to the unpredictable weather (de Leon et al., 2016).

In a breeding programme, developed or improved cultivars are assessed in a multienvironment trial (MET), which comprises potential representatives of the TPE (Cooper and DeLacy, 1994; DeLacy et al., 1996). In the MET, the term "environment" refers to a year-location combination. The objective of an MET is to determine which cultivars matched to a TPE, based on the reaction norm/expression of the cultivars per se to the environments. Thus, METs assist breeders to determine the similarity of environments and grouping similar environments in METs. The results of MET are crucial not only for selection purpose in a breeding programme but also to give advice or recommendation to farmers in deciding which cultivar is the most suitable and performs the best in their growing conditions. Since METs consist of year–location combination, the stability of cultivars is also evaluated. The stability measure is essential for breeders and farmers because, as already mentioned before, the repeatable GEI is crucial to avoid the risk of unstable cultivars in unpredictable changing weather, i.e., loss or lowering of yield. Thus, robust statistical methods are necessary to obtain accurate predictions of genotype performance such as yield, and to obtain a reliable stability measure of each cultivar across environments.

1.3 Stability measures

The MET analysis also provides an assessment of yield stability of the tested genotypes. The term "stability" refers to the behaviour of a crop in varying environments (Piepho, 1996). Environments may be locations, years, or combinations of both. Stability measures are essential for plant breeders since the breeding goal is not only to develop a high-yielding cultivar but also to develop stable cultivars for a range of environments, to meet the demand for food production. The stability measure is also beneficial for farmers in selecting which varieties to grow since the weather is changing inconsistently. Stability measures are estimated mostly for yield compared to quality traits because yield is still the most critical trait in comparison to other traits.

The stability of a genotype or cultivar is measured by variability of yields across environments, e.g. the sample variance (Piepho, 1998b). Lin et al. (1986) considered stability called static (type 1) when a genotype can give the same performance across environments. In shorter words, the reaction norm is flat, as shown in Figure 1(a) (non-plasticity). This type 1 stability is also called "biological" stability since it does not account for differences due to a different environment. Dynamic or agronomic stability (type 2) is defined when genotype performance changes in a predictable way across different environments (plasticity). Wricke (1962) and Shukla (1972) stability measures are examples of type 2 stability while Finlay and Wilkinson (1963) stability measure can show

type 1 and type 2 stability. In Finlay-Wilkinson joint-regression (FW regression), the stability is measured by the regression slope, b_i. If b_i is close to 0, then it is considered to be type 1 stability. If b_i is close to 1, then it is considered to be type 2 stability. Wricke (1962) introduced Wricke's Ecovalence (W_i) defined as the contribution of a genotype to the sum of squares GEI. Shukla (1972) stability variance σ_i^2 is a modification of W_i that σ_i^2 is an estimate of variance of *i*-th genotype across environments based on residuals of the two-way GEI. Thus, the most stable genotype is the one that has smallest σ_i^2.

Eberhart and Russell (1966) extended the FW regression with another measure called "deviation" (s_δ^2) with respect to FW regression. Hence, this method is a combination of b_i from FW regression and the s_δ^2. Eberhart and Russell (1966) stability measure s_δ^2 is known as type 3 stability. Thus, a genotype will be regarded as type 3 stability if a cultivar has the smallest s_δ^2, regardless its b_i. In practice, breeders cannot use only type 3 stability because it does not show the performance of genotype *per se*. Breeders, therefore, must consider the b_i and the mean performance of genotypes across environments. The reason is that it is possible that a genotype has type 3 stability but also has type 1 stability, which indicates that genotype either has the lowest mean or highest mean in all environments.

Lin and Binns (1988) proposed a type 4 stability. This stability not only takes into account the interaction between genotype and location but also with time, i.e., genotype×location×year interactions. Thus, type 4 stability is considered as unpredictable due to the inclusion of time, and it is the opposite of the type 2 stability. The type 2 stability defines as the changes of genotype performance to be predictable to environmental alterations because the response of genotype to environments is parallel to the mean response of all genotypes in the trial (Lin and Binns, 1988). The type 4 stability is measured by the mean squares (MS) of year nested by location (Y/L) of each genotype. The genotype with the smallest MS (Y/L) is considered as the most stable genotype.

The stability measures are important to provide information regarding the response of cultivars to unpredictable changing environments. Breeders want to select not only the cultivar that has the highest yield but also the stable one across environment (location and years). Farmers will also prefer the same cultivar behaviour since the unpredictable changing weather will risk the yield loss. Thus, an appropriate stability measure type and computation method are crucial to provide the information for breeders and farmers.

1.4 Swedish cultivar testing

Swedish cultivar testing conducts METs every year to test vast numbers of cultivars to be registered in variety registration as a precondition for the seed certification. The seed certification is a compulsory requirement to be able to enter the seeds market in Sweden and EU. Thus, a new variety can enter the seed market and cultivated as a cultivar only if it is admitted to the Swedish list of varieties or admitted to the common catalogues of varieties of agricultural plant or vegetable species (Jordbruksverket, 2015).

Furthermore, Swedish cultivar testing also assess the new plant varieties to protect the breeder's right. Breeders cannot rely on supplying existing varieties/cultivars. To meet the demand of cultivar improvement, such as enhanced quality, disease-resistance, productivity and environmental criteria, new varieties need to be developed. The Community Plant Variety Rights (CPVR) system incorporates the principle of the breeders' exemption, which allows free access to protected varieties for the development and exploitation of new plant varieties.

Thus, besides the seed certification, a new variety that will be cultivated as a cultivar has to be assessed for to ensure the breeder's right. According to the Community Plant Variety Office (CPVO), a new cultivar has to meet three criteria: distinctness, uniformity, and stability (DUS) (CPVO, 2018).

The results of this assessment can provide a cultivar recommendation for farmers in their growing regions or zones. In Sweden, the most important growing zones are depicted in Figure 2, i.e., South (A), Middle (D+E), and North (F). A zone consists of several trials/locations that represent farmers' growing conditions. Swedish cultivar testing provides cultivar recommendation on zone-based, not each trials/locations.

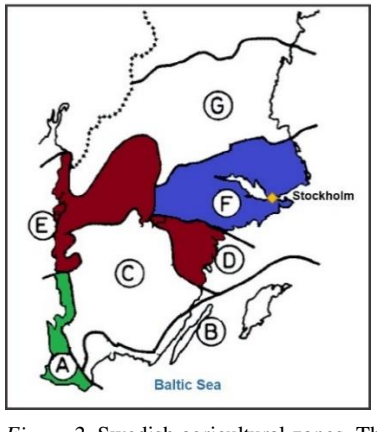

Figure 2. Swedish agricultural zones. The coloured zones indicates the zones that are used in the study. Green, south zone (A); Brown, middle zone (D+E); Blue, north zone (F).

1.5 Linear mixed models

A model is defined as a mathematical notation of the processes that give rise to the observations in a set of data (Stroup, 2012). A purely mathematical model is a deterministic device in that for a given set of inputs, it predicts the output with absolute certainty, and it leaves nothing to stochastic part (Schabenberger and Pierce, 2001). A model is considered as a statistical model when it includes a deterministic/systematic part and a stochastic/random part. A statistical model, therefore, describes the presumed impact of explanatory variables and the probability distributions associated with aspects of the process that are assumed to be characterised by random variation (Stroup, 2012). In short, a statistical model comprises three components, i.e., systematic part, which consists of quantitative and/or qualitative explanatory variables, random part (refers to residual error term), and an assumed distribution.

A linear model usually refers to a classical linear model with Gaussian error. In matrix notation, this linear model is written as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

where $\mathbf{Y}$ is an $n \times 1$ vector of observations, $\mathbf{X}$ is an $n \times k$ incidence matrix for fixed effects factors, $\boldsymbol{\beta}$ is a $k \times 1$ vector of unknown fixed effect parameters to estimate, and $\mathbf{e}$ is a vector of residual errors and is assumed homoscedastic, uncorrelated, and following $N(0, \sigma^2\mathbf{I})$. In this case, the parameter estimates of $\boldsymbol{\beta}$ are solved using ordinary least squares (OLS), and the solutions are called best linear unbiased estimation (BLUE). Thus, in the classical linear model, there is only one type of effect in the systematic part that is considered, i.e. fixed effect. The matrix structure of variance for the classical linear model is $\mathbf{V} = \sigma^2\mathbf{I}$

Linear mixed models extend the classical linear models to allow both fixed and random effects factors in one model (Eisenhart, 1947; Harville, 1976; Laird and Ware, 1982). A matrix formulation of linear mixed models is as follows:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}$$

where $\mathbf{Y}$ is a vector $(n \times 1)$ of observations, $\mathbf{X}$ is the incidence matrix for fixed effects with $(n \times k)$ matrix, $\boldsymbol{\beta}$ is a vector of unknown fixed effect parameters to estimate with $(k \times 1)$ matrix, $\mathbf{Z}$ is the incidence matrix for random effects with $(n \times p)$ matrix, $\mathbf{u}$ is a vector of unknown random effect parameters to estimate with $(p \times 1)$. Since $\mathbf{u}$ consist of random effect parameters, $\mathbf{u}$ is assumed to be $N(0, \mathbf{G})$, where $\mathbf{G}$ is the variance-covariance (VCOV) matrix of all random effects. The vector $\mathbf{e}$ consists of residual errors. The assumption of residual errors are more relaxed in the linear mixed models than in the classical linear models since it allows non-independence and heterogeneity, $N(0, \mathbf{R})$, where $\mathbf{R}$ is the VCOV matrix for the residuals. Henderson (1950,1963,1975,1984) developed mixed model equations (MME) to obtain the solutions of fixed effects

($\boldsymbol{\beta}$) and random effects ($\mathbf{u}$) for animal breeding purpose. The Henderson's MME is as follows:

$$\begin{bmatrix} \mathbf{X}^{\mathrm{T}}\widetilde{\mathbf{R}}^{-1}\mathbf{X} & \mathbf{X}^{\mathrm{T}}\widetilde{\mathbf{R}}^{-1}\mathbf{Z} \\ \mathbf{Z}^{\mathrm{T}}\widetilde{\mathbf{R}}^{-1}\mathbf{X} & \mathbf{Z}^{\mathrm{T}}\widetilde{\mathbf{R}}^{-1}\mathbf{Z} + \mathbf{G}^{-1} \end{bmatrix} \begin{bmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^{\mathrm{T}}\widetilde{\mathbf{R}}^{-1}\boldsymbol{y} \\ \mathbf{Z}^{\mathrm{T}}\widetilde{\mathbf{R}}^{-1}\mathbf{y} \end{bmatrix}$$

The solutions to MME are the BLUE for $\boldsymbol{\beta}$ and the best linear unbiased prediction (BLUP) for $\mathbf{u}$. Unlike the classical linear model, linear mixed models have variances for random effects and the residual terms. Thus, the variance for linear mixed model is written as $\mathbf{V} = \mathbf{ZGZ}' + \mathbf{R}$. The $\mathbf{G}$ matrix is the variance-covariance (VCOV) matrix for the random effects and the $\mathbf{R}$ matrix is the VCOV matrix for the residual term. The various structures of VCOV will be described in the VCOV structures section.

1.5.1 Estimating fixed effects (BLUE) and predicting random effects (BLUP)

A fixed effect factor is **estimated** with BLUE. The term "Best" means that the sampling variance is minimised. "Linear" indicates that the **estimates** are linear functions of the observed values. "Unbiased" implies that the expected values of the **estimates** are equal to their true values $E[BLUE(\boldsymbol{\beta})] = \boldsymbol{\beta}$. The effect of a factor is considered as "fixed" if we are just interested in its particular value or, in general, if a factor only has a few levels and not coming from or representing from a probability distribution (McCulloch et al., 2008), and the conclusions apply only to the particular factor levels (Lynch and Walsh, 1998). For examples, effects of different soil types, or effects of different fertilizer. In plant breeding, an individual location or a trial is considered random but the set of locations/trials is considered as a fixed effect (Bernardo, 1996). In other words, if a zone or region consists of a number of locations, then the effect of a zone is considered as fixed and the effect of locations are considered as random because the condition of locations may change from year to year. Furthermore, the mean differences among different sets of environments are considered as nuisance factors that should be taken into account for genotypes comparisons (Bernardo, 2010). In the fixed effect, when the experiment is repeated, the effect of a factor will be the same, which means the true value of fixed effect does not change in each repetition of the experiment (Blasco, 2017). The estimation of fixed effects in the linear mixed models is slightly different from the estimation in the classical linear model. In the mixed models, the fixed effects estimates are solved using generalised least squares estimation (GLSE), not OLS.

A random effect is **predicted** with BLUP. The expansion of "B" is the same with "B" in BLUE. "L" indicates the **predictions** are linear functions of the observed values. "Unbiased" implies that the expected values of the **predictions**

are equal to their true values $E[BLUP(\mathbf{u})] = E(\mathbf{u})=0$. Hence, in random effect, we would like to have prediction instead of estimation. The term "prediction" is chosen by Henderson (1984) since in the animal breeding, the interest is to evaluate the potential of breeding value of a mating between two potential parents and to predict the future records. The term "estimation" is more appropriate to estimate the value if an animal already born. Thus, it has become common term in practice to "estimate" fixed effects and to "predict" random effects (Robinson, 1991). In the opposite of fixed effect, one of the assumptions underlying random effects is when the experiment is repeated, the true value will change/not be constant. This is also the reason that "prediction" term is used for random effect. The other assumptions are: the levels of a factor are of no particular interest and represents or comes from a probability distribution. Thus, in general, the levels of a random effect factor will have many levels to represent the whole population. While in the fixed effect, the parameter to estimate is the mean of individual levels, in the random effect, a variance (dispersion parameter) is the parameter to estimate. Therefore, the conclusion applies to a population.

There are some approaches to estimate variance components for random effects such as least squares/ANOVA, maximum likelihood (ML), and residual/restricted maximum likelihood (REML). The ANOVA approach is limited because it cannot construct complex VCOV structures, the data should be balanced, and it needs to make expected mean squares tables, which are not easy to construct. The preferred used method is REML (Patterson and Thompson, 1971). REML is more favourable than ML because variance estimation via ML fails to take into account for the loss of degrees of freedom needed for estimation, which results in downward bias in variance estimates. REML accounts for the degrees of freedom that are used to estimate fixed effects. Thus, it corrects the degrees of freedom. For a simple example, an unbiased sample variance estimate should have $n-1$ for the degrees of freedom. The estimate variance with ML produces n degrees of freedom, while REML produces $n-1$. This is the reason why REML is preferable to ML. Also, when the data is balanced, the REML variance estimates are equivalent to the ANOVA approach.

Bernardo (2010) mentioned several benefits of BLUP in the plant breeding framework. First, in the MET, the better genotypes will be tested in several years while the less superior genotypes will be discarded, which results in unbalanced data. BLUP allows analysing such unbalanced data while accounting for differences in the amount of data available for each genotype. Second, BLUP uses the information for all relatives measured to improve the prediction of breeding values. For example, when a breeder wants to compare two individuals,

A and B, the comparison can be made solely on the basis performance of A and B alone. By using BLUP, the comparison will be more precise by including the information of relatives of A and relatives of B. In the MET, this feature is very useful that using BLUP, we can borrow or recovery information of the same genotype in other environments, and so exploit the genetic correlation between environments (Kleinknecht et al., 2013; Piepho et al., 2016), which improves the prediction accuracy of genotype performance compared to BLUE.

1.5.2 Variance-covariance (VCOV) structures

In the MET analysis, the assumption of homogeneity variance is hardly ever fulfilled, because genotypic variances tend to change across environments. Furthermore, genotypic correlations for pairs of these environments are not homogeneous (Bustos-Korts et al., 2016). In the linear mixed models framework, applying variance-covariance (VCOV) structures can be applied on the random effect of GEI and the residual terms to take into account this heterogeneity, and so achieve higher prediction accuracies. The VCOV for GEI effect is applied in the $\mathbf{G}_{ge}$ matrix and the VCOV of residual term is applied in the $\mathbf{R}$ matrix**.** In this section, we describe four VCOV structures: identity, compound symmetry, unstructured, and factor analytic order 1.

Identity (ID)

The ID structure assumes independence and homoscedasticity, $\mathbf{G}_{ge} = \mathbf{I}(\sigma_g^2 + \sigma_{ge}^2)$, where σ_g^2 is variance of genotype and σ_{ge}^2 is variance of GEI. The matrix form is as follows:

$$\mathbf{G}_{ge} = \begin{bmatrix} \sigma_g^2 + \sigma_{ge}^2 & 0 & \dots & 0 \\ 0 & \sigma_g^2 + \sigma_{ge}^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_g^2 + \sigma_{ge}^2 \end{bmatrix}$$

The dimension of this diagonal matrix equals the number of genotype (g) times the number of environments. In this case, all the variances in all environments are the same.

Compound symmetry (CS)

The compound symmetric (CS) model implies that both variance and covariance are homogeneous. Thus, the structure of CS is as follows:

$$\mathbf{G}_{ge} = \begin{bmatrix} \sigma_g^2 + \sigma_{ge}^2 & \sigma_g^2 & \dots & \sigma_g^2 \\ \sigma_g^2 & \sigma_g^2 + \sigma_{ge}^2 & \dots & \sigma_g^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_g^2 & \sigma_g^2 & \dots & \sigma_g^2 + \sigma_{ge}^2 \end{bmatrix}$$

Each environment has the same variance and the genetic correlation is the same between all pairs of environments.

Unstructured (US)

The unstructured VCOV structure allows both heterogeneous covariance and variance. Thus, each environment has a unique genotype variance, and each pair of environments has a unique covariance. The number of parameters needed for this VCOV structure is $(p + 1)/2$, where p is the number of environments.

$$\mathbf{G}_{ge} = \begin{bmatrix} \sigma_{e_1}^2 & \sigma_{e_{12}}^2 & \dots & \sigma_{e_{1p}}^2 \\ \sigma_{e_{12}}^2 & \sigma_{e_2}^2 & \dots & \sigma_{e_{2p}}^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{e_{1p}}^2 & \sigma_{e_{2p}}^2 & \dots & \sigma_{e_p}^2 \end{bmatrix}$$

Factor analytic order 1 (FA1)

Factor analytic (FA) structures are often more useful than the US structure for taking into account heterogeneity in complex genotype×environment models. These structures have fewer parameters than the US structure (Isik et al., 2017). We here describe the FA structure with a single multiplicative term (FA1). In this structure, the $\mathbf{G}_{ge}$ is defined as $\mathbf{\Lambda\Lambda^T + \Psi}$, where $\mathbf{\Lambda}$ is a vector of dimension $1 \times p$ that consists of loading factors λ_1 to λ_p, and $\mathbf{\Psi}$ is a $p \times p$ diagonal matrix that consists of environment-specific genotype variances (ψ_e^2), $e = 1, 2, \dots, p$. For the FA1 model with g unrelated cultivars tested in p environments, we have:

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

Hence, the VCOV structure for $\mathbf{G}_{ge}$ is:

$$\mathbf{G}_{ge} = [\mathbf{\Lambda\Lambda^T + \Psi}] = \begin{bmatrix} \lambda_1^2 + \psi_1^2 & \lambda_1\lambda_2 & \dots & \lambda_1\lambda_p \\ \lambda_2\lambda_1 & \lambda_2^2 + \psi_2^2 & \dots & \lambda_2\lambda_p \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_p\lambda_1 & \lambda_p\lambda_2 & \dots & \lambda_p^2 + \psi_p^2 \end{bmatrix}$$

The off-diagonal elements of the blocks $\boldsymbol{\Lambda\Lambda^T} + \boldsymbol{\Psi}$ are products of parameters λ_p and $\lambda_{p'}$, which refers to the e-th and e'-th environment, respectively. Therefore, the nested effects between the same genotype in different environments are correlated, while the interaction effects from different genotypes are uncorrelated.

1.6 Cross-validation (CV)

Cross-validation (CV) is a method to evaluate the performance of statistical methods by estimating test error rate (James et al., 2013). A CV is conducted to evaluate a model's performance, which is known as model assessment, and to select a model that has a proper level of flexibility, which is known as model selection (James et al., 2013). In the CV, a dataset is split into a training set and a validation set. The training set is used to train the model while the validation set is used to validate the prediction produced by the model from the training set.

Two most-used methods to conduct CV are leave-one-out (LOO) and *k*-fold CV. The LOO CV leaves one data point as validation set. Thus, if there is a set of *n* data points, then there will be *n* iterations of fitting. For example, with 10 data points, 10 iterations are done because each time one data is left out as a validation set. A *k*-fold CV divides randomly a set of data points into *k* groups, or *folds*, in an equal size. The first fold is kept for validation and the model is trained on *k*-1 folds. The process is iterated *k* times and each time a different fold or a different group of data points are used for validation. Thus, a *k*-fold CV may require less iterations than LOO CV. Figure 3 depicts the LOO CV and *k*-fold CV.

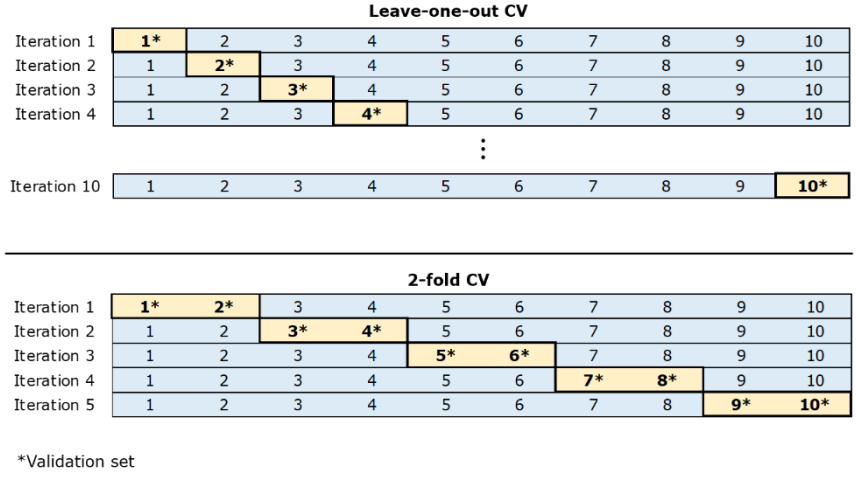

Figure 3. Examples for LOO CV and 2-fold CV for 10 data points.

The difference between the predictions from the training set and the validation set will be measured by using mean squared error (MSE). The smallest MSE of a model will be regarded as the best-performed model, and so may be selected because the best-performed model provides the highest prediction accuracy among the compared models.

1.7 Current statistical analysis in Swedish cultivar testing

The MET data structure is often large and highly imbalanced, and so causes computational problems when applying some statistical methods. Furthermore, the current statistical model needs to be assessed whether it produces accurate predictions of cultivar performance.

As already mentioned in Section 1.4, Swedish cultivar testing provides cultivar recommendation for zones, not for each trials/locations. The current statistical method analyses each zone separately to obtain the mean of each cultivar for each zone. Furthermore, the effect of cultivar in the model is fixed. The drawbacks of this statistical analysis are that the interaction of cultivar×zone is not exploited and the fixed effect of cultivar leads to the overoptimistic estimate of cultivar performance and inaccurate cultivar rankings. Patterson and Silvey (1980) mentioned that the estimate of newly recommended cultivars are, on the average, 27% too large. Thus, the cultivar performance information for the future is not accurate for farmers. For this reason, the current statistical analysis needs to be improved by developing a robust and efficient statistical method that provides more accurate predictions.

Another aspect that needs to be addressed is a stability measure of cultivar. Due to large and imbalanced data, the computation of stability measure is demanding and may lead to uninformative estimates. Thus, an efficient computation of measure of cultivar stability is also urgent to be tackled by developing a comprehensive and robust statistical method.

1.8 Aims of the thesis

This study aims to address the mentioned issues in section 1.7 and provide a gold standard for MET analysis in Swedish official cultivar testing. The overall aim of the thesis was to improve the statistical model for zone-based cultivar predictions and rankings by comparing the current-practice statistical procedures of Swedish cultivar testing with a new statistical method using BLUP through a CV study and propose a new inter-zone stability measure.

The specific aims were to:

- Investigate the performance of empirical BLUE and empirical BLUP for zone-based prediction in cultivar testing focused on the fungicide-treated subsets of the datasets (Paper I).
- Compare the performance of empirical BLUE and empirical BLUP for zone-based prediction in cultivar testing including complex variance-covariance structures in Swedish cultivar trials on all fungicide levels datasets (Paper II).
- Determine the necessity of division of agricultural zones/zonation (Papers I and II).
- Determine the best statistical analysis strategy for zone-based prediction cultivar testing, i.e., single-stage or two-stage analyses combined with complex VCOV structure focused on the fungicide-treated subsets of datasets (Paper III).
- Propose a new inter-zone stability measure based REML approach from the best model according to the CV study.

2 Materials and methods

2.1 Swedish cultivar trials datasets

The dry matter yield (DMY) of winter wheat (*Triticum aestivum* L.) and spring barley (*Hordeum vulgare* L.) from three zones, i.e., South, Middle, and North, of Swedish MET datasets were used. A detailed description of the number of locations in each zone is given in Papers II and III. The cultivar trials were laid out in a split-plot design with two replicates. The main-plot factor consisted of two levels of fungicide treatment (treated and untreated). Within each fungicide treatment, cultivars were arranged in α-designs (Patterson and Williams, 1976) with two replicates. The number of incomplete blocks varies between trials, depending on the number of cultivars.

2.2 EBLUE vs. EBLUP on fungicide-treated subsets datasets (Paper I)

The current analysis procedure of Swedish cultivar trials is done with an unweighted two-stage analysis. In the first stage, the experiment is analysed using a linear mixed model with cultivars, fungicide treatments, and cultivar×fungicide treatment interactions as fixed effects. The effects of replicates and incomplete blocks are modelled as random. 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:

$$C + F + C{\cdot}F : R + R{\cdot}F + R{\cdot}B \qquad (1)$$

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 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.

In the second stage, since the current analysis strategy is unweighted, only the adjusted cultivar means from the first stage are forwarded to the second stage. Thus, there was only a single value of DMY per cultivar and trial, and no residual plot error information was available in the second stage. By using mixed models, the adjusted cultivar means were used to obtain the final estimates of cultivar yields for each zone. The model used in the second stage is as follows:

$$C : L + C \cdot L \tag{2}$$

where L is the location/trial, which is always nested within zone/region (Z). In the current-practice, the model in Eq. 2 is fitted separately for each level of fungicide and zone. Nevertheless, since the cultivar and zone effects are fixed, we can rewrite equation 2 as:

$$C + Z + C \cdot Z : L + C \cdot L \tag{3}$$

Equation 3 is equal to equation 2 when the equation 3 uses zone-specific residual variance. The cultivar and cultivar×zone interaction effects are assigned to be fixed. Thus, it will be estimated by empirical BLUE (EBLUE), and so it is not possible to borrow information across zones. The term "empirical" indicates the actual variance components are not known and therefore are estimated from the data.

In this study, we would like to improve the prediction accuracy of yield of each genotype for each zone in the second stage. For that reason, the cultivar and cultivar×zone interaction effects were assigned to be random. Hence, the equation 2 is changed as follows:

$$Z : C + L + C \cdot L + C \cdot Z \tag{4}$$

where Z is the zone. In equation 3, the zone and cultivar×zone interaction terms are included and so accommodates all zones to be analysed in a single model and explore the interaction effect of cultivar×zone. Moreover, the cultivar and cultivar×zone interaction effects are random, which are predicted by empirical BLUP (EBLUP), and so allowing EBLUPs for a specific zone to borrow information from the other zones (Atlin et al., 2000; Kleinknecht et al., 2013; Piepho et al., 2016). Note that when there is an interaction between fixed and random effects of a factor, its effect will be considered as random. The motivation to use EBLUP has been addressed by Smith et al. (2001) because of the "deficiency in the traditional fixed cultivar-effects approach in terms of obtaining reliable predictions of future yield performance." This deficiency has

been 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." Thus, in current practice, the estimation of cultivar's yield may be too optimistic and the ranking of cultivars may be not accurate since the cultivar effect is fixed.

In this study, we searched the best model for the single-year and the multi-year series analyses via a CV study focused on improving the model in the second stage, which will be explained in section 2.5.1. The multi-year analysis refers to five-year series analysis. For this work, we started with a simple case data and focused on the fungicide-treated subsets of the datasets so the models were simple without any complex interaction terms. Five models were proposed for the single-year analysis. The five models consisted of four models for EBLUP (single-year random effect of cultivar models, SYR) and one model for EBLUE (single-year fixed effect of cultivar models, SYF). Zone-specific residual variance structures were employed in some EBLUP models to account for heterogeneity among zones. An EBLUP model without the zone term was also included in the study to assess the necessity of zonation.

For the five-year series analysis, four models, consisted of two models for EBLUP (multi-year random effect of cultivar models, MYR) and two models for EBLUE (multi-year random effect of cultivar models, MYF), were proposed. However, in the five-years series, heterogeneous residual variance was not used due to convergence problem. The EBLUE models compared in this study were models used in the current practice. The details of the models are given in Paper I. However, note that the notations in Paper I were slightly different. In the Paper I, the cultivar was coded with V because it referred to variety and the zone was coded with R because it referred to region. Thus, in principle, the variety is interchangeably to cultivar, and region is interchangeably to zone. Nonetheless, for the whole thesis, we will use C for cultivar, Z for zone, and R for replicate.

2.3 EBLUE vs. EBLUP on all fungicide levels datasets (Paper II)

In this study, we extended the model from the first study by including the fungicide factor in the second stage. Therefore, the baseline model, which is used in the current routine analysis in Swedish cultivar testing, can be written as

$$C + Z + F + C{\cdot}Z + C{\cdot}F + Z{\cdot}F + C{\cdot}Z{\cdot}F : L + C{\cdot}L \qquad (5)$$

Again, we would like to assign the effects of cultivar to be random, and so the effects of the cultivar×zone term will be random to allow borrowing information between zones. A total of 20 linear mixed models were compared for the single-

year series. 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. Since the cultivar×zone interaction term was random, several VCOV structures were applied to account for heterogeneity of this term. Zone-specific heterogeneous residual variance structures were also employed in some models to account for heterogeneity among zones. To facilitate readability, the 20 models were categorised into five groups. There were 17 SR models and 3 SF models. The details of the models and the VCOV structures are given in Table 1, Paper II.

In the five-year series, an additional factor to be included in the model is year (Y). Thus, the baseline model for the current practice, in the second stage is

$$C : L + Y + C \cdot Y \tag{6}$$

which is fitted separately for each fungicide and zone. A total of 11 models were compared for the multi-year series. One multi-year series model (M) with fixed (F) effects of cultivar is called the MF model and is the model used in current practice. The other 10 models with random (R) effects of cultivar are called MR models. The details of all the models are given in Table 2, Paper II. The MR 1 model is a basic saturated model. The next three MR models (MR 2–4) were obtained by dropping, one at a time, the single term with the smallest variance. From the model MR 5, the year×fungicide and year×zone×fungicide interactions were dropped. In models MR 6 to 8, either the cultivar×zone×fungicide interaction or the cultivar×zone×year interaction, or both these interactions, were removed. Models MR 9 and 10 are models without effects of zones, which were compared to determine whether zonation is needed or not. Due to convergence problem, no heterogeneous residual variance was used in any five-year series models.

2.4 Single-stage versus two-stage analysis for zone-based prediction (Paper III)

The multienvironment trials data can be analysed by a single-stage analysis or stage-wise analysis (two stages or more). A single-stage analysis is considered as the gold standard (Gogel et al., 2018), while the two-stage analysis will have similar results when the full VCOV matrix of the estimated cultivar means from the first stage is forwarded to the second stage (Damesa et al., 2017). Nevertheless, in practice, storing full VCOV is hard to do, and so a diagonal approximation is often used.

A single-stage analysis has an advantage from theoretical consideration since the estimation of fixed and random effects are done in a single model from plot-

level data (Piepho et al., 2012a). Nevertheless, the most common disadvantage is the computational resource, especially when the number of cultivars and environments are large and a complex variance-covariance (VCOV) structure for the cultivar×environment interaction effects is assumed (Möhring and Piepho, 2009; Welham et al., 2010).

The computational burden in the single-stage analysis motivates a stage-wise analysis that splits the analysis into two (or more) stages. Damesa et al. (2017) and Piepho et al. (2012a) reported that the stage-wise analysis was possible to substantially reduce the computational burden. In the stage-wise analysis, each trial is analysed separately using BLUE, in the first stage, to obtain adjusted cultivar means per trial. Thus, the cultivar effects are modelled as fixed. In the second stage, the adjusted cultivar means from the first stage are analysed jointly, using an appropriate mixed model, in order to compute marginal means for cultivars across environments. In this stage, the cultivar effects may be modelled as fixed or random. Piepho and Eckl (2014) mentioned another advantage of stage-wise analyses for practical analyses: it facilitates a combined analysis of different trials with different experimental designs in the first stage, and subsequently allows modelling structures for heterogeneity of variance between trials easily.

A major issue of stage-wise analysis is the choice of method to forward the information on precision (standard errors, VCOV matrix of the adjusted means) between stages to account for heteroscedasticity as well as for covariances among the adjusted means (Damesa et al., 2017; Möhring and Piepho, 2009). A general scheme of the single-stage and two-stage analysis is depicted in Figure 4.

The simulation from Möhring and Piepho (2009) indicated that weighting can improve efficiency, but the unweighted method was acceptable if the assumptions of the model were correct, i.e., when error variances are independent of the genoptype×environment interaction structure. Also, they mentioned that the performance of the method 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 due to the loss of information in estimating the estimates of cultivar performance, both overall and within environments. However, similar to Gogel et al. (2018), Welham et al. (2010) focused on prediction for individual sites, whereas the focus of our study is on means across a wider region or zone, or in other words, zone-based prediction.

In this study, we assessed the current-practice strategy in Sweden, i.e., a two-stage unweighted, with several strategies, i.e., a single-stage with heterogeneous location-specific and homogeneous residual variance , two-stage weighted with

fully-efficient weighting (forward the full VCOV from stage 1 to stage 2), and two-stage weighting with diagonal approximation, via CV study and correlation coefficient. The diagonal approximation used for weighting were Smith's weighting (Smith et al., 2001) and the average standard error of differences (AVSED) weighting (Möhring and Piepho, 2009). We also added several VCOV structures for the cultivar×zone interaction term such as compound symmetry (CS), FA, and unstructured (US).

The combination of the approaches, a single-stage or a two-stage (weighted and unweighted), weighting methods in the two-stage weighting approach, and the VCOV structures, resulted in 21 strategies to be compared. The details of the strategies are given in Table 2 (Paper III). Moreover, unlike the first two studies, in this study, we used the datasets in the plot levels of single-year datasets. Thus, the CV was conducted in a single-year series dataset. The details of the datasets used in this study are shown in Figure 1, Paper III. To our knowledge, this study is the first using cross-validation for comparing single-stage analyses with stage-wise analyses.

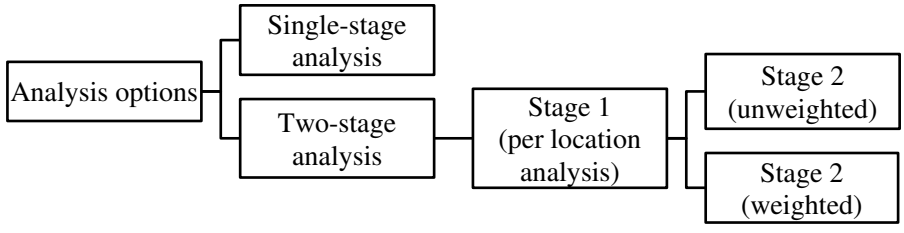

Figure 4. Scheme of the single-stage and two-stage analyses

2.5 Cross-validation study (Papers I, II, and III)

We preferred to conduct a CV study since the main objective was to select a model that provides the most accurate prediction. An information criterion like the Akaike information criterion (AIC) for model selection can be used, but it does not examine the accuracy of model prediction. Thus, a CV study is preferable to examine whether a model can produce an accurate prediction or not, and so gives a measure (MSEP) of the size of the prediction errors.

2.5.1 CV for Papers I and II

Single-year series CV

In Papers I and II, the CV study was carried out using the adjusted means of cultivars from the first stage. A 2-fold CV was used for model evaluation. In the first fold, the locations/trials were randomised 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 as the validation dataset. The reason for conducting this type of CV was the decreasing number of trials in recent years. Thus, the aim was to train the model with a small number of trials. If the CV were 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, a 2-fold CV was preferred. The illustration of the 2-fold CV is given in Figure 5.

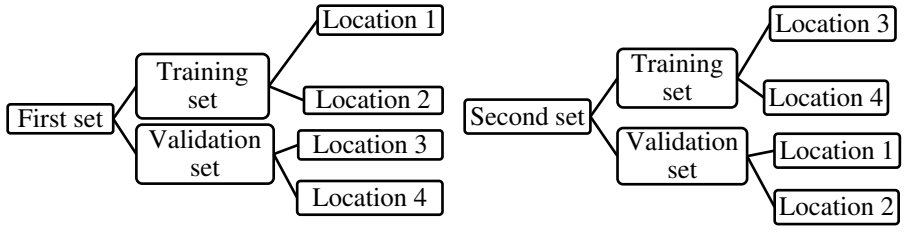

Figure 5. Illustration of single-year CV scheme.

In general, cultivar trials aim at predicting differences between tested cultivars rather than each cultivars' mean. Piepho (1998a) proposed the mean squared error of prediction (MSEP) to assess the accuracy of estimates of differences between cultivars in different 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 $(z_{vkt} - z_{v'kt})$:

$$MSEP = \frac{\sum_{t=1}^{T}\sum_{k=1}^{K}\sum_{v=1}^{V}\sum_{v'\neq i}^{V'}[y_{vkt} - y_{v'kt} - (z_{vkt} - z_{v'kt})]^2}{TKV(V-1)} \qquad (7)$$

where y_{vkt} and z_{vkt} is the observed yield and the predicted yield, respectively, of the vth cultivar in the t-th trials of j-th zones, using the k-th fungicide treatment, T is the number of trials of all zones, $\sum_{j=1}^{J} T_j$, K is the number of fungicide levels, and V is the number of cultivars.

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). In Paper I, the fungicide term was omitted because the study focused on the fungicide treated level only. The CV study was performed in SAS (SAS Institute, 2013) using PROC MIXED for the FA models and the models with the heterogeneous residuals. The PROC HPMIXED was used for the other models and to reduce the computational time.

Multi-year series CV

For the multi-year series CV, we modified a leave-one-out CV to mimic the current Swedish practice of predicting cultivar performance based on results from five years. A set of data from five consecutive years was used as a training set. Then, the following sixth year was used as a validation set as depicted in Figure 6. 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.

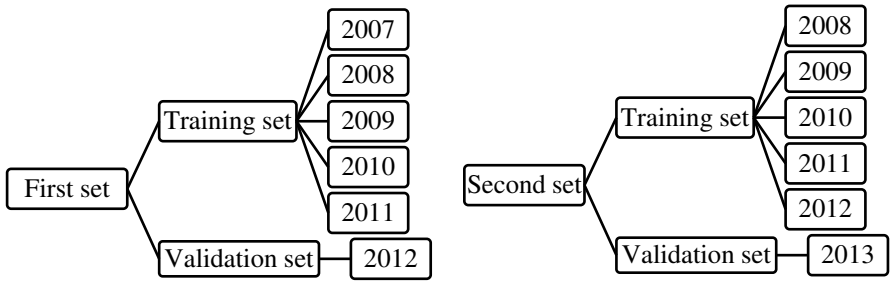

Figure 6. Illustration of multi-year CV scheme.

The CV was done in chronological order, besides to mimic the current-practice, due to the set of cultivars in the early years and recent years differ a lot. For example, when the training set consists of recent years and the validation set consists of early years, then there will be only very few cultivars in common between both sets. Consequently, most of the cultivars that are predicted in the training set would not be available in the validation set because 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.

We computed the MSEP as given in the Eq. 7. Again, the best model was the one that had 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. The multi-year series CV was performed in PROC HPMIXED in SAS (SAS Institute, 2013).

2.5.2 CV for Paper III

Most other studies comparing single-stage and stage-wise analyses used Pearson's moment-product correlation or Spearman's rank correlation between the cultivar estimates between those two analyses (Cullis et al., 2000; Damesa et al., 2017; Gogel et al., 2018; Piepho et al., 2012a). The consequence of using these correlations was the correlation coefficient estimates often are around 0.90, implying that the single-stage and stage-wise analyses provide similar results. In comparison to Pearson correlation, a CV study can measure the prediction errors of the model using MSEP, which is more desirable for choosing the model to predict cultivar performance in MET analysis.

In this study, a leave-one-out CV was performed for comparison and selection. One location was left out as a validation set and used the remaining locations as a training set. For example, when there are 10 trials in a single-year dataset, then there will be 10 folds of CV. We accumulated the discrepancies between the observed and predicted pairwise differences from the 10 folds of CV. Then, we computed the MSEP from this accumulation, and so there will be a single value of MSEP from the 10 folds of CV.

Recall from Eq.7, the MSEP in Eq.8 is similar to the MSEP proposed by Piepho (1998a), for measuring the prediction accuracy of the models for each single-year dataset. The MSEP is a standard statistic for assessing predictive accuracy as pointed out by Wallach and Goffinet (1989). Let y and z denote the observed and predicted values, respectively, I is the total number of cultivars, and J is the total number of locations. Thus, the assessment was measured based on the discrepancies between observed $(y_{cl} - y_{c'l})$ and predicted pairwise differences $(z_{cl} - z_{c'l})$:

$$MSEP = \frac{\sum_{j=1}^{J}\sum_{i=1}^{I}\sum_{i'\neq i}^{I'}\left[y_{ij} - y_{i'j} - \left(z_{ij} - z_{i'j}\right)\right]^2}{JI(I-1)} \quad (8)$$

The model that produced the smallest MSEP is considered the best model, since it predicted yield differences in the validation set most accurately, which provides the most accurate predictions per agricultural zone as a prediction for the locations within zones. The models were ranked based on the average MSEP over the five CV sets, since there were five single-year datasets. The CV study was conducted in R (R Core Team, 2018), and fitted all the models in ASReml-R 4.1.0.106 (Butler et al., 2017) to reduce computation time, and the ggplot2

package (Wickham, 2009) was used to produce plots via RStudio (RStudio Team, 2016).

2.6 A new inter-zone stability measure

The proposed inter-zone stability measure is a descriptive statistic obtained by computing a standard deviation for every single cultivar, and this standard deviation is computed of any interaction terms that comprise both cultivar and zone. The steps to calculate inter-zone stability are as follows:

1) Compute the EBLUPs of the cultivar×environment interactions, for each cultivar and in each environment.
2) Sum any EBLUPs of random effects that comprise cultivar×environment interaction effects, e.g., cultivar×zone (C·Z), cultivar×zone×fungicide (C·Z·F), and cultivar×zone×year (C·Z·Y), for each cultivar, in each environment.

   Let $\hat{a}_{ij}$ be the EBLUPs of the C·Z effect, for the ith cultivar and jth zone, $\hat{b}_{ijk}$ be the EBLUPs of the C·Z·F effect for the ith cultivar, jth zone and kth level of fungicide, and $\hat{c}_{ijkm}$ be the EBLUPs of the C·Z·Y effect for the ith cultivar, jth zone, kth level of fungicide, mth level of year. Then, the summation for the effects of C·Z, C·Z·F, and C·Z·Y for each cultivar in each zone for each fungicide treatment is:

   $$v_{ijkm} = \hat{a}_{ij} + \hat{b}_{ijk} + \hat{c}_{ijkm} \quad (9)$$

   Since v_{ijk} is the summation of two or more EBLUPs of random-effects, v_{ijk} will only be zero if $\hat{a}_{ij}$, $\hat{b}_{ijk}$, and $\hat{c}_{ijkm}$ are zero.
3) Finally, for the ith cultivar and kth level of fungicide, the inter-zone stability measure (s_{ik}) is computed as:

   $$s_{ik} = \sqrt{\frac{\sum_{j=1}^{J}\sum_{m=1}^{M}\left(v_{ijkm} - \bar{v}_{ijkm}\right)^2}{\mathrm{JM} - 1}} \quad (10)$$

where J is the number of zones and M is the number of years. As the example of the proposed inter-zone stability measure, we used the best multi-year model according to the MSEP as the fitted model for the stability measure.

3 Results and discussion

3.1 Cross-validation of statistical models on fungicide-treated subsets datasets (Paper I)

The aim of this study was to compare the EBLUP and EBLUE in simple models (only in fungicide-treated datasets). The MSEPs of the CV for the single-year series and multi-year series for both crops are listed in Tables 1 and 2, respectively. Clearly, in the single-year series, the EBLUE model (current-practice), performed comparatively poorly for winter wheat and spring barley datasets. On the other hand, as it was expected, the EBLUP models performed the best, i.e., SYR 1 with heterogeneous residuals.

In the multi-year series, again, the EBLUP model, i.e., the MYR 1 performed the best. Nonetheless, unlike the SYR 1, which employed heterogeneous residuals variance structure, the MYR 1 was, due to convergence issue, not computationally feasible to have a heterogeneous residual variance structure due to convergence issue. For that reason, we did not employ such structure.

In practice, the empirical datasets hardly satisfied the assumption of normality. However, EBLUP per se does not require normality (Searle et al., 1992), and the CV revealed that the EBLUP performed better than the EBLUE. We demonstrated the potential of borrowing strength across regions from random effects of cultivar×zone interaction, thereby increasing the accuracy of zone-based yield prediction.

In conclusion, the routine model used in the current-practice should be discontinued. The model SYR 1, i.e., the model Z : C + L + C·R, with heterogeneous residuals, should be the replacement model for the single-year series. For the multi-year series, the model MYR 1, i.e., the model Z : C + L + Y + C·Z + C·Y + Y·Z + C·Z·Y, is recommended.

Table 1. *Mean of MSEP for single-year series of winter wheat* ($N = 8$) *and spring barley* ($N = 5$).

Ranking	Model	Winter wheat	Spring Barley
		Mean	Mean
		g^2m^{-4}	g^2m^{-4}
1	SYR 1	6781	1751
2	SYR 2	6846	1766
3	SYF 1	7093	1783
4	SYR 3	7245	1814
5	SYF 2* (current method)	7407	1959

Table 2. *Mean of MSEP for multi-year series winter wheat* ($N = 6$) *and spring barley* ($N = 6$).

Ranking	Model	Winter wheat	Spring Barley
		Mean	Mean
		g^2m^{-4}	g^2m^{-4}
1	MYR 1	854685	276814
2	MYR 2	859878	278789
3	MYF 2* (current method)	938231	307994
4	MYF 1	1940205	611592

3.2 Cross-validation of statistical models on all fungicide levels datasets (Paper II)

3.2.1 Single-year series

We present parts of the list of MSEP average of winter wheat and spring barley for the single-year datasets in Table 3. The full list is given in the Paper II (Table 3). Our study shows that random-cultivar-effects model is preferable for routine zone-based yield prediction compared to fixed-cultivar-effects models. The EBLUP (SR) models achieved lower MSEPs than the EBLUP (SF) models for both crop datasets. For both crops, the SR 5 model performed the best. The current-practice model (SF 3) performed poorly for both crop datasets, and was the least performing among the SF models in spring barley 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, while in spring barley it was ranked the 16th best model. A plausible biological reason is that winter wheat is grown in winter weather conditions with large local variation, as compared to spring barley, which is sown in the springtime. Thus, spring barley is grown under less

diverse local conditions. In the winter time, the environmental conditions vary locally, from mild and humid to cold and dry, causing different stress factors to predominate (Olsen et al., 2018).

The BLUP with a complex VCOV structure, the 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 less well performing than the more parsimonious model SR 5. Therefore, the SR 5 model can be recommended for both crops.

Table 3. *Mean of MSEP from single-year CV of winter wheat* ($N = 8$) *and spring barley* ($N = 5$)

Ranking	Winter wheat		Spring barley	
	Model	**Mean**	**Model**	**Mean**
		g^2m^{-4}		g^2m^{-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
⋮	⋮	⋮	⋮	⋮
18	SF 3*	7313	SR 12	1911
19	SR 13	7826	SR 1	1914
20	SR 15	8488	SF 3*	2053

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

The investigated FA covariance structure allows heterogeneous variances and unique pairwise correlations between zones. Moreover, the FA structure is useful because it allows heterogeneous variance and covariance using fewer parameters than the unstructured covariance structure. Nonetheless, the REML 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 factor-analytic structure may be more useful when the number of zones is larger than three. Also, it is shown that the model with many interaction terms were less-performing compared to the parsimonious one.

3.2.2 Multi-year series

The MSEP means of multi-year series in winter wheat and spring barley are listed in Table 4. Again, the currently used MF model was the most unfavourable model since for both crops this model showed the largest average MSEP. The best EBLUP model was different in both crops.

In the 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. For spring barley, the MR 7 model was the best, while it performed less well in the winter wheat. The MR 5 model, which was 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 spring barley five-year series, while in winter wheat MR 10 was still ranked among the five best models, as also shown in the single-year series.

Table 4. *Mean of MSEP from multi-years CV of winter wheat* ($N = 6$) *and spring barley* ($N = 6$)

Ranking	Winter wheat		Spring barley	
	Model	**Mean**	**Model**	**Mean**
		g^2m^{-4}		g^2m^{-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
⋮	⋮	⋮	⋮	⋮
11	MF*	8596	MF*	2320

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

In the multi-year series, the prediction accuracy was not improved with the higher-order-interaction effects compared to models that are more parsimonious and straightforward to fit, as it also occurred in the single-year analysis. Employing the heterogeneous covariance structure for residual effects in the multi-year analyses may be useful in order to have variance components differing between years or zones. However, based on the single-year series CV, the model with heterogeneous variance in the **R** matrix (SR 4) did not perform well as compared to the model with homogeneous residual variance.

Furthermore, the computation time will be increased and a convergence issue may occur when applying a heterogeneous residual variance structure. A higher number of interaction terms or a more complex variance-covariance structure may cause overfitting that may decrease the accuracy of predictions. As in the

single year series, in the multi-year series, it is reasonable to choose the parsimonious models since the MSEP of these models outperformed the others. Also, the computation time will be less compared to the complex models. In the multi-year 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.

3.2.3 Use BLUP instead of BLUE – Yes, but with some notes

As we already mentioned in the first study, the empirical datasets that we used here were not perfectly normally distributed, which is showed by the residual diagnostics in the Figure S1 and S2 in the Supplemental Materials of Paper II. 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).

In practice, the variance components are unknown and must be estimated. REML estimates may be imprecise in small datasets, which makes the benefits of using random-cultivar-effects models is 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 EBLUP in small randomised complete block experiments.

We recommend striving for complete datasets for the single-year analysis. Forkman (2013) showed that analyses of incomplete datasets using generalised least squares (GLS) based on mixed models with random environmental effects can give unexpected estimates. 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 less well in that zone. In this case, the cultivars are not missing at random (MAR). If there is a 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., 2012b).

Regarding the missing data, in the multi-year 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 two years. We recommend that all cultivars should be retained in the analysis. The reason is that all cultivars involved in selection decisions should be included in the analysis to avoid selection bias, as pointed out by Piepho and Möhring (2006). Piepho and Möhring (2006) also mentioned that removal of data leads to a missing-not-at-random (MNAR) pattern that causes invalid variance component estimates. Besides, if missing data pattern is

MNAR, then EBLUP will systematically be associated with varying degree of shrinkage, which causes bias. For example, if a cultivar is very little tested, then the shrinkage of all its predicted effects will be large, and so the prediction will be less accurate.

3.2.4 Application of the best model in winter wheat datasets

We present the application of the SR 5 model in the winter wheat dataset 2016 and the MR 5 model in the winter wheat dataset 2012–2016. The ANOVA table for the fixed effects significance tests and the tables of variance components for the winter wheat datasets are given in Tables 5 and 6 in Paper II.

Table 5 presents an example of different cultivar ranking between EBLUE 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 EBLUE in some cultivars, e.g., Etana, G 0512LT3, and Brons. The smaller values using EBLUP were a consequence of "shrinkage". BLUP is a shrinkage method since information about the distribution is used, in essence, to "shrink" the effects towards zero (Galwey, 2014; Stroup, 2012). 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 to fixed effects estimation (Robinson, 1991). Means higher than the overall mean are shrunk downwards to the overall mean, as also can be seen for some cultivars such as Festival and Rivero, which were not listed among the best 10 cultivars by the EBLUP method. Therefore, the shrinkage property avoids otherwise over-optimistic estimates of cultivar performance. On the other hand, the means that are lower than the overall mean are slightly increased (shrunken upwards towards the overall mean) using SR models. In this case, 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 listed among the best 10 cultivars by the E-BLUP method.

The ranking of the cultivars is different between the EBLUE and EBLUP methods. The best cultivar according to the EBLUE method was cultivar G 0512LT3, while using the EBLUP 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 EBLUP method should be preferred due to its smaller MSEP.

Table 5. *Example of different cultivar ranking in the winter wheat 2016 from Zone A, fungicide-treated. More than half of the cultivars differed in ranking.*

Cultivar	EBLUE (SF 3)		EBLUP (SR 5)	
	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

Table 6 presents the example of different cultivar ranking between EBLUE with the MF model and EBLUP with MR 5 model in the winter wheat dataset 2012–2016. Again, we can see a considerable shrinkage in the DMY predictions using EBLUP in some cultivars, e.g., G0512LT, Lw 06W607-10, RGT Universe, and Torp. Also, the ten top-performing cultivars also differed a lot between the EBLUE and EBLUP methods. For example, G0512LT was the best cultivar according to EBLUP, while RGT Universe was the best cultivar with EBLUE.

Therefore, the multi-year example also clearly shows that the ranking between EBLUE and EBLUP differed a lot and that EBLUP provided more accurate ranking due to the shrinkage, as indicated by the lowest MSEP in the MR 5 model.

The best variety by EBLUE ranked 6 by the EBLUP method. Some varieties that were not listed among the best 10 cultivars in the EBLUE model were listed among the best 10 cultivars by the EBLUP 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 6. *Example of different winter wheat cultivar ranking in the multi-year analysis (2012–2016) from South Zone, fungicide-treated. More than half of the cultivars differed in ranking.*

Cultivar	EBLUE (MF)		EBLUP (MR 5)	
	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	-	-

3.3 Cross-validation of single-stage versus two-stage analysis (Paper III)

The first two studies confirmed that the EBLUP outperformed EBLUE. However, those studies only focused on the second stage of the analysis. The current-practice strategy is the two-stage unweighted strategy, which can be improved by single-stage analysis or two-stage weighted analysis, as mentioned by Möhring and Piepho (2009) and Welham et al. (2010). However, it should be noted that in this study the main objective is to obtain prediction accuracy for zone-based prediction, while Welham et al. (2010) aimed for individual trials. The zone-based prediction is also more favourable for breeders since breeders want to provide a cultivar that performs well in large TPE.

In this study, based on the MSEP averages in Table 7 (partly shown), the single-stage and the two-stage weighted outperformed the current-practice method. For both crops, the single-stage with identity random effects VCOV structures and location-specific residual variance structure (1S-ID-LR) performed the best, since this approach had the lowest average MSEP. However, the differences between 1S-ID-LR and the three weighted two-stage analyses (2S-ID-W-FE, 2S-ID-W-AVSED, and 2S-ID-W-S) were minor for both crops. Thus, these four analyses performed very similary.

On the other hand, the current practice analysis (2S-F-U-ZR) was the least performing analysis for winter wheat and the second least for spring barley.

Furthermore, none of the weighting methods improved the fixed-C·Z effects strategies (2S-F-AVSED and 2S-F-S) compared to the current approach.

Similar to the first two studies, the complex VCOV structures for C·Z did not improve the predictive model performance. The FA1 structure was far less performant than the ID structure for single-stage analyses as well as two-stage analyses. No model with FA structure was among the top-five performing models. In spring barley, the 1S-FA1-ID performed least well, even worse than the current practice. In general, the MSEP of US VCOV was similar to the MSEP of FA1. The exception was US in the single-stage analysis with heterogeneous residual location-specific variance (1S-US-LR), which for both crops, showed the fifth best average MSEP.

The simple two-stage unweigthed analysis, 2S-ID-U-ID, performed better than the 1S-AID in both crops. Thus, this result revealed that the simple EBLUP two-stage unweighted analysis produced better predictions than the far too simple single-stage EBLUP analysis so using adjusted means from stage 1 were more accurate than using a single-stage analysis that is neglecting heterogeneity in replicates and incomplete blocks across locations.

Table 7. *Mean of MSEP of winter wheat* ($N = 5$) *and spring barley* ($N = 5$)

Ranking	Winter wheat		Spring barley	
	Strategy	Mean	Strategy	Mean
		g^2m^{-4}		g^2m^{-4}
1	1S-ID-LR	5041	1S-ID-LR	1723
2	2S-ID-W-FE	5045	2S-ID-W-FE	1726
3	2S-ID-W-AVSED	5049	2S-ID-W-S	1727
4	2S-ID-W-S	5051	2S-ID-W-AVSED	1728
5	1S-US-LR	5057	1S-US-LR	1728
⋮	⋮	⋮	⋮	⋮
20	2S-F-S	5334	2S-F-U-ZR†	1850
21	2S-F-U-ZR†	5389	1S-FA1-ID	1870

†The current-practice analysis in Swedish cultivar testing.

3.3.1 Application of the best strategies as comparison to the current-practice strategy in winter wheat 2016 and spring barley 2015 datasets

Figure 7 shows the zone-pairwise scatter plot of each cultivar predictions (EBLUP) and estimates (EBLUE) of C·Z effect for each model and crop. Figure 7A depicts the predictions/estimates of C·Z effect of the same cultivar from South and North zones. Figure 7B presents the predictions/estimates of C·Z effect of the same cultivar from Middle and North zones, and Figure 7C shows the prediction/estimates of C·Z effect of the same cultivar from South and

Middle zones. In general, it can be seen that the EBLUP methods (1S-ID-LR, 2S-ID-W-FE, 2S-ID-W-AVSED, 2S-ID-W-S) have narrower ellipses than the EBLUE method (2S-F-U-ZR). Thus, in the EBLUP method, the cultivar predictions and rankings between each two zones were more similar than in the EBLUE method because the assumption of random effect (EBLUP) cultivar exploited the genetic correlation between zones.

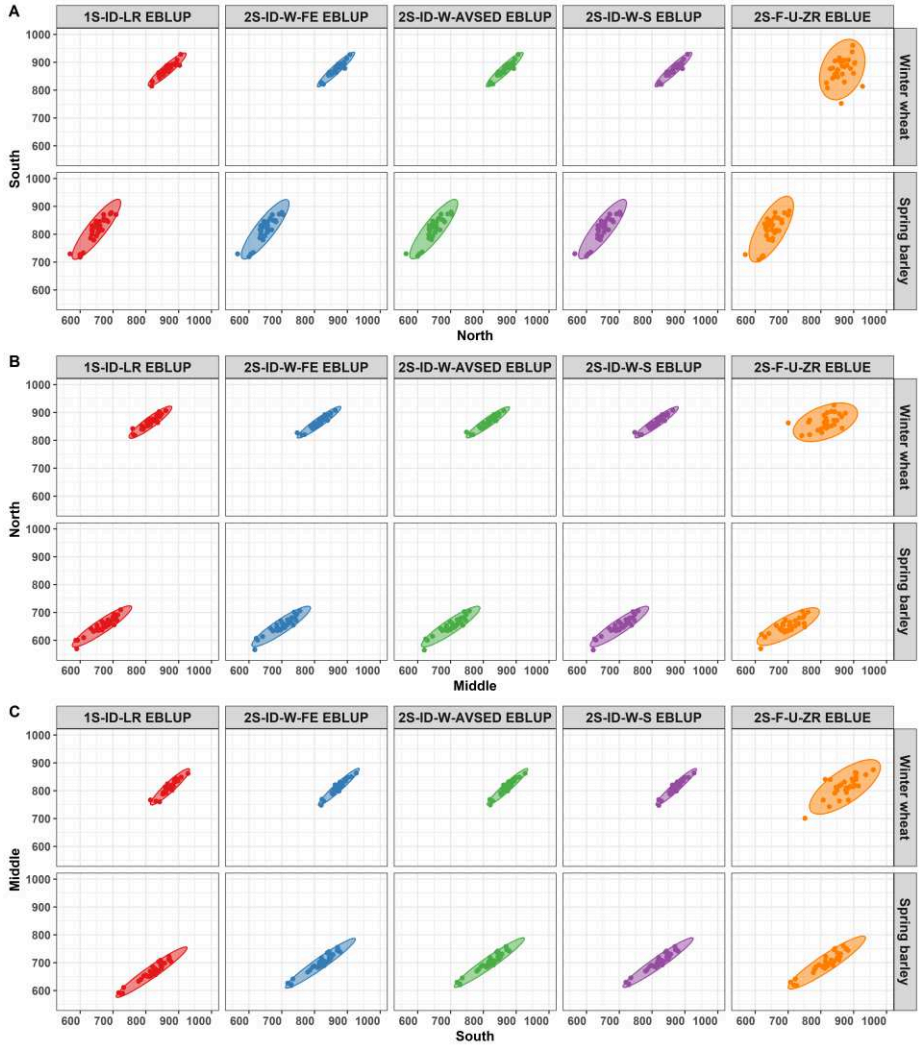

Figure 7. Zone-pairwise scatter plot of cultivar estimates of cultivar×zone (C·Z) interaction effects for four models with the smallest MSEP and current-practice model (2S-F-U-ZR EBLUE) in each cultivar. (A) Estimates cultivar×zone between North and South. (B) Estimates cultivar×zone between Middle and North. (C) Estimates cultivar×zone between South and Middle. The genetic correlation between zones is exploited in the EBLUP method compared to the EBLUE method.

The genetic correlation between zones for winter wheat dataset was 0.81 and for spring barley dataset was 0.84. Since these correlations are quite high, based on this study, the cultivar means between two zones are not very different. The genetic correlation cannot be obtained with the EBLUE method.

3.3.2 MSEP is preferable compared to correlation coefficient

Tables 8 and 9 present Pearson's product-moment correlation and Spearman's rank correlation of all adjusted cultivar predictions and estimates, using the four top-performing strategies (1S-ID-LR, 2S-ID-W-FE, 2S-ID-W-AVSED, 2S-ID-W-S) and current-practice strategy (2S-F-U-ZR), for the winter wheat and spring barley datasets, respectively. For winter wheat, both Pearson and Spearman correlations were high among the four strategies with EBLUP but were relatively low between these four strategies and the 2S-F-U-ZR. The correlations between the two-stage analyses were close to one. For spring barley, the correlations among these four strategies were relatively higher than in winter wheat, even the correlations between the strategies with EBLUP and the 2S-F-U-Z were also high.

Table 8. *Correlation among adjusted cultivar estimates of winter wheat 2016 dataset (above the diagonal: Pearson's product-moment correlation; below the diagonal: Spearman's rank correlation).*

	Strategy*				
Approach	**(1) 1S-ID-LR**	**(2) 2S-ID-W-FE**	**(3) 2S-ID-W-AVSED**	**(4) 2S-ID-W-S**	**(5) 2S-F-U-ZR**
	EBLUP	**EBLUP**	**EBLUP**	**EBLUP**	**EBLUE**
(1)	1.0000	0.9894	0.9904	0.9895	0.8986
(2)	0.9866	1.0000	0.9997	0.9997	0.9227
(3)	0.9881	0.9987	1.0000	0.9999	0.9243
(4)	0.9872	0.9991	0.9997	1.0000	0.9244
(5)	0.8889	0.9125	0.9156	0.9144	1.0000

*1S-ID-LR, single-stage analysis; 2S-ID-W-FE, two-stage fully-efficient; 2S-ID-W-AVSED, two-stage analysis with AVSED weights (Möhring and Piepho, 2009); 2S-ID-W-S, two-stage with Smith's diagonal weights (Smith et al., 2001); 2S-F-U-ZR, current practice method.

Table 9. *Correlation among adjusted cultivar estimates of spring barley 2015 dataset (above the diagonal: Pearson's product-moment correlation; below the diagonal: Spearman's rank correlation).*

	Strategy*				
Approach	**(1) 1S-ID-LR**	**(2) 2S-ID-W-FE**	**(3) 2S-ID-W-AVSED**	**(4) 2S-ID-W-S**	**(5) 2S-F-U-ZR**
	EBLUP	**EBLUP**	**EBLUP**	**EBLUP**	**EBLUE**
(1)	1.0000	0.9841	0.9812	0.9812	0.9784
(2)	0.9721	1.0000	0.9998	0.9998	0.9977
(3)	0.9704	0.9995	1.0000	1.0000	0.9978
(4)	0.9704	0.9996	1.0000	1.0000	0.9978
(5)	0.9644	0.9964	0.9964	0.9964	1.0000

*1S-ID-LR, single-stage analysis; 2S-ID-W-FE, two-stage fully-efficient; 2S-ID-W-AVSED, two-stage analysis with AVSED weights (Möhring and Piepho, 2009); 2S-ID-W-S, two-stage with Smith's diagonal weights (Smith et al., 2001); 2S-F-U-ZR, current practice method.

In comparison to the MSEPs in Table 7, when exclusively using the Pearson and Spearman correlation, as presented in Tables 8 and 9, it is difficult to determine that the approaches with random effects of cultivar performed better than the one with fixed effects, especially in spring barley. Kobayashi and Salam (2000) mentioned a cogent reason that correlation was not satisfactory for model evaluation since the mean squared deviation (MSD) were easier to interpret and more useful for direct comparison between model output and measurement. For that reason, an additional evidence was needed, which was the MSEP from the CV study. The MSEP obtained in our CV study clearly showed that the four strategies with random effects of cultivar were more accurate than the strategy of current practice. Furthermore, the MSEP revealed that models with complex VCOV structures were likely overfitted. As also shown in our previous two studies, the MSEP values were larger for these complex models than for models with simpler VCOV structures. Based on these results, it is clear that the MSEP shows better discrimination between the different strategies than correlation coefficients. MSEP measures the predictive accuracy (Gauch et al., 2003), and is considered to be more informative than the correlation coefficient.

The MSEP among the 1S-ID-LR, 2S-ID-W-FE, 2S-ID-W-AVSED, and 2S-ID-W-S strategies were very small. For that reason, the choice of strategy depends on computational resources (Gogel et al., 2018). The current software and computational resources (Windows 10, 64-bit operating system, 16GB RAM) only took a couple of minutes for the single-stage approach. When the computational resources are limited, then employing the two-stage analysis is preferable. Another benefit in the two-stage analysis is that we can check whether any errors or correction needed from the results of the first-stage before proceed to the second-stage. Nonetheless, the two-stage fully-efficient analysis needs more memory allocation for conducting the analysis in stage 2 and obtaining the EBLUPs due to the full VCOV is passed to stage 2.

3.3.3 Why is zone-based prediction preferable to individual locations?

As Damesa et al. (2017) pointed out, it was more informative to obtain predictions per agro-ecological zone (larger TPE) than predictions for individual locations. This is because farmers are interested in the cultivar that performs well on average across broad environmental conditions and the next growing season (the next growing season can be considered as a new environment that no trial has previously been conducted in). Another reason is that when a prediction is made for individual farmer's field, the predictions of closest trial location can be used. However, in this case, the valid standard errors for the predictions cannot be achieved since the pattern of interaction between a farmer's field, which is

the target site, and the nearest trial location and the corresponding with years are unknown. Nonetheless, if predictions made for zones or a whole TPE, the valid inferences are possible to be obtained due to the availability of random sample of trial locations and years for that TPE *per se* (Damesa et al., 2017).

Furthermore, from a breeder's perspective, prediction of cultivar performance in a specific site is rarely of interest. Swedish official cultivar trials has the same objective, i.e., to recommend well-performing cultivars for each zone, not for individual trial locations. Thus, accurate information regarding which cultivars perform well within zones or perform above average across locations is essential, for farmers as well as for breeders.

3.3.4 Why not using BLUP in every stage?

The cultivar effects should be modelled as random when the primary goal is selection of the best cultivars from the population under study, and the effects and residuals presumably follow normal distributions. In this case, BLUP will give predictions of cultivar rankings that are close to the true cultivar rankings (McCulloch et al., 2008, p.309; Searle et al., 1992, p.268, 269 and 273). The shrinkage feature of BLUP avoids over-optimistic predictions of top-performing cultivars and over-pessimistic predictions of less-performing cultivars.

Furthermore, with random effects in the C·Z interaction, the accuracy of predictions within zones is improved due to borrowing of information across zones (Atlin et al., 2000; Kleinknecht et al., 2013; Piepho et al., 2016). The borrowing information refers to recovering or using of information across zones. As pointed out by Lee et al. (2017, p.144), "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 and Stein, 1992).

Nevertheless, we discourage using BLUP in the first stage when the two-stage analysis is used. Using BLUP in the first stage leads to double shrinkage, since BLUP is also used in the second stage. If BLUP were to be used in the first stage, predictions have to be unshrunk before proceeding to the second stage (Smith et al, 2001), but it is not obvious how this "unshrinking" should be accounted for in the weighting applied in the second stage. Some progress could be made by taking recourse to so-called "deregressed" proofs as used in animal breeding (Calus et al., 2016), but we consider this step is unnecessary in plant breeding. Dairy cattle breeders, for example, are often forced to use BLUPs in all stages, essentially because bulls do not give milk and a bull's merit can only be referred by "borrowing strength/information" from its female relatives.

3.4 A new inter-zone stability measure analysis

In this study, a new inter-zone stability measure is proposed. The proposed inter-zone stability measure has the similarity to those proposed by Wricke (1962), Shukla (1972), and Denis et al. (1997), i.e., a stability variance of each cultivar across environment based on the cultivar×environment interactions terms, and can be extended by including year term as proposed by Lin and Binns (1988). Piepho (1999) demonstrated the computation of Shukla's stability (1972) by using a mixed-models approach. This was done by computing cultivar-specific variance in the interaction genotype×environment term. Piepho et al. (2016) also demonstrated Shukla's stability (1972) by using a mixed-models approach for cultivar×zone and cultivar×zone×year for two different seeding time of wheat separately. The stability measure that was obtained from cultivar×zone×year was considered similar as type 4 stability since it included the year term.

The drawback of this approach is that the computation time increases as the number of cultivars increases, and the convergence issues will occur due to the amount of data and the model complexity. Besides, it can occur that the stability estimates go to non-positive definite variance due to the model complexity and convergence issues. Piepho et al. (2016) performed the stability analysis separately for each seeding time, so it did not determine the cultivar stability by including the seeding time term.

We propose a new stability measure that applies the same EBLUP model for analysing the MET data. In other words, there is no need to fit other models or covariance structures to compute the stability measure. In other words, if we have other terms, e.g., seeding time or fungicide level, we can obtain the stability measure directly from one model. The conceptual difference of the proposed inter-zone stability measure, as compared with other measures, is that the proposed measure does not directly involve the variance components estimates in the computation of the stability measure and is specifically not the square root of any variance parameter estimate of the fitted model.

Instead, the stability measure is only a descriptive measure of the variability in the estimated interaction effects for each cultivar. Since the stability measure is the standard deviation computed from the coefficients of random-effects/EBLUPs that includes all cultivar×zone interaction terms, it allows the assumption of a model with homogeneous cultivar×zone random interaction effects. Thus, in contrast to the Shukla (1972) approach, the proposed inter-zone stability measure does not require a model with "stability variances" or heterogeneous variance of random cultivar×zone interaction effects, which may not always be easy to fit, especially with a large number of cultivars and a large number of environments. The standard deviation is preferred to the variance, since this is a measure on the same scale as the observations.

3.4.1 Application in the winter wheat 2012–2016 dataset

The inter-zone stability measure of each cultivar in the winter wheat 2012–2016 dataset within each fungicide treatment is presented in Table 10. Computing the stability measure based on the multi-year dataset is more reliable than computing it based on a single-year dataset. Leon and Becker (1988) reported that there is no reliable estimation of phenotypic stability based on a single-year analysis. Stability measure of multi-year and single-year series may differ much because of their using different models and information. Also, it is possible that in the next year, the weather changes drastically. In that case, the stability measure between two different years become very distinct. The stability measure should be based on the multi-year series, since it will be more reliable than based on a single-year.

Table 10. *The DMY stability measure of each cultivar in each level of fungicide treatment based on five-year series winter wheat dataset (2012–2016). The stability measure is a standard deviation of each cultivar based on the combination of zones and years.*

Cultivar	Fungicide	Stability
		g m^{-2}
Dante	Untreated	20.52
Dante	Treated	5.87
Torp	Untreated	15.69
Torp	Treated	13.98
RGT Hasseth	Untreated	22.61
RGT Hasseth	Treated	10.03
R 11224 RAGT	Untreated	17.52
R 11224 RAGT	Treated	8.64
Etana	Untreated	13.48
Etana	Treated	7.87
Ohio	Untreated	14.51
Ohio	Treated	5.18
Lw 08DH642-26	Untreated	3.18
Lw 08DH642-26	Treated	7.03
Lw 06W607-10	Untreated	21.32
Lw 06W607-10	Treated	21.10
Maradona	Untreated	8.38
Maradona	Treated	10.32
Informer	Untreated	7.91
Informer	Treated	5.98

The stability measure was computed using the multi-year model, i.e., the MR 5 model. The multi-year stability measures were obtained based on the EBLUP coefficients of cultivar×zone (C·Z), cultivar×zone×fungicide (C·Z·F), and cultivar×zone×year (C·Z·Y). This proposed stability measure can be related to

the type 4 stability (Lin and Binns, 1988) because the year term is included in the model.

Since the stability measure is a standard deviation of each cultivar based on combinations of zones and years, the smaller the value, the more stable a cultivar is. For example, for cultivar Dane, with fungicide treated, is relatively more stable than without the fungicide, since the deviation with fungicide treatment is smaller (5.87 g m^{-2}) than the untreated (20.52 g m^{-2}). The computation of the stability measure for Dante, as an example, is available in the Appendix.

In general, one has to fit this four-way term, i.e., cultivar×zone×fungicide× year (C·Z·F·Y), to obtain the stability measure of each cultivar for each fungicide across zones and years. The drawbacks with fitting four-way term are; (1) the computation time will increase as the number of cultivar increases and because the complex interaction term has to be fitted, and (2) non positive definite stability estimate may occur, which leads to zero values, and this somewhat makes no sense that the stability of a cultivar is zero.

The proposed stability measure can be computed using the same model that is used for the MET analysis so if the simpler model is adequate for the MET analysis, the stability can be computed without changing any factor nor VCOV structure of the factor in the model. We have shown that the stability for each cultivar in each fungicide treatment across the years can be obtained easily without the four-way interaction terms (C·Z·F·Y), since the best model according to the CV was the model that excluded the four-way interaction term. Thus, it reduces the computation time and can provide a non-zero stability measure of each variety.

In practice, the proposed new inter-zone stability measure requires that any missing data are missing at random (MAR) and the dataset is not highly imbalanced. If missing data are not missing at random and the method for analysis is EBLUP, then there will systematically be varying degree of shrinkage, which causes bias in the relative assessments of stability. For example, if a cultivar is very little tested, then the shrinkage of all its predicted effects will be large, and so this cultivar will appear more stable than it actually is. An alternative can be the options proposed by Edwards and Jannink (2006) and Orellana et al. (2014), who showed that hierarchical Bayesian methods are useful to model heterogeneity of both residual and cultivar×environment interaction variances. The Bayesian method is an appealing alternative, since, in our experience, a model with heterogeneous variance for cultivars is not easily fitted by REML to small datasets. A further study would be valuable to compare the new REML-based inter-zone stability measure and the Bayesian approach.

4 Conclusions

The general conclusions from this study are:

1. The CV revealed that current-statistical method of Swedish cultivar testing, which is using EBLUE for cultivar and cultivar×zone terms has to be abandoned and replaced with EBLUP (random effects) to improve the zone-based prediction accuracy and cultivar rankings.
2. The two-stage unweighted analysis strategy needs to be replaced with either a two-stage weighted or a single-stage analysis.
3. The new inter-zone stability measure has the salient features that it does not need a different model to compute the cultivar stability. Thus, this measure requires little computational time.

The detailed conclusions of each part of this study are given in the following subsections.

4.1 Cross-validation on fungicide-treated subsets datasets

- The EBLUP models performed better than the EBLUE model.
- For the routine analysis of single-year, the recommended model was the SYR 1 model, Z : C + L + C·R, with heterogeneous residuals.
- For the routine analysis of multi-year, the recommended model was the MYR 1 model, Z : C + L + Y + C·Z + C·Y + Y·Z + C·Z·Y.

4.2 Cross-validation on all fungicide levels datasets

- The current model for routine analysis performed less well compared to the EBLUP models. Thus, the currenty used model should be discontinued in routine analysis.

- Winter wheat data were more heterogeneous within zones compared to spring barley, which explained that the model without zonation performed better in winter wheat datasets than in spring barley datasets. However, the model that includes zone still performed better than the one without zone. Thus, zonation is definitely necessary.
- Based on the MSEP, using more interaction terms, (e.g., F·L, C·L·F, C·F·Y, or C·Z·F·Y) or fitting more complex VCOV structures was not necessary.
- For the routine analysis of single-year series, the recommended model was the SR 5 model, Z + F + Z·F : C + L + C·Z + C·F + C·Z·F.
- For the routine analysis of multi-year series, the recommended model was 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.

4.3 Cross-validation for single-stage versus two-stage analysis

- The MSEP from the CV study provided a direct measure of prediction accuracy of single-stage and two-stage strategies compared to merely using the correlation coefficient.
- The two-stage weighting analysis (fully-efficient, AVSED and Smith's diagonal weighting) performed similarly to the single-stage analysis with location-specific residual variances. Thus, the loss of information due to diagonal approximate weighting was negligible.
- The decision of using a single-stage or a two-stage analysis depends on the computational resources.
- The benefit of using two-stage analysis is the possibility to check any errors that were produced in the first-stage and make corrections before proceeding to the second-stage.
- Complex VCOV structures were not necessary, because there were only three zones. Moreover, the complex VCOV caused over-fitting of the model.
- As also shown in our two first studies, the effects of C and C·Z interaction effects were better assigned random than fixed, since it improved accuracy of zone-based prediction through borrowing of information across zones.
- Prediction for zones is more useful and informative for farmers and breeders than prediction for individual locations, since zones cover broader TPEs.

4.4 The new inter-zone stability measure

- The proposed inter-zone stability measure is easily computed directly from the predictions of the random effects, without any computational burden, and may include other factors that are involved in the genotype×environment interactions term, e.g., seeding time and fungicide level.
- The stability measure adds additional valuable information for cultivar recommendation about the stability of the cultivars across zones.
- We recommend using a multi-year dataset to obtain the cultivar stability, since such dataset comprises more information than a single-year dataset.

References

Acquaah, G. 2012. *Principles of plant genetics and breeding, Second edition*, Wiley-Blackwell.West Sussex.

Atlin, G. N., R. J. Baker, K. B. McRae & X. Lu 2000. Selection response in subdivided target regions. *Crop Science,* 40**:** 7-13. doi:10.2135/cropsci2000.4017

Bernardo, R. 1996. Best linear unbiased prediction of maize single-cross performance. *Crop Science,* 36**:** 50-56. 10.2135/cropsci1996.0011183X003600010009x

Bernardo, R. 2010. *Breeding for quantitative traits in plants, Second Edition*, Stemma Press.Woodbury

Blasco, A. 2017. *Bayesian data analysis for animal scientists: The basics,* Springer International Publishing. Cham.

Bustos-Korts, D., M. Malosetti, S. Chapman & F. van Eeuwijk 2016. Modelling of genotype by environment interaction and prediction of complex traits across multiple environments as a synthesis of crop growth modelling, genetics and statistics. *In:* Yin, X. & P. C. Struik (eds.) *Crop systems biology: Narrowing the gaps between crop modelling and genetics.* Cham: Springer International Publishing.

Butler, D. G., B. R. Cullis, A. R. Gilmour, B. J. Gogel & R. Thompson. 2017. ASReml-R reference manual, version 4. University of Wollongong. Wollongong.

Calus, M. P. L., J. Vandenplas, J. ten Napel & R. F. Veerkamp 2016. Validation of simultaneous deregression of cow and bull breeding values and derivation of appropriate weights. *Journal of Dairy Science,* 99**:** 6403-6419. doi:10.3168/jds.2016-11028

Comstock, R. 1977. Quantitative genetics and the design of breeding programme. *In:* Pollak, E., O. Kempthorne & J. T.B. Bailey, eds. International Conference on Quantitative Genetics, 1977. Ames: Iowa State Univ., Iowa State Press, 705–718.

Cooper, M. & I. H. DeLacy 1994. Relationships among analytical methods used to study genotypic variation and genotype-by-environment interaction in plant breeding multi-environment experiments. *Theoretical and Applied Genetics,* 88**:** 561-572. 10.1007/bf01240919

Cooper, M. & G. L. Hammer 1996. *Plant adaptation and crop improvement*, CAB International.Wallingford, UK.

Cooper, M., C. D. Messina, D. Podlich, L. R. Totir, A. Baumgarten, N. J. Hausmann, D. Wright & G. Graham 2014. Predicting the future of plant breeding: Complementing empirical evaluation with genetic prediction. *Crop Pasture Sci.,* 65(4)**:** 311–336. doi:10.1071/CP14007

CPVO 2018. Protecting new plant varieties in Europe. *In:* Office, C. P. V. (ed.). Angers Cedex 2, France: Community Plant Variety Office.

Cullis, B. R., A. Smith, C. Hunt & A. Gilmour 2000. An examination of the efficiency of Australian crop variety evaluation programmes. *The Journal of Agricultural Science,* 135**:** 213-222.

Damesa, T. M., J. Möhring, M. Worku & H. P. Piepho 2017. One step at a time: Stage-wise analysis of a series of experiments. *Agronomy Journal,* 109**:** 845-857. doi:10.2134/agronj2016.07.0395

de Leon, N., J.-L. Jannink, J. W. Edwards & S. M. Kaeppler 2016. Introduction to a special issue on genotype by environment interaction. *Crop Science,* 56**:** 2081-2089. 10.2135/cropsci2016.07.0002in

DeLacy, I. H., K. E. Basford, M. Cooper, J. K. Bull & G. McLaren 1996. Analysis of multi-environment trials—an historical perspective. *In:* Cooper, M. & G. L. Hammer (eds.) *Plant adaptation and crop improvement.* Wallingford: CAB International.

Denis, J.-B., H.-P. Piepho & F. A. van Eeuwijk 1997. Modelling expectation and variance for genotype by environment data. *Heredity,* 79**:** 162. doi:10.1038/hdy.1997.139

DeWitt, T. J. & S. M. Scheiner 2004. *Phenotypic plasticity: Functional and conceptual approaches*, Oxford University Press. Oxford.

Eberhart, S. A. & W. A. Russell 1966. Stability parameters for comparing varieties1. *Crop Science,* 6**:** 36-40. doi:10.2135/cropsci1966.0011183X000600010011x

Edwards, J. W. & J.-L. Jannink 2006. Bayesian modeling of heterogeneous error and genotype × environment interaction variances. *Crop Science,* 46**:** 820-833. doi:10.2135/cropsci2005.0164

Eisenhart, C. 1947. The assumptions underlying the analysis of variance. *Biometrics,* 3**:** 1-21. 10.2307/3001534

FAO 2017. The future of food and agriculture – trends and challenges. Rome: Food and Agriculture Organization of the United Nations.

Fehr, W. 1987. *Principles of cultivar development: Theory and technique*, Macmillian Publishing Company. Ames.

Finlay, K. & G. Wilkinson 1963. The analysis of adaptation in a plant-breeding programme. *Australian Journal of Agricultural Research,* 14**:** 742-754. https://doi.org/10.1071/AR9630742

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. doi:10.1080/02664763.2013.825703

Forkman, J. & H. P. Piepho 2013. Performance of empirical blup and bayesian prediction in small randomized complete block experiments. *The Journal of Agricultural Science,* 151**:** 381-395. doi:10.1017/S0021859612000445

Galwey, N. W. 2014. *Introduction to mixed modelling: Beyond regression and analysis of variance, Second Edition*, John Wiley & Sons, Ltd. West Sussex.

Gauch, H. G., J. T. G. Hwang & G. W. Fick 2003. Model evaluation by comparison of model-based predictions and measured values. *Agronomy Journal,* 95**:** 1442-1446. doi:10.2134/agronj2003.1442

Gogel, B., A. Smith & B. Cullis 2018. Comparison of a one- and two-stage mixed model analysis of Australia's National Variety Trial Southern Region wheat data. *Euphytica,* 214**:** 44. doi:10.1007/s10681-018-2116-4

Harville, D. 1976. Extension of the Gauss-Markov theorem to include the estimation of random effects. *The Annals of Statistics,* 4**:** 384-395.

Henderson, C. R. 1950. Estimation of genetic parameters. *Annals of Mathematical Statistics,* 21**:** 309–310.

Henderson, C. R. 1963. Selection index and expected genetic advance. *In:* Hanson, W. D. & H. F. Robinson, eds. Statistical Genetics and Plant Breeding, 1963. Washington, DC: National Academy of Science—National Research Council Publication, 141–163.

Henderson, C. R. 1975. Best linear unbiased estimation and prediction under a selection model. *Biometrics,* 31**:** 423–447.

Henderson, C. R. 1984. *Applications of linear models in animal breeding*, University of Guelph. Guelph.

Isik, F., J. Holland & C. Maltecca 2017. Multi environmental trials. *Genetic data analysis for plant and animal breeding.* Cham: Springer International Publishing.

James, G., D. Witten, T. Hastie & R. Tibshirani 2013. *An introduction to statistical learning: With applications in R*, Springer New York. New York.

James, W. & C. Stein 1992. Estimation with quadratic loss. *In:* Kotz, S. & N. L. Johnson (eds.) *Breakthroughs in statistics: Foundations and basic theory.* New York, NY: Springer New York.

Jordbruksverket. 2015. *Plant variety catalogues* [Online]. Available: http://www.jordbruksverket.se/swedishboardofagriculture/engelskasidor/crops/plantvarieties/plantvarietycatalogues.4.4e88d23a14e47fc2869eeb10.html [Accessed 3 June 2019].

Kang, M. S. & D. P. Gorman 1989. Genotype × environment interaction in maize. *Agronomy Journal,* 81**:** 662-664. 10.2134/agronj1989.00021962008100040020x

Kleinknecht, K., J. Möhring, K. P. Singh, P. H. Zaidi, G. N. Atlin & 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 Science,* 53**:** 1384-1391. doi:10.2135/cropsci2013.02.0073

Kobayashi, K. & M. U. Salam 2000. Comparing simulated and measured values using mean squared deviation and its components. *Agronomy Journal,* 92**:** 345-352. 10.2134/agronj2000.922345x

Laird, N. M. & J. H. Ware 1982. Random-effects models for longitudinal data. *Biometrics,* 38**:** 963-974. 10.2307/2529876

Lee, Y., J. A. Nelder & Y. Pawitan 2017. *Generalized linear models with random effects: Unified analysis via H-likelihood, Second Edition*, CRC Press/Taylor & Francis Group. Boca Raton.

Lin, C. S. & M. R. Binns 1988. A method of analyzing cultivar×location×year experiments: A new stability parameter. *Theoretical and Applied Genetics,* 76**:** 425-430. 10.1007/bf00265344

Lin, C. S., M. R. Binns & L. P. Lefkovitch 1986. Stability analysis: Where do we stand? *Crop Science,* 26**:** 894-900. doi:10.2135/cropsci1986.0011183X002600050012x

Lynch, M. & B. Walsh 1998. *Genetics and analysis of quantitative traits*, Sinauer Associates. Sunderland

Malosetti, M., J.-M. Ribaut & F. A. van Eeuwijk 2013. The statistical analysis of multi-environment data: Modeling genotype-by-environment interaction and its genetic basis. *Frontiers in Physiology,* 4**:** 44. doi:10.3389/fphys.2013.00044

McCulloch, C. E., S. R. Searle & J. M. Neuhaus 2008. *Generalized, linear, and mixed models, Second edition*, John Wiley & Sons, Inc. Hoboken, New Jersey.

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

Olsen, A. K. B., T. Persson, A. Wit, L. Nkurunziza, E. Sindhøj & H. Eckersten 2018. Estimating winter survival of winter wheat by simulations of plant frost tolerance. *Journal of Agronomy and Crop Science,* 204**:** 62-73. doi:10.1111/jac.12238

Orellana, M., J. Edwards & A. Carriquiry 2014. Heterogeneous variances in multi-environment yield trials for corn hybrids. *Crop Science,* 54**:** 1048-1056. doi:10.2135/cropsci2013.09.0653

Patterson, H. D. 1997. Analysis of series of variety trials. *In:* Kempton, R. A. & P. N. Fox (eds.) *Statistical methods for plant variety evaluation.* London: Chapman & Hall.

Patterson, H. D. & V. Silvey 1980. Statutory and recommended list trials of crop varieties in the United Kingdom. *Journal of the Royal Statistical Society. Series A (General),* 143**:** 219-252. doi:10.2307/2982128

Patterson, H. D. & R. Thompson 1971. Recovery of inter-block information when block sizes are unequal. *Biometrika,* 58**:** 545-554. 10.2307/2334389

Patterson, H. D. & E. R. Williams 1976. A new class of resolvable incomplete block designs. *Biometrika,* 63**:** 83-92. 10.1093/biomet/63.1.83

Piepho, H.-P., M. F. Nazir, M. Qamar, A.-u.-R. Rattu, Riaz-ud-Din, M. Hussain, G. Ahmad, Fazal-e-Subhan, J. Ahmad, Abdullah, K. B. Laghari, I. A. Vistro, M. S. Kakar, M. A. Sial & M. Imtiaz 2016. Stability analysis for a countrywide series of wheat trials in Pakistan. *Crop Science,* 56**:** 2465-2475. doi:10.2135/cropsci2015.12.0743

Piepho, H. P. 1996. Analysis of genotype-by environment interaction and phenotypic stability. *Genotype-by-environment interaction and phenotypic stability.* Boca Raton: CRC Press.

Piepho, H. P. 1998a. Empirical best linear unbiased prediction in cultivar trials using factor-analytic variance-covariance structures. *Theoretical and Applied Genetics,* 97**:** 195-201. doi:10.1007/s001220050885

Piepho, H. P. 1998b. Methods for comparing the yield stability of cropping systems. *Journal of Agronomy and Crop Science,* 180**:** 193-213. doi:10.1111/j.1439-037X.1998.tb00526.x

Piepho, H. P. 1999. Stability analysis using the SAS system. *Agronomy Journal,* 91**:** 154-160. doi:10.2134/agronj1999.00021962009100010024x

Piepho, H. P., A. Büchse & K. Emrich 2003. A hitchhiker's guide to mixed models for randomized experiments. *Journal of Agronomy and Crop Science,* 189**:** 310-322. doi:10.1046/j.1439-037X.2003.00049.x

Piepho, H. P. & T. Eckl 2014. Analysis of series of variety trials with perennial crops. *Grass and Forage Science,* 69**:** 431-440. doi:10.1111/gfs.12054

Piepho, H. P. & J. Möhring 2006. Selection in cultivar trials—is it ignorable? *Crop Science,* 46**:** 192-201. doi:10.2135/cropsci2005.04-0038

Piepho, H. P., J. Möhring, T. Schulz-Streeck & J. O. Ogutu 2012a. A stage-wise approach for the analysis of multi-environment trials. *Biometrical Journal,* 54**:** 844-860. doi:10.1002/bimj.201100219

Piepho, H. P., E. R. Williams & L. V. Madden 2012b. 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

R Core Team. 2018. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria.

Robinson, G. K. 1991. That BLUP is a good thing: The estimation of random effects. *Statistical Science,* 6**:** 15-32.

RStudio Team. 2016. Rstudio: Integrated development environment for R. RStudio, Inc. Boston, MA.

SAS Institute. 2013. SAS system for windows 9.4. SAS Inst. Inc. 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. *Animal Science Journal,* 89**:** 876-879. doi:10.1111/asj.13016

Schabenberger, O. & F. J. Pierce 2001. *Contemporary statistical models for the plant and soil sciences*, CRC Press. Boca Raton.

Searle, S. R., G. Casella & C. E. McCulloch 1992. *Variance components*, John Wiley & Sons. New York

Shukla, G. K. 1972. Some statistical aspects of partitioning genotype-environmental components of variability. *Heredity,* 29**:** 237. doi:10.1038/hdy.1972.87

Smith, A., B. R. Cullis & A. Gilmour 2001. Applications: The analysis of crop variety evaluation data in Australia. *Australian & New Zealand Journal of Statistics,* 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.

Wallach, D. & B. Goffinet 1989. Mean squared error of prediction as a criterion for evaluating and comparing system models. *Ecological Modelling,* 44**:** 299-306. https://doi.org/10.1016/0304-3800(89)90035-5

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

Welham, S. J., B. J. Gogel, A. B. Smith, R. Thompson & B. R. Cullis 2010. A comparison of analysis methods for late-stage variety evaluation trials. *Australian & New Zealand Journal of Statistics,* 52**:** 125-149. doi:10.1111/j.1467-842X.2010.00570.x

Wickham, H. 2009. *Ggplot2: Elegant graphics for data analysis*, Springer-Verlag. New York.

Wilkinson, G. N. & C. E. Rogers 1973. Symbolic description of factorial models for analysis of variance. *Journal of the Royal Statistical Society. Series C (Applied Statistics),* 22**:** 392-399. doi:10.2307/2346786

Wricke, G. 1962. Über eine Methode zur Erfassung der ökologischen Streubreite in Feldversuchen. *Zeitschrifi für Pflanzenzüchtung,* 47**:** 92–96.

Popular science summary

Background

Swedish official cultivar testing conducts multienvironmental trials (MET) every year to evaluate the performance, i.e., yield, of a vast number of cultivars in different environmental conditions because different cultivars perform differently in various environmental conditions, known as genotype× environment interactions (GEI). The aim of a MET is to provide accurate information of cultivar performance so that a recommendation of which cultivar that performs the best in a farmer's growing condition can be available. A robust and reliable statistical procedure is needed to fulfil this aim. Thus, in this thesis, the current-practice statistical model of Swedish official cultivar testing used for analysing MET data was assessed via cross-validation (CV).

A CV study is conducted for model selection purpose. The idea of CV is to split a dataset into a training set and a validation set. The models that are proposed will be fitted to the training set. The prediction values from the fitted model in the training set will be subtracted with the values in the validation set. Then, these differences will be squared and the average computed, known as the mean squared error of prediction differences (MSEP). Thus, the model that produces the smallest MSEP is considered as the best model that provides the best prediction accuracy.

Problems

The current-practice statistical method in Swedish official cultivar testing uses fixed cultivar effects, known as best linear unbiased estimation (BLUE). BLUE has drawbacks that the estimation of yield will be too optimistic since it cannot handle missing data properly, and so decrease the accuracy of the cultivars' yield information. Note that, in the MET data, the degree of imbalanced is usually high. We proposed to use a random cultivar effects, known as best linear unbiased prediction (BLUP), because BLUP will handle missing data better than BLUE. For example, BLUP will use all information of cultivars

in other environments, and so it maximises the prediction of the cultivars. Moreover, BLUP is a prediction that predicts the future performance of the cultivars, while BLUE is merely an estimation of what has been done in the MET, since BLUE assumes the estimation will not change when the MET is repeated. BLUE is inappropriate, since MET should predict the future performance of the crop. The environmental conditions, such as weather, unpredictably changes from year to year.

The analysis step of the routine procedure in Swedish official cultivar testing was also assessed. In general, there are single-step and two-step analyses. The single-step analysis has the benefit that it only needs a one-time analysis with a single model for the results. However, the computational time may increase due to the amount of data and the complexity of the model. The current method is a two-step analysis without any precision measure (unweighted). In this method, individual trials/locations in each zone are analysed separately. Then, the results from step 1, i.e., the adjusted means of each cultivar at each location are passed to step 2 to make a final analysis for the zone level. The drawback of this method is that without carrying the precision measure from step 1, the analysis results in step 2 are less precise, due to loss of information from step 1. Besides, with many missing data, the precision measure is essential to "weight" each trial/location as a "correction".

What was done in this study?

We wanted to improve prediction accuracy in three Swedish agricultural zones, i.e., south, middle, and north of Sweden. Thus, a zone consists of several locations. In other words, a zone is a compilation of several locations that have similar geographical or climate conditions. Thus, a CV study was performed to compare the current statistical method, which uses BLUE, with BLUP for zone-based prediction.

We also compared the current practice of a simple two-step analysis with the single-step and several two-step weighted analyses with different precision measures via CV. The single-step analysis was able to be included in this study, since the availability of developed software that able to run the single-step analysis in a short time.

A new inter-zone stability measure was proposed to assess the stability of cultivars. Stability measure is important because they provide information regarding the performance of cultivars across zones and years. Farmers would like to have not only high yield but also a stable cultivar across the years. Many stability measures are available. However, the computation method and time are sometimes a burden. For that reason, we proposed this stability measure that is easier and more flexible to compute.

The results

The results of this thesis showed that BLUP outperformed BLUE, and the two-step unweighted procedure performed less well compared to the single-step and the two-step analyses. Thus, the routine procedure of the current-practice should be replaced by using BLUP in combination with, either a single-step or a two-step weighted analysis. The proposed new inter-zone stability measure can be computed easily. Moreover, the stability measure provides more information for farmers to choose not only a high-yield but also a stable cultivar across years since the stability measure may be obtained from a multi-year series analysis.

Acknowledgements

I thank my supervisor Johannes Forkman, for the invaluable opportunity to work on this project. You have believed in me and always been supportive. Also, you have allowed me to learn things by doing.

I thank my other supervisors, Hans-Peter Piepho, Jannie Hagman, and Jesper Rydén for sharing your knowledge, and providing constructive comments and suggestions.

I thank the member of this project, Karl-Oskar Andersson, Anders Ericsson, Alf Ceplitis, and Jannie Hagman for the great support and input.

I thank Magnus Haling from Swedish official cultivar testing for the helpful discussion.

I thank my colleagues at the unit of Applied Statistics and Mathematics, Razaw, Claudia, Ulf; you have always been thoughtful and gave the opportunity to broaden my network and experience in the Applied Statistics field.

I thank Dietrich and Tomas, who have shared their life experience and pieces of advice every time we had lunchtime.

I thank my officemate, Cigdem. You are not only a great officemate, but also a marvellous friend and teacher.

My friends in Crop Production Ecology, James, Xiangyu, Hui, and Elsa; it has been a great time to have fika and discussion with all of you. You all are awesome.

I thank my Indonesian friends, especially Fahry, Michael, Mira, Ng, Rahmanu, Reza, and Sanka, for this great friendship.

I thank Stiftelsen lantbruksforskning (SLF) – Swedish farmers' foundation for agricultural research, which funded this project.

Finally, I thank my parents. You have always supported and allowed me to pursue my passion.

Appendix

The dry matter yield (DMY) coefficients of cultivar×zone×fungicide (C·Z·F), cultivar×zone (C·Z), cultivar×zone×year (C·Z·Y), and the summation of the coefficients for cultivar Dante in each level of fungicide treatment for computing the stability measure based on five-year series winter wheat dataset (2012–2016) fitted with the MR 5 model.

Year	Cultivar	Region	Fungicide	EBLUP coefficients			Summation
				C·Z·F	C·Z·F	C·Z·Y	
2013	Dante	South	Untreated	256.51	17.87	-72.45	201.93
2013	Dante	Middle	Untreated	-77.54	-8.05	-20.00	-105.59
2014	Dante	South	Untreated	256.51	17.87	123.09	397.47
2014	Dante	Middle	Untreated	-77.54	-8.05	-48.27	-133.86
2014	Dante	North	Untreated	-110.99	-6.27	-78.47	-195.73
2015	Dante	South	Untreated	256.51	17.87	102.85	377.23
2015	Dante	Middle	Untreated	-77.54	-8.05	-0.90	-86.50
2015	Dante	North	Untreated	-110.99	-6.27	24.59	-92.68
2013	Dante	South	Treated	-99.29	17.87	-72.45	-153.87
2013	Dante	Middle	Treated	6.70	-8.05	-20.00	-21.35
2014	Dante	South	Treated	-99.29	17.87	123.09	41.68
2014	Dante	Middle	Treated	6.70	-8.05	-48.27	-49.62
2014	Dante	North	Treated	55.81	-6.27	-78.47	-28.93
2015	Dante	South	Treated	-99.29	17.87	102.85	21.44
2015	Dante	Middle	Treated	6.70	-8.05	-0.90	-2.25
2015	Dante	North	Treated	55.81	-6.27	24.59	74.12

*The stability measure of Dante for each fungicide is computed based on eight values of the summation.