UNIVERSITÄT HOHENHEIM

FACULTY OF AGRICULTURAL SCIENCES

Institute of Crop Science

Department of Biostatistics (340c)

Prof. Dr. Hans-Peter Piepho

Weighting methods for variance heterogeneity in phenotypic and genomic data analysis for crop breeding

Dissertation

submitted in fulfillment of the regulations to acquire the degree

“Doktor der Agrarwissenschaften”

(Dr.sc.agr. / Ph.D. in Agricultural Sciences)

to the

Faculty of Agricultural Sciences

presented by:

Tigist Mideksa Damesa

from

Addis Ababa, Ethiopia

Submitted in: April, 2019

This thesis was accepted as a doctoral thesis (Dissertation) in fulfilment of the regulations to acquire the doctoral degree “Doktor der Agrawissenschaften” by the Faculity of Agricultural Sciences at University of Hohenheim on July 10, 2019.

Date of the oral examination: September 26th, 2019

Examination Committee

Chairperson of the oral examination: Prof. Dr. Thilo Streck

Supervisor and reviewer: Prof. Dr. Hans-Peter Piepho

Co-Reviewer: Dr. Johannes Forkman

Additional examiner: Prof. Dr. Bettina Haussmann

Acknowledgments

First and foremost, I am very much grateful to my research supervisor Prof. Dr. Hans-Peter Piepho for his advice and continuous support of my P.h.D research. I appreciate all his contributions of ideas and time throughout my study.

Many thanks to Jens Hartung for his unreserved help, for reading the manuscripts, and for being co-author of publications.

Thanks to Prof. Dr. Bettina Haussmann and Prof. Dr. Thilo Streck for being the exam committee.

I am grateful to Dr. Gezahegn Bogale (Ethiopian Institute of Agricultural Research, Melkassa) for providing maize trial data. I also acknowledge Dr. Mosisa Worku for sharing maize trial data.

Thank you to Dr. Johannes Forkman for the collaboration and for being the exam committe.

I would like to thank Dr. Mosisa Worku, Dr. Manje Gowda, Dr. Yoseph Beyene, Dr. Biswanatha Das and Dr. Kassa Fentaye for being co-authors of the publications.

I am also greatful to members of the Biostatistics group. I appreciate the help of Mrs. Zeynep Akyildiz from the beginning until the end of my project.

Most sincere thanks to my family, who encouraged and supported me to follow my dreams.

The financial support from German Academic Exchange Service (DAAD) throughout my research period is gratefully acknowledged.

Table of Contents

Abbreviations

AIC	Akaike information criterion
AM	association mapping
AR(1)	one-dimensional first-order autoregressive model
AR(1)×AR(1)	two-dimensional first-order autoregressive model
BLUE	best linear unbiased estimation
BLUP	best linear unbiased prediction
CIMMYT	International Maize and Wheat Improvement Center
CS	compound symmetry
CSH	heterogeneous compound symmetry
CV	cross validation
EVCDTH12	evaluation of CIMMYT drought tolerant hybrids in 2012
FA	factor-analytic
GBLUP	genomic best linear unbiased prediction
GBS	genotyping-by-sequencing
GEBV	genomic estimated breeding values
GEI	genotype by environment interaction
GLMM	generalized linear mixed model
GS	genomic selection
GWAS	genome-wide association studies
LD	linkage disequilibrium
LMM	linear mixed model
MEI	marker-by-environment interaction

MET	multi-environment trial
ML	maximum likelihood
MSD	mean squared difference
NNA	nearest-neighbor analysis
Non-QPM	non-quality protein maize
PC	principal component
PCA	principal component analysis
PN	power normal
POM	power-of-the-mean
QTL	quantitative trait loci
REML	restricted maximum likelihood
TPE	target population of environments
UN	unstructured
UNR	unstructured correlation

Chapter 1

General introduction

1.1 Multi-environment field trials

The main objective of plant breeding and variety testing is the development of high quality genotypes in terms of yield, and other important characteristics such as disease resistance and drought tolerance. The performance of a given genotype is determined by the genetic make-up of the plant, the environment and genotype-by-environment interaction, where an environment represents a site or site-year combination. In order to control environmental factors and make reliable selections of well performing genotypes, trials are usually replicated at several sites, and over several years and/or seasons (Cochran, 1937; Yates and Cochran, 1938; Comstock and Moll, 1963; Gauch, 1992; Talbot, 1997). Such trials are known as multi-environment trials (MET). Data from MET are used to investigate the average performance of genotypes in a range of environments, representing a target population of environments (TPE) (Atlin et al., 2000), and are also used to measure stability of traits accurately (Crossa, 1990).

MET data give rises to different sources of within-trial and between-trial variation, and there is usually heterogeneity of variance at both of these levels. If these sources of variation and the variance heterogeneity are not accounted for, inefficient estimates of genotype effects may result, which adversely affects selection gain (Edwards et al., 2015). In order to account for all sources of variation and obtain reliable results, choice of good experimental design and appropriate analysis, accounting for any heterogeneity of variance, are crucial (Fisher, 1935; Cullis et al., 1998; Smith et al., 2001; Piepho et al., 2012a).

1.2 Accounting for within-trial variation and heterogeneous error variance

Data from field trials shows substantial variation that arises from multiple sources. Some examples of such sources of variability are soil moisture gradients, variation in experimental procedure, and other factors like disease and drought. These sources of variability should be separated from genotype mean estimates. Usually, field variability is controlled using proper experimental designs along with the corresponding design-based analysis (Fisher, 1935). In addition, a number of studies showed that analyses of field trials with models which account for spatial correlation are superior to traditional purely randomization-based analyses (Gilmour et al., 1997; Schabenberger and Pierce, 2002; Piepho et al., 2008; Piepho and Williams, 2010; Müller et al., 2010; Sripathi et al., 2017). Spatial modeling approaches are therefore gaining popularity in plant breeding. Spatial models can be categorized into two kinds, i.e., isotropic and anisotropic. Isotropy means the spatial variation depends only on the distance between observations whereas anisotropy means the spatial correlation depends both on the distance and direction. In this study, we fitted isotropic one-dimensional models assuming auto-regressive (AR), exponential, spherical, and Gaussian covariance structures and assuming that correlation exists only along rows, but different rows were independent. For anisotropic modelling we considered the geometric exponential, spherical, and Gaussian and the two-dimensional AR(1)×AR(1) covariance structures.

Data from agricultural field trials are often analysed based on classical linear model assumptions for the error term. For example, the baseline (randomization-based) model assumes independent error terms with homogeneous variance. By contrast, most spatial models assume dependent error effects, but still assuming constant variance. This study is concerned with the analysis of agricultural field trials when the asumption of homogeneous variance is violated. In variety performance trials, it is often observed that within-trial error variance differs between enviroments. If data analysis is done without considering the variance heterogeneity, then the analysis results may be misleading and may change the conclusion of the study compared to an approprate one.

To account for the variance heterogeneity problem, there are various techniques available. Variance modelling and data transformation are two of the common methods (Box and Cox, 1964; Carroll and Ruppert, 1988; Piepho, 2009). Variance modelling allows the analysis of data with unequal variance per experimental unit. One popular variance model assumes that the variance is proportional to the power of the mean (Carroll and Ruppert, 1988). Data transformation techniques also help to handle the variance heterogeneity problem (Box and Cox, 1964; Lee et al., 2008; Piepho, 2009). However, even if a data transformation resolves the variance heterogeneity problem for a single trial, it is unsatisfactory when it comes to analysis of series of field trials, especially when the optimal transformation differs between trials. The reason is that back-transformation to original scale is not easy (Freeman and Modarres, 2006).

This thesis proposes and demonstrates methods for analyzing MET data when the classical assumption of within-trial homogeneous variance is violated based on variance modeling approaches. Furthermore, an extension of the approach to simultaneously handle spatial variation along with heterogeneity of variance is considered.

1.3 Analysis methods for MET

There exist several statistical methods for analyzing MET data (Finlay and Wilkinson, 1963; Kempton, 1984; Piepho, 1997; Piepho et al., 1998; Smith et al., 2001). Linear mixed models (LMM) provide a convenient approach for analysis of MET, because they can handle the complexities of MET such as unbalancedness, unequal variances and spatial correlation. LMM for MET data can be fitted in two different ways: either as a single-stage analysis or as a stage-wise analysis. In single-stage analysis a combined analysis of raw plot data is considered and all source of variation are estimated simultaneously. This approach is considered to be the gold standard of MET analyses (Smith et al., 2001, 2005; Piepho et al., 2012). However, single-stage analysis may require large computation time, due to the fact that MET often produce large datasets and require complex variance-covariance structures to be fitted. As a result, MET data are often analyzed using a stage-wise approach, in which genotype means are first computed from individual trial analyses and then in the next stage these means are combined for a joint analysis using a mixed model.

In MET analysis, error and genotype-by-environment interaction (GEI) variance are usually heterogeneous between trials (Frensham et al., 1997; Cullis et al., 1998). In stage-wise analysis the error variance in the second stage is considered known but is replaced by its residual maximum likelihood (REML) estimate from the first-stage (individual trial) analysis. To account for heterogeneous error variances, a weighting approach is used for the joint analysis (Smith et al., 2001; Piepho et al., 2012a). The weights are derived from the variance-covariance matrix of the adjusted genotype means computed in the first stage.

To fit heterogeneous GEI variances, different approaches have been proposed. Examples are multiplicative models and factor-analytic (FA) variance structures for the interaction variance (Gogel et al., 1995; Piepho, 1997; Smith et al., 2015; Smith et al., 2018). In this study we are concerned only with approaches to account for error variance heterogeneity between trials, whereas for random GEI effects we used the simplest model.

Stage-wise analysis is an approximation for single-stage analysis. Stage-wise analysis will very closely resemble single-stage analysis if the full information is forwarded from the first stage to the second stage using an appropriate weighting method. This thesis explores methods that deal with variance heterogeneity between MET data in the most efficient way, which involves weighting based on mixed model approaches using the full information from previous stages.

1.4 Weighted genomic selection and genome-wide association studies

In modern plant breeding different types of marker-based procedures are applied to increase genetic gain and improve the quality of genotypes. Marker-assisted selection (MAS) and genomic selection (GS) have become important tools for breeders to select superior genotypes. MAS is an indirect type of selection that uses molecular markers in linkage disequilibrium (LD) with quantitative trait loci (QTL). Linkage mapping (LM) and genome-wide association studies (GWAS) are the two commonly used methods to identify markers for MAS (Yu et al., 2006; Oraguzie et al., 2007). GS is another type of selection for improving plant breeding using whole genome molecular markers to predict genomic estimated breeding values (GEBV) of both phenotyped and unphenotyped genotypes (Meuwissen et al., 2001; Hayes and Goddard, 2010; Gowda et al., 2015). In plant breeding, MET data are central to select genotypes using observed data (phenotype), and also for marker-based selection (MAS and GS). Therefore an appropriate analysis of phenotype data is indispensible to obtain accurate and reliable results for both GWAS and GS. In phenotypic MET analysis some researchers use weights and some do not. However, when the researcher's objective is to do GS or GWAS analyses, adjusted means obtained from MET are almost invariably forwarded to the actual GS or GWAS analyses without any weighting method being applied (Shikha et al., 2017; Edriss et al., 2017). In this thesis weighted and unweighted methods are compared for GS and GWAS analyses.

1.5 Objectives

The main objective of this thesis is to use existing statistical methods and determine a best approach for handling within-trial and between-trial variance heterogeneity in phenotypic MET and genomic data analyses using weighting methods. The specific objectives are:

1. Propose a method to account for within-trial variance heterogeneity in the case of MET
2. Compare spatial versus baseline models
3. Determine the best approach for account for variance heterogeneity and within-trial spatial correlation at the same time
4. Evaluate if accounting of spatial variation and heterogeneous error improves the analysis or not
5. Demonstrate the application of a new weighting method called the fully efficient method in stage-wise analysis of MET
6. Compare the performance of fully efficient weighting with diagonal weighting and unweighted analysis of MET
7. Evaluate weighted versus unweighted methods when analysis of MET is extended to GS and GWAS analysis

1.6 Outline of the thesis

In Chapter 2, statistical approaches for the simultaneous handling of dependent and heteroscedastic errors for the case of MET are demonstrated. In Chapter 3, the use of the fully efficient weighting method for stage-wise analysis of three different types of MET is illustrated and its performance evaluated and compared to other weighting methods. In Chapter 4 the use of weighting methods in stage-wise analysis of GS and GWAS and its comparison with the unweighted stage-wise analysis is discussed. Chapter 5 provides a general discussion of the thesis.

Chapter 2

Modelling spatially correlated and heteroscedastic errors in Ethiopian maize trials

Tigist Mideksa Damesa[1], Jens Möhring[1], Johannes Forkman[2], Hans-Peter Piepho[1]*

[1]Biostatistics Unit, Institute of Crop Science, University of Hohenheim, Fruwirthstrasse 23, 70599 Stuttgart, Germany.

[2]Department of Crop Production Ecology, Swedish University of Agricultural Sciences, Uppsala, Sweden.

Published in Crop Sci. 58:1575–1586 (2018)

The original publication is available at

doi: 10.2135/cropsci2017.11.0693

2.1 Abstract

The precision of estimates of genotype means and genotype comparisons in agricultural field trials can be increased by using an appropriate experimental design and spatial modelling techniques. Both randomization-based and spatial analysis usually make the assumption of homogeneous variance. But in reality this assumption may not generally hold true. If this is ignored, erroneous estimates of the precision of fixed effect estimates can result, therefore some remedy should be sought in case heterogeneity of variance is detected. The objective of this study is to investigate methods of analysis accounting for possible variance heterogeneity along with the spatial trend if any. The methods are explored using three maize trials from Ethiopia. We consider the Box-Cox transformation to stabilize variance and variance models allowing for heterogeneity. For variance modelling we use the power-of-the-mean (POM) and exponential models. The Box-Cox transformation was found to be successful in stabilizing the variance but estimating genotype means and their standard error on the original scale is challenging. The POM and exponential variance models, which avoid this problem, were

found to effectively deal simultaneously with both spatial correlation and heterogeneity of variance.

In plant breeding field experiments, there are many biotic and abiotic sources of variability that can adversely affect mean estimates of the genotypes. There are cases for which the variability might have a spatial trend, and if not controlled, this will result in poor estimation and ranking of genotypic performance. Appropriate statistical design and modeling approaches help to address these challenges.

Blocking techniques, replication, and randomization have traditionally been used for controlling field variability (Fisher, 1935; Edmondson, 2005). However, the associated randomization-based analyses do not fully exploit the presence of spatial correlation among field plots. Different spatial analysis methods, such as nearest-neighbor analysis and various autoregressive models [AR(1), linear variance, etc.] (Gleeson and Cullis, 1987; Cullis and Gleeson, 1991; Gilmour et al., 1997; Piepho et al., 2008; Piepho and Williams, 2010; Müller et al., 2010; Sripathi et al., 2017) are available that are based on the assumption that near plots are more highly correlated than more spatially separated plots (Schabenberger and Pierce, 2002).

In linear models, estimation of the unknown fixed effect parameters usually makes the assumption of homogeneous variance, meaning that if this assumption is not attained, there will be a loss of efficiency. The existence of variance heterogeneity in experimental field trials is not unusual. Overlooking this problem will result in inaccurate inferences on the fixed effects (Carroll and Ruppert, 1988; Littell et al., 2006). The heterogeneity of variance may exist between treatments, or it may be due to a variance–mean relationship; depending on the type of heterogeneity, there are different remedial measures. Here, we will focus on the most common case, where the variance is a function of the mean.

A nonlinear variance-stabilizing transformation can be tried as a remedy (Box and Cox, 1964; Carroll and Ruppert, 1988; Sakia, 1992; Piepho, 2009). Although this approach may be successful in stabilizing the variance, thus allowing a valid analysis on the transformed scale, interpretation of the estimate on the transformed scale may be difficult. Moreover, back-transformation to the original scale is not straightforward. Usually the inverse of the Box–Cox transformation is calculated, which leads to an estimate of the median on the original scale (Piepho, 2009). This naive back-transformation is an adequate approach for a single trial, but more thought is needed when aiming to integrate results from a series of trials, a common task in stage-wise analysis of multienvironment trials (MET). The reason is that different transformations may be needed in different trials, which complicates the integration of results across trials on the original scale. In MET, the combined analysis is often done using a two-stage analysis, in which each trial is analyzed independently in the first stage and adjusted genotype mean estimates are saved. In the second stage, a joint analysis may be done using mixed models (Möhring and Piepho, 2009; Piepho et al., 2012; Piepho and Eckl, 2014; Damesa et al., 2017). For integrating trial results across environments, all estimated genotype means should be in the original units of measurement. For a single trial, use of a median estimate is unproblematic; however, for combined analysis of MET where a linear model is assumed for the estimates computed per trial, an estimate of the expected value is more suitable than an estimate of the median. In addition, an estimate of the variance-covariance matrix of the adjusted means on the original scale is needed, and this is difficult to obtain when a transformation is involved. Freeman and Modarres (2006) studied the moments of the power normal distribution and derived an expression for the expected value and variance on the original scale when the parameter of the Box–Cox transformation is between zero and one. For other values of the transformation parameter, however, no simple equations are available both for the expected value and variance-covariance matrix on the original scale.

If the variance increases or decreases in relation to the mean and approximate normality can be assumed on the observed scale, modeling of the variance as a function of the mean is an alternative and more flexible approach that avoids the complications of data transformation. Essentially, this approach assigns relative weights to the observation depending on their predicted mean. The power-of-the-mean (POM) and exponential models are two examples for variance modeling (Carroll and Ruppert, 1988). An advantage of this approach is that mean estimates are obtained directly on the observed scale, thus facilitating two-stage analysis to integrate results from MET.

The variance modeling approach is applied only to resolve problems related to heteroscedasticity, assuming that approximate normality and additivity hold on the original scale, whereas a nonlinear transformation is applied in the hopes to not only stabilize a heterogeneous variance, but also to help fix other linear model assumption failures such as non-additivity and non-normality in the original scale. For example, multiplicative effects and lognormal distributions on the original scale imply additive effects and normal distribution on the log scale. The transform-both-sides method (Carroll and Ruppert, 1988), which applies the same transformation to both sides of the model equation, is another option when the data shows both skewness and nonconstant variance on the original scale. However, this option is not considered here because it is fraught with the same difficulties as the Box–Cox transformation in getting back to the original scale.

In this study, we consider three maize trials from Ethiopia. Inspection of the residuals from the randomization-based analysis indicates that there is variance heterogeneity in all these three trials, with the variance decreasing as the mean increases. The main objectives of this study are (i) to propose a method to fit spatial models in case of heterogeneity of variance, (ii) to fit spatial models both along with heterogeneous variance and after stabilizing the variance with Box–Cox transformation for the maize trial data, and (iii) to evaluate if the spatial modeling with heterogeneous variance improves the analysis when no data transformation is used.

2.2 Material and methods

2.2.1 Data

For illustration, we use three different drought tolerance maize trials, which were obtained from the Melkassa Center of the Ethiopian Institute of Agricultural Research. In the beginning of our study, we considered a 4-yr multisite dataset and first analyzed all individual trials using the randomization-based model and checked the assumption of variance homogeneity. Out of all those trials, we selected these three, which showed severe heterogeneity of variance as compared with the other trials. All three trials were laid out according an a-lattice design. The trials are from the low-moisture-stress mega-environment in Ethiopia, were performed to introduce and test adaptation of drought-tolerant maize, and were conducted in 2011 (EVDC11A, evaluation of drought-tolerant varieties in the 2011 trial season), 2012 (EITCHYB12, evaluation of intermediate top-cross hybrids in 2012), and 2014 (ENHNVT14B, evaluation of normal maize hybrids under national variety trials). Trial EVDC11A had 46 early-maturing maize crosses, laid out in 92 plots of size 6.375 m2. There were two complete replicates, each laid out in two rows and 23 columns. Each row corresponds to an incomplete block. Trial EITCHYB12 was performed in the main rainy season of the year 2012. This experiment had 56 genotypes planted in 112 plots of 7.875 m2.

It had two complete replicates each laid out in four rows and 14 columns per each row. Each replicate had eight incomplete blocks of size seven. The third trial, ENHNVT14B, comprised 32 early normal hybrid maize lines, planted in 64 plots of 6.3 m2. This trial also had two complete replicates, where each replicate had one row (rows correspond to replicates), 32 columns, and two incomplete blocks of size 16. In all three datasets, columns ran parallel to the direction of maize rows. In all cases, the shape of incomplete blocks was rectangular with several columns and one row.

2.2.2 Statistical methods

Baseline model

The randomization-based model for an α-lattice design is used as the baseline model. The randomization-based model is a design-based model defined by the group of permutations underlying the randomization (Bailey and Brien, 2016). The model can be written as

$$y_{ijh} = \mu + \tau_i + \gamma_j + b_{jh} + e_{ijh}\,, \tag{1}$$

where y_{ijh} is the observed yield of the *i*-th genotype in the *j*-th replicate and *h*-th block, μ is an overall intercept, τ_i is the fixed main effect of the *i*-th genotype, γ_j is the fixed effect of the *j*-th complete replicate, b_{jh} is the random effect of the *h*-th block nested within the *j*-th complete replicate, and e_{ijh} a residual effect corresponding to y_{ijh}.

To assess the assumption of constant variance, plots of studentized residuals versus predicted values were scrutinized. If the constant error variance assumption holds true, this plot should show a horizontal band with constant variability along the vertical axis across predicted values (Atkinson, 1985; Carroll and Ruppert, 1988; Montgomery et al., 1992). Any departure from this expected pattern, e.g., an increase of the variance with the mean (out-ward opening funnel) or a variance increase as the mean decreases (in-ward opening funnel), indicates violation of the constant-variance assumption and suggests the need for remedies such as transformation and variance modelling. Therefore all checking of assumptions is based on a visual assessment of residual plots. We prefer this approach to significance testing of assumptions (Kozak and Piepho, 2018).

Remedial measures for variance heterogeneity

Box-Cox transformation

Box and Cox (1964) consider a parametric family of nonlinear transformations given by

$$y(\lambda) = \begin{cases} \dfrac{y^{\lambda} - 1}{\lambda} & \text{for} \quad \lambda \neq 0 \\ & , \\ \log(y) & \text{for} \quad \lambda = 0 \end{cases} \qquad (2)$$

where λ is a transformation parameter to be determined from the data. For example, the square root and cube root transformation correspond to λ values of 1/2 and 1/3, respectively. If the transformation parameter takes the value $\lambda = 1$, this indicates that no transformation is needed. The best value of λ is estimated by maximum likelihood (ML), assuming normality and a model with constant variance on the transformed scale, through a grid search over a range of values for λ. The need of the transformation for the data is tested by a likelihood ratio test comparing the deviance for the optimally transformed and the untransformed data (Piepho and Ogutu, 2003; Piepho, 2009). A SAS macro based on the MIXED procedure was used to determine λ by the ML method (Piepho, 2009). Since the method assumes homogeneity of variance on the transformed scale, the Box-Cox transformation usually also stabilizes the variance, in addition to achieving approximate normality.

Variance modelling

Another approach to account for variance heterogeneity is to model the variance as a function of the mean, leaving the data untransformed. Our preliminary study of residuals suggested that the POM and exponential models could be used. In the POM model, the variance is assumed to be proportional to the power of the mean, whereas in the exponential model the variance is assumed to be an exponential function of the mean. The general variance model as a function of the mean can be written as

$$Var\left(y_{ijh} \mid E\left[y_{ijh}\right]\right) = \sigma^2 V\left(E\left[y_{ijh}\right]\right), \tag{3}$$

where σ^2 is unknown scale/variance parameter, $E\left[y_{ijh}\right] = \mu + \tau_i + \gamma_j + b_{jh}$ is the conditional expected value of y_{ijh}, given the effects for treatments, replicates, and incomplete blocks and $V\left(E\left[y_{ijh}\right]\right)$ is the variance function, which is equal to $V_{P\_ijh} = \left|E\left[y_{ijh}\right]\right|^{\theta_1}$ in the case of the POM model, and $V_{E\_ijh} = \exp\left(\theta_2 E\left[y_{ijh}\right]\right)$ for the exponential function. Note that V_P_ijh and V_E_ijh represent the POM and exponential variance functions, respectively. The parameters θ_1 and θ_2 are the variance function parameters to be estimated, and $E\left[y_{ijh}\right]$ is the mean corresponding to y_{ijh}. The POM variance function is also a characteristic of the Tweedie family of distributions. Tweedie distributions are families of exponential dispersions used to model responses with non-negative values using generalized linear models (Tweedie, 1947, 1984; Jørgensen, 1987; Peel et al., 2012; Wood and Fasiolo, 2017). For the POM model, $\theta_1 = 0$ corresponds to a Normal distribution with constant variance, $\theta_1 = 1$ corresponds to a Poisson distribution, and $\theta_1 = 2$ corresponds to a Gamma or lognormal distribution. For the exponential model, $\theta_2 = 0$ also corresponds to the homogeneous-variance model. Usually, the values of the variance parameters θ_1 and θ_2 are not known and are estimated from the data, e.g., by ML or by restricted maximum likelihood (REML) (Carroll and Ruppert, 1988).

Modelling spatial variability in mixed linear models

Gilmour et al. (1997) identify three major components of spatial variation in a field experiment that they denote as natural or local, extraneous and global. The extraneous and global components are accounted for through the block, row and column effects. For the local trend, the residual e_{ijh} in equation (1) can be decomposed as $e_{ijh} = \eta_{ijh} + \varepsilon_{ijh}$, where η_{ijh} represents the local trend and ε_{ijh} is the remaining error. Collecting the plot errors e_{ijh} into a vector e, and the random block effects into a vector u, we may represent the residual variance by $Var(e) = R$, which will be needed later when introducing heterogeneity. The random block effects u and residuals e_{ijh} (η_{ijh} and ε_{ijh}) are assumed to be mutually independent and each have mean zero and constant variance.

Finding the best spatial model requires fitting different models for each individual trial and selecting the one that best fits the data, the reason being that it is impossible to find one model that is efficient and appropriate for all trials (Gilmour et al., 1997; Piepho and Williams, 2010). Over-fitting is a main problematic issue when various models are to be tried, therefore models should be selected strategically (Burnham and Anderson, 1998). One suggestion to avoid over-fitting is first to model the data with the randomization-based model (baseline-model), and then to extend this by adding spatial model components only when this improves the fit (Williams, 1986; Williams et al., 2006; Piepho and Williams, 2010; Piepho et al., 2011).

One-dimensional isotropic model for local trend

In a time series context, a first-order autoregressive (AR) model is a representation of the dependent variable as a linear combination of its previous values. Its spatial version is a representation of data at location *l* as a linear function of nearest neighbor values. The AR model can be fitted in one dimension, i.e., by assuming the correlation exists either along rows or along columns. Models for the local trend can also be fitted using the exponential, spherical, and Gaussian models which in their basic form are isotropic, i.e., the spatial variation among observations depends only on the distance between them (Schabenberger and Pierce, 2002; Schabenberger and Gotway, 2005). Each isotropic spatial model has three variance parameters called sill, range, and nugget. We fitted isotropic one-dimensional models for an AR, exponential, spherical, and Gaussian model assuming correlation exists only along rows, but different rows were independent.

Two-dimensional anisotropic model for local trend

When fitting spatial models across two dimensions, one must cater for situations where the variation among observations depends both on the distance as well as directions; this phenomenon is called anisotropy (Schabenberger and Gotway, 2005). Geometric anisotropy is a simple form of anisotropy which occurs when the semivariogram range differs between directions, and this can be defined, e.g., for the exponential, spherical and Gaussian models (Gleeson and Cullis, 1987; Cressie, 1991; Zimmerman and Harville, 1991). In addition to the three parameters sill, range and nugget for isotropy model, geometric anisotropy models require two additional parameters which are anisotropy angle and anisotropy ratio. Geometric anisotropy can be reduced to an isotropic model by a linear transformation of the coordinate system (Schabenberger and Gotway, 2005). The two-dimensional autoregressive model AR(1)×AR(1) is another type of anisotropic model (Gilmour et al., 1997). In this study we consider the geometric exponential, spherical, and Gaussian and the two-dimensional AR(1)×AR(1) anisotropic models which can be fitted using the MIXED procedure of SAS. We fitted all these anisotropic models assuming correlation exist across the whole field and across replicates.

Joint modelling of spatially correlated and heteroscedastic errors

To model the spatial correlation along with the heterogeneity of variance the variance-covariance structure R of plot errors can be formulated as follows (Carroll and Ruppert, 1988):

$$R = R_M^{1/2} A R_M^{1/2}, \tag{4}$$

where A represents the spatial correlation matrix and R_M is a diagonal matrix whose diagonal elements are $\sigma^2 V(E[y_{ijh}])$, which is the variance function for the *ijh*-th observation, where $E[y_{ijh}]$ is the expected value of y_{ijh}.

There are several methods of variance function estimation, i.e., methods to estimate parameters σ^2 and θ_1 or θ_2. Pseudo-likelihood estimation is one of the standard methods (Carroll and Ruppert, 1988) and it is based on the idea that the conditional expected value of y_{ijh} can be replaced by the current estimate, possibly obtained from unweighted generalized least squares methods. Using the current estimate of the linear predictor, the variance parameters are then estimated using likelihood methods. The pseudo-likelihood estimation technique depends on the mean-variance relationship; it does not make other parametric assumptions (Carroll and Ruppert, 1988). In this study we apply the pseudo-likelihood method assuming additionally that our data follow a normal distribution. For scalar variance parameters θ_1 and θ_2 of the POM and exponential variance models respectively, the pseudo-likelihood estimate can be computed using a grid search approach in a reasonable range of values, where the optimal value of θ (θ_1 or θ_2) is chosen to be the value of the parameters which maximizes the likelihood over the grid. To obtain an efficient estimate an iteration process has been suggested, requiring at least two iterations (Carroll and Ruppert, 1988). The usual pseudo-likelihood method is based on ML estimation but this method does not account for the loss of degrees of freedom due to estimating the fixed effects. However, REML can be used to account for the bias, leading to a residual pseudo-likelihood approach. For given values of the variance parameters θ_1 and θ_2 for the variance functions V_{P_ijh} and V_{E_ijh}, weights $w_{V\_Pijh} = 1/V_{p\_ijh}$ and $w_{V\_Eijh} = 1/V_{E\_ijh}$ can be used to fit the POM and exponential models along with a spatial correlation model. Since the variance of an observation is defined as the product of the inverse of the weights and the scale parameter σ^2, the weights need to be standardized so that the scale parameter is identifiable. One possible standardization method is to divide each weight by mean of all weights ($\overline{w}_{V_P}$ and $\overline{w}_{V_E}$, for POM and exponential models, respectively), so that the mean of standardized weight variable equals one. Thus, the standardized weights for the *ijh*-th observations are $w_{zV_P(ijh)} = 1/\left(\overline{w}_{V_P} V_{P\_ijh}\right)$ and $w_{zV_E(ijh)} = 1/\left(\overline{w}_{V_E} V_{E\_ijh}\right)$ for the POM and exponential models, respectively. With these standardized weights, the variance functions become $V(E[y_{ijh}]) = 1/w_{zV_P(ijh)}$ and $V(E[y_{ijh}]) = 1/w_{zV_E(ijh)}$, respectively.

In our experience, if the random inter-block variance does not converge to zero and we do not make it proportional to the residual variance, this will likely result in convergence problems. Thus, we assume that block and error variances are proportional to one another. This assumption also seems quite natural and plausible, because if the error variance is a function of the mean then so should be the block variance. In order to let the variance of a random effect u (i.e. the block in our example) be proportional to the residual variance, we extended the random-effect model by multiplying the random effect u by the square root of the *ijh*-th value of the variance function, denoted here as $s_{(VP\_ijh)} = \sqrt{1/w_{zV_P(ijh)}}$ and $s_{(VE\_ijh)} = \sqrt{1/w_{zV_E(ijh)}}$, for the POM and the exponential model, respectively. For the random effect u with a constant variance σ_u^2, we then have

$\mathrm{var}(s_{(VP\_ijh)}u) = V(E[y_{ijh}])\sigma_u^2 = s_{(VP\_ijh)}^2\sigma_u^2$ and

$\mathrm{var}(s_{(VE\_ijh)}u) = V(E[y_{ijh}])\sigma_u^2 = s_{(VE\_ijh)}^2\sigma_u^2$ for the POM and the exponential model, respectively. We developed a SAS macro called ***%fit_variance_function*** to estimate all the parameters needed to fit the POM and exponential variance models along with the spatial models, using a grid search procedure as detailed in the Appendix. For all three datasets the restricted pseudo-likelihood estimate for θ_1 and θ_2 was computed on the grid of values with bounds chosen so that the optimal value was found within these bounds. We used a step size of 0.1 for both the POM and exponential models. With all models we use two iterations, which is the minimum suggested by Carroll and Ruppert (1988).

To choose the best fitting model we use the Akaike information criterion (AIC), with smaller value indicating better fit (Akaike, 1974). However, to choose the optimal value of θ from the range of θ values for a given model we use the deviance (-2 times residual log-likelihood). The reason for using the deviance rather than the AIC is that for a given model (spatial or randomization-based) the number of parameters does not change as we screen different values of θ. The preferred best model with the optimal θ is the one with the smallest deviance.

Moreover, if the estimate of a spatial covariance parameter of a model is close to zero, and the model has convergence problems, we do not report that model. Our modeling approach is based on the algorithm presented in Figure 1.

Algorithm:

```
1.  Choose grid bounds $\theta_{\min}$, and $\theta_{\max}$, step-size $\Delta\theta$, and number of iterations n
2.  Analyse the data assuming there is no relation between means and variances
3.   For $\theta = \theta_{\min}$, to $\theta = \theta_{\max}$ by $\Delta\theta$ do
         For $j$=1 to n do
             3.1    Calculate weight $w_i$ for $\theta_i$, using current mean estimates (from Step 2 when $j$=1
                    and from Step 3.2 of previous iteration for j>1)
             3.2     Fit model using weight from Step 3.1 to obtain new estimates of means
         End
     End
4.   Choose the optimal value of the variance parameter θ with lowest deviance from grid of values $\theta_i$
     tried in Step 3
```

Fig. 1: General algorithm for fitting variance model either with the baseline or spatial models

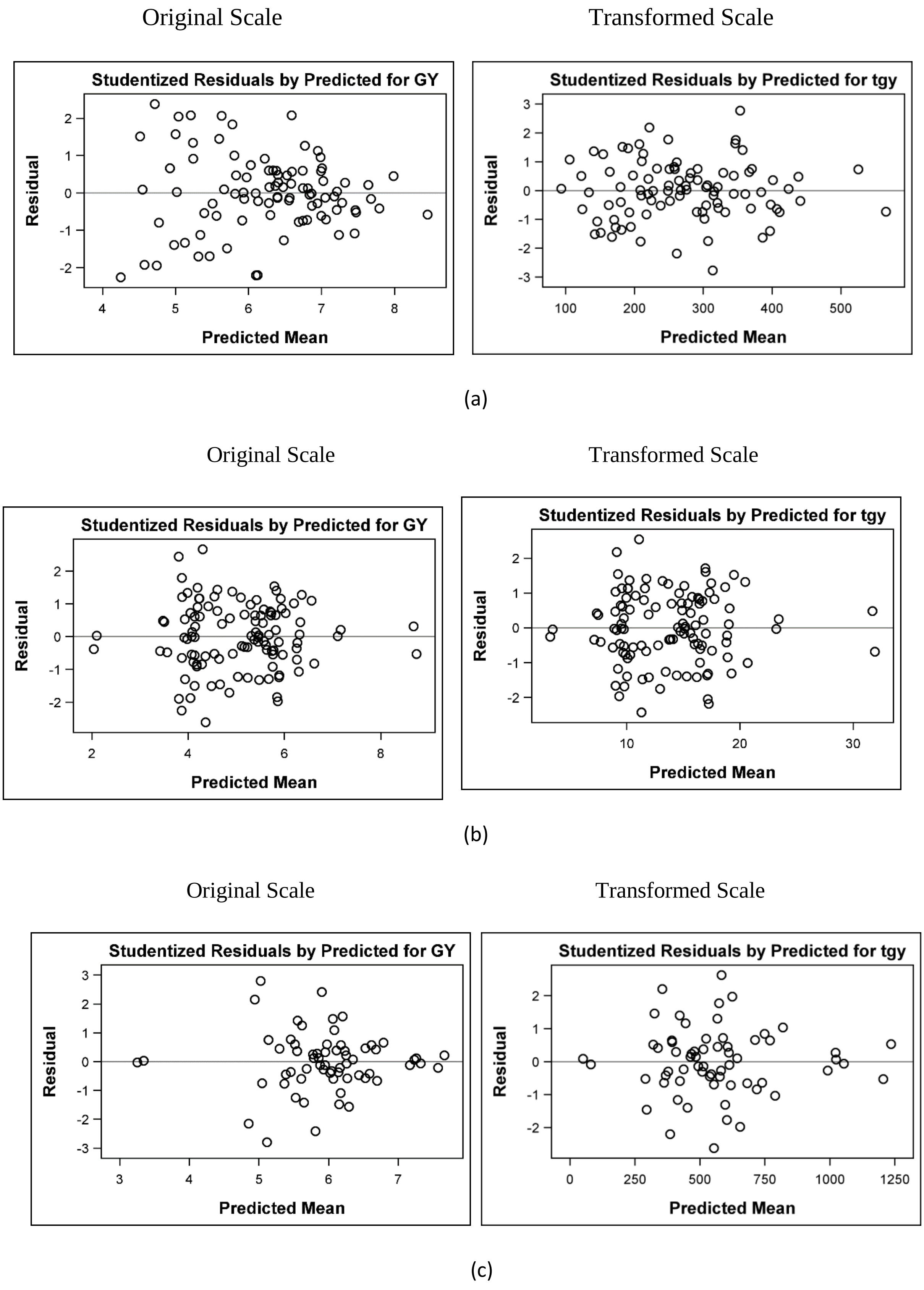
2.3 Results

2.3.1 Preliminary checking of assumptions

Inspection of the studentized residuals from fitting of the baseline model for the three data sets indicates that all three datasets violate the constant variance assumption, i.e., the variance decreases as the mean increases (e.g., Fig. 2, left side).

2.3.2 Box-Cox transformation

For all datasets application of the Box-Cox transformation fixes the problem of variance heterogeneity as can be seen in the residual plots (Fig. 2). Note that inspection and comparison of residual plots is always somewhat subjective. In our examples, transformation reduced heteroscedasticity, especially in dataset EVDC11A. The difference in the residual plots before and after transformations are not very pronounced for Figs. 2b and 2c, but we still believe that there is improvement after transformation. The Box-Cox transformation parameters based on the baseline models estimated by ML are 3.0, 1.6, and 3.5 for trials EVCD11A, EITCHYB12 and ENHNVT14B, respectively. In all three datasets the drop in deviance is significant as compared to the model with untransformed data (Table 1). Even though the Box-Cox transformation is moderately successful in stabilizing the variance, in all three cases the transformation parameter λ is not in the interval between 0 and 1 (Table 1), in which case back-transformation to the expected value on the original scale is impractical (Freeman and Modarres, 2006). Because of this limitation, the alternative option of modeling variance heterogeneity on the observed original scale is considered (Carroll and Ruppert, 1988).



(a)

(b)

(c)

Fig.2. Plots of the studentized residual versus predicted mean for grain yield in the original scale (GY, tonnes per hectare) (left side) and transformed grain yield (TGY) in the Box-Cox

transformed scale (right side). (a) EVDC2011A maize trial data, (b) EITCHYB2012 maize trial data and (c) ENHNVT2014B maize trial data.

Table 1. Values of transformation parameter λ and deviance of the Box-Cox transformation for baseline model.

Trial Name	Lambda λ	Deviance for untransformed data	Deviance for transformed data	Drop in deviance
2011/EVDC11A	3.0	235.5	213.675	21.825*
2012/EI-TCHYB	1.6	227.1	223.118	3.982*
2014/ENHNVT14B	3.5	82.3	72.755	9.545*

λ is an optimal value of the Box-Cox transformation parameter determined from the data using ML.

* Significant based on $\chi^2_{\alpha=0.05;\, df=1} = 3.84$.

2.3.3 Modelling the variance heterogeneity along with spatial structure

Example 1- Trial EVDC11A

The baseline model (independent and constant errors model) with random block effects was extended to allow for spatially correlated errors with and without the variance model. The geometrically anisotropic spherical model with correlation across the replicates and without nugget for the POM variance model was found to be the best. Comparing a spatial model with homogeneous variance and the same spatial models with heterogeneous variance, the spatial model with heterogeneous variance had the best AIC among all of the fitted models (100%) for both the POM as well as the exponential model. Comparing the variance models, the POM model had smaller AIC for 81.8% of the fitted models and the exponential model fits better than POM only for about 18.2% of the fitted models (Table 2).

Table 2: Deviance values and optimal θ's of EVDC11A maize trial dataset for the baseline and spatial models assuming homogeneous variance and using POM and exponential variance modeling.

Models	Range of correlation	Nugget	Homogeneous variance	Variance model			
				POM††		Exponential	
			Deviance	RLL	θ_1	Deviance	θ_2
Baseline			182.3	171.1	-4.7	170.2	-0.8
AR1	block	No	172.5	164.2	-3.4	165.1	-0.6
AR1	block	Yes	172.2	164.6	-3.4	165.3	-0.5
Exp	block	No	172.5	164.2	-3.4	165.1	-0.6
Exp	block	Yes	172.2	164.6	-3.4	165.3	-0.5
Gau	block	No	174.6	166.0	-3.1	167.3	-0.5
Gau	block	Yes	172.4	165.2	-3.1	166.0	-0.5

Sph	block	No	172.2	159.5	-3.6	165.7	-0.6
Sph	block	Yes	172.3	163.7	-3.3	164.4	-0.6
AR1	column	No	182.0	168.8	-4.5	168.3	-0.8
AR1	column	yes	182.0	168.8	-4.6	168.2	-0.8
AR1×AR1	whole field	No	172.5	163.4	-3.3	164.5	-0.5
AR1×AR1	whole field	yes	172.0	163.6	-2.9	164.4	-0.5
Expga	whole field	No	172.2	154.3	-5.1	155.2	-0.7
Expga	whole field	yes	-	-	-	-	-
Gauga	whole field	No	171.4	163.5	-3.1	165.0	-0.5
Gauga	whole field	yes	169.6	-	-	-	-
Sphga	whole field	No	180.2	162.9	-3.3	156.3	-0.7
Sphga	whole field	yes	-	-	-	-	-
AR1×AR1	replicates	No	171.7	164.6	-3.5	165.3	-0.6
AR1×AR1	replicates	yes	171.6	164.5	-3.4	165.2	-0.6
Expga	replicates	No	169.1	162.7	-2.8	163.2	-0.4
Expga	replicates	yes	-	-	-	-	-
Gauga	replicates	No	168.8	159.7	-2.8	160.3	-0.5
Gauga	replicates	yes	-	155.2	-3.4	155.5	-0.5
Sphga	replicates	No	170.7	153.7	-5.0	159.0	-0.3
Sphga	replicates	yes	NC	NC	NC	NC	NC

†† POM, power of the mean variance model; NC, not converged; AR1, one-dimensional autoregressive model; dashed line (-), are models with no valid fit for one or more of the variance component(s) because they are estimated to be either zero or in the border and Hessian matrix is not positive definite; AR1×AR1, two-dimensional anisotropic autoregressive model; Exp, one-dimensional isotropic exponential model; Gau, one-dimensional isotropic Gaussian model; Sph, one-dimensional isotropic spherical model,

Expga, two-dimensional geometric anisotropic exponential model; Gauga, two-dimensional geometric anisotropic Gaussian model; Sphga, two-dimensional geometric anisotropic spherical model.

Example 2 - EITCHYB12

The AIC (Table 3) reveals that the baseline model with the exponential variance performed better than the other models. Among the fitted spatial models, for both the POM and the exponential variance model, 100% of the spatial models had a smaller AIC than the same spatial model with homogeneous variance. When comparing the two variance models, the exponential model was better than the POM for 87.5% of the fitted models, while the POM model performed better than the exponential model for only 12.5% of the models.

Table 3: Deviance values and optimal θ's of EITCHYB12 maize trial data for the baseline and spatial models assuming homogeneous variance and using POM and exponential variance modeling.

			Homogeneous	Variance model			
	Range of		variance	POM††		Exponential	
Model	correlation	nugget	Deviance	Deviance	θ_1	Deviance	θ_2
Baseline			193.1	190.6	-1.8	189.8	-0.4
AR1	block	no	192.8	189.1	-2.3	188.6	-0.5
AR1	block	yes	-	-	-	-	-
Exp	block	no	-	-	-	-	-
Exp	block	yes	-	-	-	-	-
Gau	block	no	193.1	190.6	-1.8	189.8	-0.4
Gau	block	yes	193.1	-	-	-	-
Sph	block	no	-	-	-	-	-

Sph	block	yes	199.1	-	-	-	-
AR1	column	no	191.9	190.2	-1.7	189.5	-0.4
AR1	column	yes	190.0	188.4	-1.5	187.8	-0.4
AR1	row	no	193.1	190.1	-2.2	189.4	-0.5
AR1	row	yes	-	-	-	-	-
Exp	row	no	-	-	-	-	-
Exp	row	yes	192.6	-	-	-	-
Gau	row	no	-	190.6	-1.8	-	-
Gau	row	yes	-	-	-	-	-
Sph	row	no	-	-	-	-	-
Sph	row	yes	192.8	189.9	-1.4	189.3	-0.3
AR1×AR1	whole field	no	191.7	-	-	-	-
AR1×AR1	whole field	yes	190.0	-	-	-	-
Expga	whole field	no	191.7	187.4	-1.1	189.8	-0.3
Expga	whole field	yes	188.3	-	-	184.7	-0.5
Gauga	whole field	no	201.1	-	-	-	-
Gauga	whole field	yes	198.7	-	-	-	-
Sphga	whole field	no	-	-	-	-	-
Sphga	whole field	yes	-	-	-	188.4	-0.2
AR1×AR1	replicates	no	191.9	-	-	-	-
AR1×AR1	replicates	yes	188.6	NC	NC	NC	NC
Expga	replicates	no	191.8	190.4	-1.4	-	-
Expga	replicates	yes	188.3	-	-	-	-
Gauga	replicates	no	-	-	-	-	-

Gauga	replicates	yes	187.4	186.5	-0.7	185.7	-0.4
Sphga	replicates	no	-	-	-	-	-
Sphga	replicates	yes	189.3	188.0	-1.5	188.2	-0.2

†† POM, power of the mean variance model; NC, not converged; AR1, one-dimensional autoregressive model; dashed line (-), are models with no valid fit for one or more of the variance component(s) because they are estimated to be either zero or in the border and Hessian matrix is not positive definite; AR1×AR1, two-dimensional anisotropic autoregressive model; Exp, one-dimensional isotropic exponential model; Gau, one-dimensional isotropic Gaussian model; Sph, one-dimensional isotropic spherical model, Expga, two-dimensional geometric anisotropic exponential model; Gauga, two-dimensional geometric anisotropic Gaussian model; Sphga, two-dimensional geometric anisotropic spherical model.

Example 3 - ENHNVT14B

For this dataset, the one-dimensional autoregressive model with correlation across column, without nugget and exponential variance, was found to be the best model. All the spatial models with the heterogeneous variance outperformed the same spatial model with homogeneous variance for both POM and exponential models (Table 4). As regards the variance model comparisons, 100% of the exponential models had the smaller AIC compared to the corresponding POM models.

Table 4: Deviance values and optimal θ's of ENHNVT14B maize trial dataset for the baseline and spatial models assuming homogeneous variance and using POM and exponential variance modelling.

Model	Range of correlation	Nugget	Homogeneous variance	Variance model			
				POM††		Exponential	
			Deviance	Deviance	θ_1	Deviance	θ_2
Baseline			87.3	80.4	-7.4	78.9	-1.4

AR1	Block	No	86.8	81.0	-6.0	79.6	-1.2
AR1	Block	Yes	-	80.8	-6.7	79.5	-1.3
Exp	Block	No	87.3	-	-	79.7	-1.2
Exp	Block	Yes	-	80.8	-6.8	79.5	-1.3
Gau	Block	No	87.3	80.9	-6.1	79.5	-1.2
Gau	Block	Yes	-	80.3	-7.8	78.5	-1.2
Sph	Block	No	-	-	-	-	-
Sph	Block	Yes	-	-	-	-	-
AR1	Column	No	84.8	74.6	-6.6	72.3	-1.4
AR1	Column	Yes	84.8	-	-	-	-
AR1	Row	No	86.8	80.7	-6.7	79.4	-1.3
AR1	Row	Yes	-	80.4	-7.5	79.3	-1.3
Exp	Row	No	-	80.7	-6.6	79.4	-1.3
Exp	Row	Yes	-	80.4	-7.4	79.3	-1.3
Gau	Row	No	-	80.7	-6.7	79.3	-1.3
Gau	Row	Yes	-	80.6	-8.0	79.1	-1.3
Sph	Row	No	-	-	-	-	-
Sph	Row	Yes	-	-	-	-	-
AR1×AR1	whole field	No	-	74.4	-7.3	72.6	-1.4
AR1×AR1	whole field	Yes	-	-	-	70.6	-1.8
Expga	whole field	No	-	73.7	-6.3	72.5	-1.1
Expga	whole field	Yes	-	-	-	-	-
Gauga	whole field	No	83.9	NC	NC	73.5	-1.3
Gauga	whole field	Yes	-	-	-	-	-

Sphga	whole field	No	-	-	-	-	-
Sphga	whole field	Yes	-	-	-	-	-
AR1×AR1	replicates	No	-	-	-	-	-
AR1×AR1	replicates	Yes	-		NC	NC	NC
Expga	replicates	No	-	-	-	-	-
Expga	replicates	Yes	-	-	-	-	-
Gauga	replicates	No	-	-	-	-	-
Gauga	replicates	Yes	-	-	-	-	-
Sphga	replicates	No	-	-	-	-	-
Sphga	replicates	Yes	-	-	-	-	-

†† POM, power of the mean variance model; NC, not converged; AR1, one-dimensional autoregressive model; dashed line (-), are models with no valid fit for one or more of the variance component(s) because they are estimated to be either zero or in the border and Hessian matrix is not positive definite; AR1×AR1, two-dimensional anisotropic autoregressive model; Exp, one-dimensional isotropic exponential model; Gau, one-dimensional isotropic Gaussian model; Sph, one-dimensional isotropic spherical model, Expga, two-dimensional geometric anisotropic exponential model; Gauga, two-dimensional geometric anisotropic Gaussian model; Sphga, two-dimensional geometric anisotropic spherical model.

2.4 Discussion

The analysis of data with linear mixed models is usually based on the assumption of homogeneity of error variance and this assumption may be violated. Techniques used when such problems are encountered fall into two broad categories: weighting and data transformation. Weighting can be performed after determining the weights using an appropriate variance model. The aim of this study was to explore the application of the Box-Cox transformation and variance modeling as means of possible remedy when the constant variance assumption does not hold true in field trials. It has been shown for three examples that both Box-Cox transformation and modeling heterogeneous variance resulted in a better model fit than the homogeneous variance model. For instance, after applying the variance model, the variance of studentized residuals became constant for all three trials (Fig. 3).

If variance heterogeneity is observed, fitting a model which assumes homogeneous variance can be expected to be inefficient. In a two-stage analysis of MET, the variances are assumed to be heterogeneous between trials. This variance heterogeneity can be handled in the second stage by using proper weighting techniques of the variance estimate from the first stage. The REML estimates of the variance and co-variances of adjusted genotypes from the first stage are used to compute the weights in the second stage (Damesa et al., 2017), meaning that taking the remedial action for within-trial variance heterogeneity will improve the estimates of the weights for the second-stage analysis. If the variance heterogeneity is ignored, then the error will be twofold in case of two-stage analysis. This is because, firstly, estimation of adjusted genotype means in the first stage will not be efficient and, secondly, the variance-covariance matrix which is supposed to be used for determining the weights in the second stage analysis, will be wrongly specified, thus producing inefficient mean estimates and biased standard errors in the second stage. The same problems would come into play in a single-stage analysis.

There are several reasons which make the Box-Cox transformation attractive for practical analysis, some of which are its efficiency and its flexibility, its being a generalization of the most common transformations such as logarithmic, square root, cubic root, quadratic, and also its applicability in mixed models (Piepho, 2009). Besides all of these advantages, the difficulty in reporting on the original scale, or interpreting the meaning of the estimated fixed effect parameters in the transformed scale is its major drawback, particularly in MET analysis.

In variance modelling, it is possible to assume a distribution family has a constant coefficient of variation. The gamma and the lognormal are the two most commonly used distributions for this strategy, and both have variance function corresponding to $\theta_1 = 2$ in the POM model. However, such a prior distributional assumption may not be appropriate, and it is recommended to estimate the mean variance relating parameter from the data (Carroll and Ruppert, 1988).

In many cases, when non-constant variance occurs, usually the variance increases as the mean increases. However, in all three datasets in this study variance and mean are inversely related. This could be the result of different degrees of response of varieties to moisture stress that is prevalent in the low-land mega-environments studied here. Since this study is a drought tolerance evaluation, low yield can be an indicator for susceptible varieties whereas high yield is an indicator for drought tolerance varieties. Therefore the inverse variance-mean relationship could also be interpreted as implying that the variances are larger for varieties that are sensitive to drought and smaller for varieties that are resistant to drought (Fig. 2, left side). However, as we only have the yield data at our disposal for these trials, it is not possible to provide a more detailed biological explanation for the observed negative relationship. Other causes could be unstudied environmental factors (abiotic and biotic) like variation in soil (moisture content, fertility) or attack by insects or animals (Gomez and Gomez, 1984). If the heterogeneous variance between entries is the cause, one solution to handle this problem is to use a model with heterogeneous variance between entries. However, in all three datasets each entry is replicated only twice, and therefore the estimate of variance for each entry would be a poor estimate of variance. Generally with only two or three replications, the estimate of variance will be very inaccurate (Carroll and Ruppert, 1988). For example, in our attempts to fit a model with entry-specific variance, there were computational problems for all of the three example datasets (result not shown), indicated by an infinite likelihood at the initial iteration or a non-positive definite Hessian matrix. Edwards and Jannink (2006) proposed Bayesian hierarchical models to address the problems of estimation of variance from a small number of observations per treatment. They found that Bayesian methods can improve estimation through borrowing of information from neighboring observations and allow accounting for heterogeneity of error variances when variance is estimated from few observations per treatment. This is an interesting alternative to the variance modelling proposed in our paper. In fact, one could combine both approaches and do the variance-mean modelling in a Bayesian framework.

Usually, when the variance and mean are inversely related, for the Box-Cox method a transformation parameter value $\lambda > 1$ stabilizes the variance, whereas for the exponential and POM models variance parameters $\theta < 0$ help to appropriately down-weight portions of the data which are highly variable and extract more information from portions of the data that are more precise. However, the optimal value of the variance parameter should be determined by an appropriate estimation method.

Accounting for variance heterogeneity and correlation of neighboring plots simultaneously can be necessary in analysis of agricultural field trials. Both issues can be resolved jointly using a suitable modelling approach, as we have demonstrated in this study. According to randomization theory, ignoring the correlation of neighboring plots is not a mistake, because randomization breaks any spatial dependencies (Piepho et al., 2013); however, modelling spatial correlation can be an opportunity to improve the precision of the analysis.

For this particular study the variance modeling remedy works well for the variance heterogeneity problem, but this may not generally be the best solution for other studies. Therefore, considering both a Box-Cox transformation and variance modelling, and possibly other options on a case-by-case basis is generally advisable.

Care should be taken in applying variance models for small sample sizes. We do not recommend this approach when the number of varieties is too small. As far as we know there is no clear threshold value for the minimum number of observations for variance model. Bootstrapping and simulation have been suggested for identifying the smallest sample size required for variance modelling (Carroll and Ruppert, 1988), but we have not pursued this.

One of the challenges in fitting spatial and variance models in SAS PROC MIXED is the lack of an option to specify them together in the residual variance-covariance matrix R. A possibility that we considered was using the GLIMMIX procedure with a user-defined variance function and estimating the variance parameters using a pseudo-likelihood estimation method, searching over a grid of values, to optimize the whole model, but this approach was not functional due to persistent convergence problems. We therefore implemented our proposed procedure from scratch in the macro ***%fit_variance_function***.

As we have seen from the three datasets, the spatial models were best when combined with the heterogeneous variance model. In two of the dataset the best model was from the POM variance model, whereas for the third dataset the exponential model was the best variance model. Which variance model needs to be chosen obviously depends on the type of data, so fitting a number of promising candidate models and then selecting the best among them will usually be necessary.

2.5 Appendix

The macro ***%fit_variance_function***

This macro estimates the variance function parameters θ_1 and θ_2 for the POM and exponential model, respectively, from the data set using a grid search approach over a range of given values. The macro involves two calls of the PROC MIXED procedure. In the first step, the baseline or spatial model is fitted and the conditional predicted values for each observation are saved. Next, using the predicted values and a range of different values for θ, a range of different weights are computed for each theta. Weights are standardized before they are submitted to the second PROC MIXED call, in which the variance is modeled as a weight for the baseline or spatial model. Conditional predicted values for each θ are saved to replace predicted values from the former PROC MIXED call. The second step is repeated over the grid of values for θ_1 and θ_2 for *n* iterations. Separate analyses are performed for each value of θ_1 and θ_2 over the grid. The optimal value of θ_1 or θ_2 is the one which results in the minimum deviance. The number of steps for the grid search as well as the limits of the grid can be chosen by the user.

2.6 Supplementary material

Supplementary materials comprising the SAS macro ***%fit_variance_function.sas*** and SAS example files (Examples power of the mean variance model.sas, Examples exponential variance model.sas and Examples homogeneous variance model) are available online.

Chapter 3

One Step at a Time: Stage-Wise analysis of series of experiments

Tigist Mideksa Damesa[1], Jens Möhring[1], Mosisa Worku[2], Hans-Peter Piepho[1*]

[1]Biostatistics Unit, Institute of Crop Science, University of Hohenheim, Fruwirthstrasse 23, 70599 Stuttgart, Germany.

[2]International Maize and Wheat Improvement Center (CIMMYT), Nairobi, Kenya.

Published in Agron. J. 109:845–857 (2017)
The original publication is available at

doi:10.2134/agronj2016.07.0395

3.1 Abstract

Multi-environment trials (MET) can be analyzed using single-stage or stage-wise analysis. Single-stage analysis is fully efficient, meaning that the estimators can be expected to be as close as possible to the corresponding true genotypic values, and so is often deemed preferable to two-stage analysis. However, two-stage analysis is often favored in practice over single-stage analysis in case of large datasets because of the larger computational burden of the latter and because the former allows separate analyses of individual trials in the first stage accounting for any specifics of each trial. In this study we demonstrate the similarities of results of single-stage and two-stage analysis when information on mean estimates and the associated variance-covariance matrix is forwarded from the first stage to the second stage using four examples with maize (*Zea mays* L.) trial data from Ethiopia. A new fully efficient and an approximate two-stage method with diagonal weighting matrix are used for weighting in the second stage. We extend the method to three-stage analysis for MET when sites are stratified by agro-ecological zones and demonstrate how to obtain best linear unbiased predictions (BLUP) of genotype effects per zone using the information from neighboring zones. Two macros which compute weights for use in the fully efficient and diagonal weighting approaches are provided.

Many trials are replicated in multiple environments in order to broaden the inference space. For example, plant breeding and variety trials are typically performed at multiple sites and in several years (Yates and Cochran, 1938; Cochran, 1937; Comstock and Moll, 1963; Gauch, 1992; Talbot, 1997). A joint analysis of such multi-environment trials (MET) can be done in a single stage by a linear mixed model (LMM) for the plot data (Smith et al., 2001, 2005). Such an analysis is commonly considered to be fully efficient because all sources of variation can be accounted for simultaneously in a single model and the analysis provides best linear unbiased estimates (BLUE) of all fixed effects, as well as best linear unbiased predictions (BLUP) of all random effects under that assumed single-stage model (Searle et al., 1992). An alternative method of analysis is to proceed in two stages, where in the first stage genotype means are computed per trial and in the second stage genotype means from all trials are subjected to a joint analysis. In principle, the stage-wise approach can also be extended to more than two stages (Piepho et al., 2012a).

In both cases, individual trials are first analyzed separately, paying due attention to all specifics of a trial, including outlier detection, the particular experimental design and randomization scheme used, and selection of a preferred analysis model among contending candidate models (purely randomization-based, spatial, with covariate adjustments, etc.). In two-stage analysis, only the means and some measure of precision (standard errors, variance-covariance matrix of the means or diagonal elements of the inverse of this matrix) are saved from the first stage and carried forward to the second stage. By contrast, in a single-stage analysis, the preferred analysis models identified for each individual trial are integrated into an overall model for analysis of the MET plot data, which is fitted in a single stage. The computational burden for single-stage analysis is typically larger than for stage-wise analysis because both the size of the dataset submitted to an analysis across environments and the complexity of the model are larger in single-stage analysis. How much of an advantage the alleviated computational burden by using stage-wise analysis affords depends on several factors, including the size of the dataset, the designs and models used for the individual trials and the complexity of the single-stage model. Moreover, stage-wise analysis is convenient for practical analysis, because it facilitates a combined analysis of different trials with different design and modelling structures and also allows for heterogeneity of variance between trials (Piepho and Eckl, 2014).

Researchers wanting to analyse MET are frequently faced with the question whether to use a single-stage or stage-wise analysis. In this paper it will be argued that, while single-stage analysis can justly be regarded as the gold standard, a stage-wise analysis, if done properly, is perfectly valid and typically very close to a single-stage analysis.

Several papers have been written comparing single-stage and two-stage analysis (Möhring and Piepho, 2009; Welham et al., 2010; Piepho et al., 2012a; Schulz-Streeck et al., 2013a). This in-depth treatment will not be repeated here. Instead of giving very detailed theoretical background, the key results, facts and arguments justifying a stage-wise analysis will be briefly reviewed and the important practical implications discussed. The main purpose of this study is to illustrate stage-wise analysis with typical examples using a new weighting method. This weighting method differs from previous weighting methods in that it carries the full variance-covariance matrix from the first stage to the next stage instead of using a diagonal weighting matrix. Also, it is simpler than an alternative approach, based on rotation (Piepho et al., 2012a), which is slightly more complicated than what we propose here, though results are identical. To the best of our knowledge, a weighting method using the full variance-covariance matrix from the previous stage without rotation has not been used before in the context of series of trials. We provide two macros that can be used to get weights for stage-wise analysis by the new method and by a diagonal method that was suggested previously (Smith et al., 2001). Four worked examples serve to illustrate the similarity between single-stage and stage-wise analysis.

3.2 Statistical methods for trials at multiple sites in a single year

3.2.1 Single-stage analysis

The randomization-based model for analysis of the series of experiments laid out as generalized lattice designs is (Calinski et al., 2005)

$$y_{ijkm} = \phi + g_i + s_j + (gs)_{ij} + r_{jk} + b_{jkm} + e_{ijkm}, \tag{1}$$

where ϕ is a general intercept, g_i is the fixed main effect of the *i*-th genotype, $s_j \sim N(0, \sigma_s^2)$ is the random main effect of the *j*-th site, $(gs)_{ij} \sim N(0, \sigma_{gs}^2)$ is the random interaction effect of the *i*-th genotype and the *j*-th site, $r_{jk} \sim N(0, \sigma_{r(j)}^2)$ is the random effect of the *k*-th replicate within the *j*-th site, $b_{jkm} \sim N(0, \sigma_{b(j)}^2)$ is the random effect of the *m*-th block nested within the *k*-th replicate at the *j*-th site, and $e_{ijkm} \sim N(0, \sigma_{e(j)}^2)$ is the residual plot error associated with the observation y_{ijkm}. Note that the variances for replicates, blocks and error are site-specific here, which is usually a realistic assumption (So and Edwards, 2011) and also allows a two-stage analysis to be fully equivalent to single-stage analysis (Piepho et al., 2012a).

3.2.2 Fully efficient two-stage analysis

The term fully efficient two-stage analysis refers to a two-stage analysis that forwards the full variance-covariance matrix of adjusted means obtained in the first stage to the next stage. For analysis of individual sites (first stage), it is convenient to re-write model (1) as

$$y_{ijkm} = \mu_{ij} + r_{jk} + b_{jkm} + e_{ijkm}, \tag{2}$$

where $\mu_{ij} = \phi + g_i + s_j + (gs)_{ij}$ is the conditional expected value of the *i*-th genotype $(i = 1,...,q)$ at the *j*-th site $(j = 1,..., p)$. We here regard μ_{ij} as a fixed effect for site-wise analysis, i.e. the analysis is conditional on the site-specific effects s_j and $(gs)_{ij}$. Collecting expected values μ_{ij} at the *j*-th site into a vector $\mu_j = (\mu_{1j}, \mu_{2j},..., \mu_{qj})^T$ and plot observations into the vector y_j , we have $\text{var}(\hat{\mu}_j) = \Omega_j = (X_j^T \Sigma_j^{-1} X_j)^{-1}$, where $\hat{\mu}_j$ is the generalized least squares estimator of μ_j, given by $\hat{\mu}_j = (X_j^T \Sigma_j^{-1} X_j)^{-1} X_j^T \Sigma^{-1} y_j$, X_j is a full-rank treatment design matrix for μ_j

at the *j*-th site and $\Sigma_j = \mathrm{var}(y_j)$ is a non-singular variance-covariance matrix of the plot data at the *j*-th site, which depends on the experimental design and the variances $\sigma^2_{r(j)}$, $\sigma^2_{b(j)}$ and $\sigma^2_{e(j)}$.

In the second stage, we can fit the model

$$\hat{\mu}_{ij} = \mu_{ij} + f_{ij} = \phi + g_i + s_j + (gs)_{ij} + f_{ij}, \tag{3}$$

where f_{ij} is the residual of the *i*-th genotype in the *j*-th site and $\mathrm{var}(f_j) = \Omega_j$ with $f_j = (f_{1j}, f_{2j}, ..., f_{qj})^T$. In practice, Ω_j is replaced by its residual maximum likelihood (REML) estimate from the first stage. To fit the model in the second stage, we need the variance-covariance matrix of $f = (f_1^T, f_2^T, ..., f_p^T)^T$ given by

$$\mathrm{var}(f) = \begin{pmatrix} \Omega_1 & 0 & \cdots & 0 \\ 0 & \Omega_2 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \Omega_p \end{pmatrix} = \bigoplus_{j=1}^{p} \Omega_j = \Omega. \tag{4}$$

Plugging in the estimate of Ω_j from the first stage, we can then estimate the fixed genotype means across environments, $\theta_i = \phi + g_i$, and the variances σ_s^2 and σ_{gs}^2 at the second stage based on (3), thus providing estimates of all parameters of the single-stage model (1), if analyses of both stages are taken together. It is shown in Piepho et al. (2012a) that both analyses are fully equivalent provided the same variance component values are used for all random effects. This is also why we denote this two-stage approach as fully efficient. For theoretical details the reader is referred to that paper. The fully efficient method described here is essentially the same as that in Piepho et al. (2012a), except that we omitted the rotation; all other results in that paper apply equally, especially those on the equivalence of single-stage and stage-wise analysis as derived from the mixed model equations.

Any numerical differences between resulting estimates of genotype means θ_i only result from numerical differences in the variance component estimates under single-stage and two-stage analysis. We also note that we have used a simple model for the random genotype-environment effects, but the approach also works with more complex models such as factor-analytic variance-covariance structures for genotype-environment effects (Piepho, 1997).

For illustration we here use the PROC MIXED procedure of the SAS system to perform all analyses. To fit the model (3) in the second stage we can use the code in Box 1 (Piepho and Eckl, 2014; Electronic Appendix).

```
ods output lsmeans=mean_twostage_stagetwo_full_1
covparms=cp_twostage_stagetwo_full_1 ;

proc mixed data= mean_twostage_stageone_full_1w;

class genotype site row;

model estimate=genotype;

random int genotype/sub=site;

repeated row/sub=site type=lin(1)
ldata=mean_twostage_stageone_full_1w;

lsmeans genotype/diffs ;

parms (1)(1)(1)/hold=3;

run;
```

Box 1: SAS code for stage-two analysis of a fully efficient two-stage analysis.

In this code, mean_twostage_stageone_full_1w specified with the data option in PROC MIXED and the LDATA option to the REPEATED statement is a dataset containing the adjusted genotype-site means $\hat{\mu}_{ij}$ and the corresponding estimate of Ω from the first stage in a suitable format as detailed in the Appendix, GENOTYPE and SITE are variables representing the genotypes and sites, ROW is a sequential number indexing genotype-site means in the dataset mean_twostage_stageone_full_1w, and ESTIMATE is the response variable carrying the adjusted genotype-site means. Note that the REPEATED statement (as well as the RANDOM statement) specifies SITE as a subject effect with the SUBJECT option, so the blocks of Ω are processed by sites, which entails savings in memory and computing time compared to a coding not making use of the SUBJECT option. Generally, where possible, it is important that the REPEATED statement and all RANDOM statements share the same subject effect, so that levels of that effect are recognized as independent subjects. The smaller the size of the subjects and the more subjects there are, the better. In this example, the shared subject effect is SITE because correlations among observed data occur only within sites. We here exploit the fact that under the assumed LMM, observations from different sites are independent. Thus, inversions of variance-covariance matrices needed during REML iterations can be performed by sites, which saves computing time. In the Supplemental Material we provide a macro ***%get_one_big_omega***, which assembles the estimate of Ω in a form suitable for use with the code in Box 1 based on site-wise first-stage analyses in which estimates of Ω_j are obtained using the COV option to the LSMEANS statement for estimating genotype means at each site.

3.2.3 Two-stage analysis with diagonal weight matrix

An alternative approach to the fully efficient two-stage analysis described above was proposed by Smith et al. (2001), who suggested to fit the second-stage model assuming that $\mathrm{var}\left(f_{ij}\right)=\left(\omega^{ij}\right)^{-1}$, where ω^{ij} is the *i*-th diagonal element of Ω_j^{-1}. The rationale for this suggestion is that the mixed model equations for (3) depend linearly on $\Omega^{-1}=\bigoplus_{j=1}^{p}\Omega_j^{-1}$, which can be approximated by a diagonal matrix with diagonal elements equal to ω^{ij}. The SAS code in Box 2 can be used to perform this approximate analysis at the second stage.

```
ods output lsmeans=mean_twostage_stagetwosmith_1
covparms=cp_twostage_stagetwosmith_1 ;

proc mixed data= mean_twostage_stageonesmith_1w;

class genotype site;

model estimate=genotype;

random int genotype/sub=site;
```

```
lsmeans genotype / cov;

weight weight_smith;

parms (1)(1)(1)/hold=3;

run;
```

Box 2: SAS code for second stage of an approximate two-stage analysis, using the weights proposed by Smith et al. (2001).

In this code, all variables are as defined for Box 1 and weight_smith is the variable in the dataset mean_twostage_stageonesmith_1w holding the weights. In the Supplemental Material we provide a SAS macro %***get_Smith_weights*** that can compute these weights based on the same site-wise first-stage analyses as under the fully efficient two-stage analysis. A brief description of this macro is available in the Appendix. Using the diagonal approximation in the second stage leads to savings in computing time compared to the fully efficient two-stage analysis.

3.2.4 Statistical methods for trials at multiple sites and in multiple years

Again assuming a generalized lattice design, the first-stage model for the trial in the *j*-th site and *h*-th year is given by

$$y_{ijhkm} = \mu_{ijh} + r_{jhk} + b_{jhkm} + e_{ijhkm}, \quad \text{where} \tag{5}$$

$$\mu_{ijh} = \phi + g_i + s_j + a_h + (gs)_{ij} + (ga)_{ih} + (sa)_{jh} + (gsa)_{ijh}, \tag{6}$$

in which ϕ is a general intercept, g_i is the fixed main effect of the *i*-th genotype, $s_j \sim N(0,\sigma_s^2)$ is the random main effect of the *j*-th site, $a_h \sim N(0,\sigma_a^2)$ is the random main effect of the *h*-th year, $(gs)_{ij} \sim N(0,\sigma_{gs}^2)$ is the random two-way interaction of the *i*-th genotype and the *j*-th site, $(ga)_{ih} \sim N\left(0, \sigma_{ga}^{2}\right)$ is the random two-way interaction effect of the *i*-th genotype and the *h*-th year, $(sa)_{jh} \sim N\left(0, \sigma_{sa}^{2}\right)$ is the random two-way interaction effect of the *j*-th site and the *h*-th year, $(gsa)_{ijh} \sim N\left(0, \sigma_{gsa}^{2}\right)$ is the random three-way interaction effect of the *i*-th genotype, the *j*-th site and the *h*-th year, $r_{jhk} \sim N\left(0, \sigma_{r(jh)}^{2}\right)$ is the random effect of the *k*-th replicate within the *j*-th site and *h*-th year, $b_{jhkm} \sim N\left(0, \sigma_{b(jh)}^{2}\right)$ is the random effect of the *m*-th block nested within the *k*-th replicate at the *j*-th site and *h*-th year, and $e_{ijhkm} \sim N\left(0, \sigma_{e(jh)}^{2}\right)$ is the error associated with the observation y_{ijhkm}. Note that, as before, the variances for replicate, block and error depend on the site-year combination and hence are trial-specific. When the experiment is laid out in randomized complete blocks, we drop the incomplete block effect. Complete blocks are then represented by the complete replicate effect. A stage-wise analysis computes genotype means per trial (year-site combination) in the first stage and then fits model (6) to these means across years and sites in stage two.

3.2.5 Extending the model when sites are stratified into zones

If sites are stratified by zone, model (5) for the observed data remains the same, however, the conditional expected value in equation (6) needs modification. Specifically, each effect involving site in (6) needs to be replaced by two effects, the one involving zone and the other one involving site nested within zone. Thus, equation (6) can be extended as

$$\begin{aligned}\mu_{ij(q)h} &= \phi + g_{i} + z_{q} + (zs)_{j(q)} + a_{h} + (gz)_{iq} + (zgs)_{ij(q)} + (ga)_{ih} + (za)_{qh} \\ &+ (zsa)_{jh(q)} + (zga)_{ih(q)} + (gzsa)_{ij(q)h}\end{aligned}, \tag{7}$$

where all effects involving sites (*s*) in (6) have been replaced by two effects, i.e. one involving zone (*z*) instead of site and the other involving site nested within zone (*zs*). Moreover, g_{i} is the main effect of the *i*-th genotype, and $(gz)_{iq}$ is the interaction of the *i*-th genotype and *q*-th zone. To borrow strength across zones when estimating mean genotype yields for a specific zone (Piepho and Möhring, 2005; Kleinknecht et al., 2013; Piepho et al., 2016a), we modeled

g_i and $(gz)_{iq}$ as random, e.g., assuming that both g_i and gz_{iq} have a normal distribution with mean zero and a constant or heterogeneous variance. Thus, we may obtain estimates of genotype mean in zone *q*

$$\mu_{iq} = \phi + g_i + z_q + (gz)_{iq} \tag{8}$$

using BLUP. BLUP is an estimation method for random effects in LMMs, which minimizes the mean squared error under the assumed model and it entails shrinkage, meaning that the estimate of genotype effects will tend to fall back towards the mean of all genotypes. So BLUPs of good performers tend to be smaller than the corresponding BLUEs, while BLUPs of bad performers tend to be elevated compared to the corresponding BLUEs (Robinson, 1991; Searle et al., 1992). Moreover, in case of correlated genetic effects, BLUP allows exploiting information from correlated observations. In our case, we consider the effect h_{iq} of the *i*-th genotype in the *q*-th zone

$$h_{iq} = g_i + (gz)_{iq} \,. \tag{9}$$

For a given genotype *i*, these effects are correlated between zones *q*, due to the genotype main effect g_i shared between different zones. Thus, when estimating the effect of the *i*-th genotype in the *q*-th zone by BLUP, we are also making use of information on the same genotype from the other zones.

We can estimate effects in (7) in a single stage, in two stages or in three stages. Two-stage analysis proceeds in the same way as previously, with (7) fitted in the second-stage. Three-stage analysis considers effects g_i and $(gz)_{iq}$ as fixed in the second stage to compute estimates of means in (8) and the associated variance-covariance matrix. In the third stage, equation (8) is fitted to these means taking $h_{iq} = g_i + (gz)_{iq}$ as random and using the variance-covariance matrix of adjusted means from the second stage for weighting. In our example, we have two zones, so we need to consider 2×2 variance-covariance structures of the form

$$\mathrm{var}\begin{pmatrix} h_{i1} \\ h_{i2} \end{pmatrix} = \Gamma = \begin{pmatrix} \sigma_{g1}^2 & \sigma_{g12} \\ \sigma_{g12} & \sigma_{g2}^2 \end{pmatrix}, \tag{10}$$

where σ_{g1}^2 is the variance of the *i*-th genotype in zone 1, σ_{g2}^2 is the variance of the *i*-th genotype in zone 2 and σ_{g12}^2 is the covariance between effects of the *i*-th genotype in zones 1 and 2. Equation (10) may be denoted as an unstructured variance-covariance model. Alternatively, we may impose a specific structure. Modeling g_i and $(gz)_{iq}$ as random with the assumptions $g_i \sim N(0, \sigma_g^2)$ and $(gz)_{iq} \sim N(0, \sigma_{gz}^2)$ results in a compound symmetry (CS) variance structure. The CS model has two parameters, i.e., a constant variance and a constant covariance. The CS variance-covariance structure of (9) can be written as

$$\Gamma = \begin{pmatrix} \sigma_g^2 + \sigma_{gz}^2 & \sigma_g^2 \\ \sigma_g^2 & \sigma_g^2 + \sigma_{gz}^2 \end{pmatrix}. \tag{11}$$

An extension of this model is the heterogeneous compound symmetry (CSH) model, which has a different variance parameter for each diagonal element (zone). For two zones, it has the representation

$$\Gamma = \begin{pmatrix} \sigma_{g1}^2 & \sigma_{g1}\sigma_{g2}\rho \\ \sigma_{g1}\sigma_{g2}\rho & \sigma_{g2}^2 \end{pmatrix}. \tag{12}$$

Because there are only two zones in our example, the CSH model is just a re-parameterization of the unstructured model, and it also has the same specification here as the unstructured model parameterized in terms of variances and correlations (UNR). All of these structures (CS, CSH, UN and UNR) are available in SAS.

A stage-wise analysis fits equation (7) across years and sites in stage two in order to compute genotype means per zone by BLUE. The BLUE of genotype means at the second stage will be used for comparison with BLUP from two-stage and three-stage analysis. In stage three of three-stage analysis, the linear predictor (8) is fitted to genotype-zone means using the random-effects specification in (9) and (10).

3.3 Example 1: Trials conducted at multiple sites in a single year

3.3.1 The dataset

Twenty-two different genotypes of non-quality protein maize (*Zea mays* L.) (non-QPM) were evaluated in the Ethiopian preliminary national variety trials of maize (EVCDTH12; Evaluation of CIMMYT drought tolerant hybrids in 2012 main rainy season). The trials aim to identify high yielding, adapted hybrids for low-moisture stress areas. The experiment was conducted during the period from July 1, 2012 to December 25, 2012 in the low moisture stress area at four sites (Dhera, Melkassa, Mieso and Ziway). The experimental designs used at all sites were α-designs with 11 incomplete blocks of size two in each replicate. Each trial had three replicates. The plot size was 7.5 m^2 with six planted rows. This data is made available as dataset Example1 in the Supplemental Material.

3.3.2 Results

Single-stage analysis and two-stage analysis were performed. Variance component estimates for single-stage and two-stage analyses agree reasonably well (Table 1). The estimated means in Table 2 (columns 1 and 3) show that the fully efficient two-stage analysis carrying the full variance-covariance matrix of adjusted means forward from stage one yields identical results to single-stage analysis provided the same variance component values are used as expected from theory (Piepho et al., 2012a). When variances are estimated separately in each type of analysis, adjusted genotype means from single-stage analysis show correlations larger than 0.99 with those of two-stage analyses (Table 3).

Table 1: Variance component estimates for single-stage analysis, fully efficient two-stage analysis, and two-stage analysis with diagonal weights (Smith et al., 2001) (Example 1: EVCDTH12 maize trial dataset).

Variance parameter	Fully efficient two-stage	Smith et al. approximation two-stage	Single-stage
σ_s^2	10.4537	10.4227	10.4543
σ_{gs}^2	0.1272	0.1279	0.1053
$\sigma_{r(1)}^2$	0.1004	0.1004	0.08817
$\sigma_{r(2)}^2$	1.4012	1.4012	1.3829
$\sigma_{r(3)}^2$	0	0	0
$\sigma_{r(4)}^2$	0.01413	0.01413	0.01442
$\sigma_{b(1)}^2$	0.2504	0.2504	0.3312
$\sigma_{b(2)}^2$	0.4645	0.4645	0.4747
$\sigma_{b(3)}^2$	0	0	0
$\sigma_{b(4)}^2$	0.07197	0.07197	0.06953
$\sigma_{e(1)}^2$	1.2363	1.2363	1.3467
$\sigma_{e(2)}^2$	0.2020	0.2020	0.1936
$\sigma_{e(3)}^2$	1.0549	1.0549	1.1531

$\sigma^2_{e(4)}$	0.1126	0.1126	0.1112

Table 2: Adjusted genotype estimates when (1) the full information of estimates and their corresponding measure of precisions and (2) estimates and diagonal weights (Smith et al., 2001), are carried forward from the first stage to the second stage of the analysis. Analyses (3), (4), and (5) are single-stage analyses, where (3) and (4) use the variance-covariance matrix of mean estimates from (1) and (2), respectively; and (5) is single-stage analysis when the variances components are estimated directly from the plot data (Example 1: EVCDTH12 maize trial dataset).

Genotype	(1) Fully efficient two-stage	(2) Smith et al. approximation two-stage	(3) Single-stage variance estimate from (1)	(4) Single-stage variance estimate from (2)	(5) Single-stage variances re-estimated
1	5.135	5.113	5.135	5.135	5.148
2	5.510	5.431	5.510	5.509	5.544
3	5.147	5.152	5.147	5.147	5.187
4	4.593	4.584	4.593	4.593	4.589
5	4.845	4.812	4.845	4.845	4.817
6	4.692	4.705	4.692	4.692	4.659
7	4.663	4.678	4.663	4.663	4.636
8	4.405	4.388	4.405	4.405	4.358
9	5.055	5.027	5.055	5.056	5.033
10	4.839	4.803	4.839	4.839	4.841
11	4.542	4.503	4.542	4.542	4.539
12	4.890	4.910	4.890	4.890	4.870
13	4.850	4.859	4.850	4.850	4.841

14	4.325	4.280	4.325	4.326	4.293
15	4.448	4.415	4.448	4.448	4.472
16	4.343	4.328	4.343	4.342	4.371
17	4.072	4.086	4.072	4.072	4.068
18	4.978	4.916	4.978	4.978	4.945
19	4.711	4.724	4.711	4.711	4.715
20	4.808	4.819	4.808	4.807	4.861
21	4.128	4.135	4.128	4.128	4.133
22	4.577	4.581	4.577	4.576	4.628

Table 3: Correlation among adjusted genotype means (above diagonal: Pearson's product-moment correlation; below diagonal: Spearman's rank correlation). When (1) the full information of estimates and their corresponding measure of precisions and (2) estimates and diagonal weights (Smith et al., 2001), are carried forward from the first stage to the second stage of the analysis. Analyses (3), (4), and (5) are single-stage analyses, where (3) and (4) use the variance-covariance matrix of mean estimates from (1) and (2), respectively; and (5) is single-stage analysis when the variances components are estimated directly from the plot data (Example 1: EVCDTH12 maize trial dataset).

	(1)	(2)	(3)	(4)	(5)
	Fully efficient two-stage	Smith et al. approximation two-stage	Single-stage variance estimates from (1)	Single –stage variance estimate from (2)	Single-stage variances re-estimated
(1)	1	0.99707	1.00000	1.00000	0.99674
(2)	0.99661	1	0.99707	0.99706	0.99477
(3)	1.00000	0.99661	1	1.00000	0.99674
(4)	1.00000	0.99661	1.00000	1	0.99665

(5)	0.98984	0.99548	0.98984	0.98984	1.00000

3.4 Example 2: Extending Example 1 to allow for trials-specific analysis models for post-blocking and residual error

This example is presented to illustrate the performance when different analysis models (randomization-based baseline model, spatial models and models with post-blocking for row, column and column nested within replicate effects) are used for individual trials, using the dataset of Example 1. First, the baseline model is extended by effects for row, column or column nested within replicate. Taking the optimal model from these candidate models for each site, local spatial trends are modelled by one-dimensional and two-dimensional auto-regressive models. For the one-dimensional case we assume that a correlation exists either within rows, within columns or within columns nested within replicates. The two-dimensional AR(1)×AR(1) model is fitted assuming that correlation extends across the whole field. For all autoregressive models, the autocorrelation parameter was constrained to be non-negative (Piepho et al., 2015). The best model for each individual trial was selected using the Akaike information criterion (AIC) (Table 4).

3.4.1 Results

The estimates of the variance components using the fully efficient and diagonal weighting are quite similar with this method as well, however, compared to Example 1 the variance component estimate using single-stage analysis were somewhat more different from the stage-wise analyses (Table S1). The correlations of adjusted genotype means using the different approaches (Table S2) are slightly smaller than in Example 1, but are all greater than 0.98 (Table 5), indicating close similarity of single-stage and two-stage analysis.

Table 4. AIC values for the baseline and different extended models for each site (Example 2). AIC values for the model with the best fit are given in bold.

Model	AIC from analysis of site			
	1	2	3	4
Baseline	173.7	134.3	153.4	79.9
Baseline model plus post-blocking:				
Row	**162.4**	135.9	153.4	79.9
Col	175.7	134.3	153.4	79.9
Col(rep)	173.7	136.3	153.4	**74.6**
Row+col	164.4	135.9	153.4	79.9
Row+col(rep)	163.0	137.9	153.4	76.5
Best post-blocking model with spatial add-on component[§]:				
AR(1) along row	162.4	**130.1**	**152.7**	74.6
AR(1) along col	162.4	134.3	153.4	74.6
AR(1) along col(rep)	163.0	136.3	153.4	76.6
AR(1) along row + nugget	164.4	131.7	153.7	76.6
AR(1) along col + nugget	164.4	136.3	155.4	76.6
AR(1) along col(rep) + nugget	164.4	138.3	155.4	78.6
AR(1) ×AR(1)	162.4	130.1	154.7	74.6
AR(1) ×AR(1) + nugget	164.4	131.7	155.7	78.6

Baseline: Baseline model with all randomization based effects including independent error effects; row, col, col(rep), row+col and row+col(rep) are models extending the baseline model by post-blocking terms for row, column or column nested within replicate; AR(1): first order auto-regressive correlations between plots along the mentioned experimental unit; + nugget models include an additional independent error effect; AR(1)×AR(1): two-dimensional autoregressive variance-covariance structure, so correlations extended along both rows and columns.

§ All spatial models include effects of the best post-blocking model. If post-blocking was not effective, the baseline model is used for augmentation with a spatial error component.

Table 5: Correlation among adjusted genotype means (above diagonal: Pearson's product-moment correlation; below diagonal: Spearman's rank correlation) when (1) the full information of estimates and their corresponding variance-covariance matrix and (2) estimates and diagonal weights (Smith et al., 2001), are carried forward from the first stage to the second stage of the analysis. Analyses (3), (4), and (5) are single-stage analyses, where (3) and (4) use the variance-covariance matrix of mean estimates from (1) and (2), respectively, and (5) is single-stage analysis when the variances components are estimated directly from the plot data (Example 2).

Approach	(1)	(2)	(3)	(4)	(5)
	Fully efficient two-stage	Smith et al. approximation two-stage	Single-stage variance estimates from (1)	Single –stage variance estimate from (2)	Single-stage
(1)	1.0000	0.9894	1.0000	0.9996	0.9819
(2)	0.9955	1.0000	0.9894	0.9917	0.9905
(3)	1.0000	0.9955	1.0000	0.9996	0.9819
(4)	0.9977	0.9921	0.9977	1.0000	0.9862
(5)	0.9842	0.9887	0.9842	0.9864	1.0000

3.5 Example 3: Trials at multiple sites and in multiple years

3.5.1 The Dataset

During the 1997 and 1998 main cropping seasons, twenty different maize varieties of East African and CIMMYT origin were tested at nine sites. These sites represent two of the maize-producing mega-environments (zones) in Ethiopia; viz. the low (low-mid) altitude sub-humid zone and the high altitude sub-humid zone (Fig. 1). Randomized complete block designs with three replicates and two-row plots were used at all sites and in both years. Each row was 5.1 meter in length, the space between rows was 0.75 m and the distance between plants was 0.3 m. The recommended management was applied in each site. This data is made available as dataset Example3 in the Supplemental Material.

3.5.2 Results

Results demonstrate the similarity of single-stage and two-stage analysis for the multi-site and multi-year dataset using the fully efficient two-stage analysis and the approximate two-stage method of Smith et al. (2001). The variance parameter estimates are approximately equal for the three methods (Table S3). Likewise, the estimated adjusted genotype means are quite similar for single-stage versus two-stage analysis (Table S4), as also indicated by the correlations presented in (Table 6).

Table 6: Correlation among adjusted genotype means (above diagonal: Pearson's product-moment correlation; below diagonal: Spearman's rank correlation) when (1) the full information of estimates and their corresponding variance-covariance matrix and (2) estimates and diagonal weights (Smith et al., 2001) are carried forward from the first stage to the second stage of the analysis. Analyses (3), (4), and (5) are single-stage analyses, where (3) and (4) use the variance-covariance matrix of mean estimates from (1) and (2), respectively; and (5) is single-stage analysis when the variances components are estimated directly from the plot data for the multi-site and multi-year maize trial dataset (Example 3).

	(1)	(2)	(3)	(4)	(5)
	Fully efficient two-stage	Smith et al. approx. two-stage	Single-stage variance estimates from (1)	Single-stage variance estimates from (2)	Single-stage
(1)	1	1.0000	1.0000	1.0000	0.9999
(2)	1.0000	1	1.0000	1.0000	0.9999
(3)	1.0000	1.0000	1	1.0000	0.9999
(4)	1.0000	1.0000	1.0000	1	0.9999
(5)	1.0000	1.0000	1.0000	1.0000	1

3.6 Example 4: Extending the model for Example with sites stratified into zones

Here we perform single-stage, two-stage and three-stage analysis using the data of Example 3. The CS, CSH, UN, and UNR variance structures were imposed for the correlation between zones. Among the fitted variance-covariance structures the CS model performed better (had a smaller AIC value) than the other models, therefore we summarize the result to show the similarity of single-stage and stage-wise analysis using the CS variance structure. For CSH and UN, the single-stage analysis did not converge, so only results of CS and UNR are presented (Table 7).

Table 7: Akaike Information Criterion (AIC) and -2 residual log-likelihood (-2LL) values with different variance structures for fully efficient two-stage, three-stage and single-stage analysis for the multi-site and multi-year maize trial dataset (Example 4).

Covariance structure †	Two-stage analysis ‡		Three-stage analysis ‡		Single-stage analysis ‡	
	AIC	-2LL	AIC	-2LL	AIC	-2LL
CS	984.1	972.1	73.8	69.8	2965.8	2901.8
CSH	985.4	971.4	75.1	69.1	--ᶲ	--ᶲ
UN	985.4	971.4	75.1	69.1	--ᶲ	--ᶲ
UNR	985.4	971.4	75.1	69.1	2967.2	2901.2

† CS, compound symmetry; CSH, heterogeneous compound symmetry;

UN, unstructured; UNR, unstructured correlations.

‡ AIC, Akaike Information Criterion; -2LL, -2 residual log likelihood

ᶲ Did not converge from CSH and UN

3.6.1 Results

Since we only have two zones, the CSH, UN, and UNR models have exactly equal variance-covariance and correlation values for the stage-wise analysis. The estimated genetic correlations between zones are large, which indicates a close relation of the two zones in terms of the adjusted genotype means (Table 8, 9 and 10).

Table 8: Values of genotypic variance (on the diagonal), correlation (above diagonal) and covariance (below diagonal) for the fully efficient three-stage analysis, for compound symmetry (CS), heterogeneous compound symmetry (CSH), unstructured (UN) and unstructured correlation (UNR) variance structure for the multi-site and multi-year maize trial dataset (Example 4).

	Covariance structure †							
	CS		CSH		UN		UNR	
Zone	1	2	1	2	1	2	1	2
1	0.3325	0.7994	0.2615	0.8185	0.2615	0.8185	0.2615	0.8185
2	0.2658	0.3325	0.2577	0.3791	0.2577	0.3791	0.2577	0.3791

† CS, compound symmetry; CSH, heterogeneous compound symmetry;

UN, unstructured; UNR, unstructured correlations.

Table 9: Values of genotypic variance (on the diagonal), correlation (above diagonal) and covariance (below diagonal) for the fully efficient two-stage analysis, with compound symmetry (CS), heterogeneous compound symmetry (CSH), unstructured (UN), and unstructured correlations (UNR) variance structure for the multi-site and multi-year maize trial dataset (Example 4).

	Covariance structure †							
	CS		CSH		UN		UNR	
Zone	1	2	1	2	1	2	1	2
1	0.3318	0.7999	0.2611	0.8194	0.2611	0.8194	0.2611	0.8194
2	0.2654	0.3318	0.2575	0.3782	0.2575	0.3782	0.2575	0.3782

† CS, compound symmetry; CSH, heterogeneous compound symmetry;

UN, unstructured; UNR, unstructured correlations.

Table 10: Values of genotypic variance (on the diagonal), correlation (above diagonal) and covariance (below diagonal) for the single-stage analysis, with compound symmetry (CS) and unstructured correlations (UNR) variance structures for the multi-site and multi-year maize trial dataset (Example 4).

	Covariance structure †			
	CS		UNR	
Zone	1	2	1	2
1	0.3332	0.8008	0.2625	0.8197
2	0.2668	0.3332	0.2585	0.3789

† CS, compound symmetry;

UNR, unstructured correlations.

Overall, the variance-covariance parameter estimates are very similar for the single-stage, two-stage and three-stage analysis (Table S5). There are also close similarities between the BLUPs of single-stage, two-stage and three-stage analysis (Tables S6) as quantified by the correlations larger than 0.96 in Table 11. The correlations of BLUEs and BLUPs are smaller with values between 0.92 and 0.98 (Table 11).

Table 11: Correlation among adjusted genotype means using best linear unbiased prediction (BLUP) and best linear unbiased estimation (BLUE) (above diagonal: Pearson’s product-moment correlation; below diagonal: Spearman’s rank correlation). BLUEs are computed using (1) single-stage analysis (BLUE_1), (2) fully efficient two-stage analysis (BLUE_FE2), and (3) diagonal weights two-stage analysis (BLUE_Smith2), whereas BLUPs are computed based on (4) single-stage analysis (BLUP_1), (5) fully efficient two-stage analysis (BLUP_FE2), (6) fully-efficient three-stage analysis (BLUP_FE3), (7) diagonal weights two-stage analysis (BLUP_Smith2), and diagonal weights three-stage analysis (BLUP_Smith3). Results are for the zoned multi-site and multi-year maize trial dataset (Example 4).

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	BLUE_1	BLUE_FE2	BLUE_Smith2	BLUP_1	BLUP_FE2	BLUP_FE3	BLUP_Smith2	BLUP_Smith3
(1)	1	0.9999	0.9999	0.9821	0.9821	0.9819	0.9826	0.9678
(2)	0.9981	1	1.0000	0.9816	0.9817	0.9815	0.9822	0.9672
(3)	0.9979	0.9998	1	0.9811	0.9812	0.9810	0.9818	0.9667
(4)	0.9503	0.9462	0.9452	1	0.9999	0.9999	0.9999	0.9971
(5)	0.9448	0.9407	0.9398	0.9994	1	1.0000	1.0000	0.9972
(6)	0.9448	0.9407	0.9398	0.9994	1.0000	1	0.9889	0.9972
(7)	0.9477	0.9435	0.9428	0.9996	0.9994	0.9994	1	0.9972
(8)	0.9180	0.9133	0.9124	0.9901	0.9916	0.9916	0.9914	1

The differences and comparatively low correlations between BLUP and BLUE of genotype effects in Figs. 2, 3, and 4 imply that there is a considerable environmental variation and BLUP borrows a substantial amount of information across zones. However, by contrast BLUE can not borrow information across zones. Note that BLUPs of genotype-zone means in three-stage analysis are compared to BLUEs of genotype-zone means computed at the second stage (i.e., a third stage is not needed with BLUE). In this example as well as in the above three examples, we found the estimated variance-covariance matrix for the random effects to be non-positive definite in some stages of the analysis, which was due to some variance estimates being zero. While a message to this effect is printed in the log window, this is no reason for concern; zero variance component estimates are not uncommon. Effects with zero variance are effectively removed from the model and the resulting analysis is fine.

3.7 Discussion

It is shown in Piepho et al. (2012a) that two-stage and single-stage analysis yield fully equivalent results provided that (i) the same values are used for all relevant variance parameters and the full information on all effect estimates and their associated estimated variances and covariances are carried forward from the first to the second stage, (ii) the same model assumptions are used for all effects, and (iii) all effects for which estimates are carried forward are formally regarded as fixed in the first stage. These results naturally carry over to more than two stages, the requirement being that all effects for which estimates are carried forward in any stage are formally modeled as fixed up to that stage. This was illustrated in the present paper using MET data for maize in Ethiopia. Thus, providing the full equivalence of models, any discrepancies in genotype mean or effect estimates only arise from differences in the variance-covariance parameter estimates. A further cause of differences between both analyses arises when the variance-covariance matrix of estimated effects from the first stage is approximated by a diagonal matrix (Smith et al., 2001) rather than carried forward in full as was also illustrated in this paper.

In our study the numerical differences are very small regarding the resulting genotype mean estimates, and this has also been found in other work by our group (Piepho and Möhring, 2005; Piepho et al., 2012a; Schulz-Streeck et al., 2013a; Piepho and Eckl, 2014). Therefore, we believe that for the types of data we typically see, a stage-wise analysis is perfectly valid and acceptable for most practical purposes. The main advantage of stage-wise analysis is that analysis of individual trials with different designs can be done for all trials at the same time with their corresponding appropriate models, whence the adjusted means and the associated variance-covariance matrix of adjusted means can be stored away once and for all for later processing in a stage-two analysis. We found that even with our relatively simple examples there were convergence problems in single-stage analysis, particularly when different variance-covariance structures were imposed (Table 5). In addition, the time taken by stage-wise analysis was smaller than the time taken by single-stage analysis. For instance, with the CS variance structure single-stage analysis for Example 4 took thirty-three hours, two-stage analysis took approximately two minutes and three-stage analysis took less than one minute on a standard desktop computer (Windows 7, 64 bit operating system, 4GB RAM) .

In general two-stage analysis can always be used provided the individual trials allow a separate analysis, as will be the case when designs with proper randomization and (partial) replication are used. So whenever single-stage analysis is inconvenient or computationally too demanding, stage-wise analysis can be recommended. The fully-efficient weighting method is the preferred one because it carries all information forward to the next stage, but it is computationally more demanding than use of diagonal weights. When computational resources are limiting, diagonal weights can be used, and in our experience the loss of information compared to a fully efficient analysis is usually negligible (Möhring and Piepho, 2009).

A key question with the models we consider here is whether the genotype factor is fixed or random. This decision depends on the objectives of the experiment. For example, if the objective is selection of the best genotypes from a population of genotypes under study and it is reasonable to assume that genotype effects at least approximately follow a normal distribution, then genotype effects can be considered as random and BLUP will be the best method of estimation to obtain ranks of the genotypes which are very close to the true rankings of the genotype effects (Searle et al. 1992, p.264). On the other hand if the objective of the analysis is to obtain significance tests for the difference between pairs of genotypes, then BLUE is an appropriate method (Smith et al., 2005). In variety testing in Ethiopia, it is customary to take genotypes as fixed and compute adjusted means across environments. But we have given an example where genotypes were taken as random in order to exploit correlations between zones for making zone-specific predictions. Such predictions (BLUPs) have been shown to be more accurate than zone-specific mean estimates (BLUEs) assuming fixed effects, which can not borrow strength across zones (Kleinknecht et al., 2013). The approach does require that there is a sufficient number of genotypes to estimate genotypic variances and covariances and the distribution of effects can reasonably be assumed to be approximately normal. Genotypes are also modeled as random in genomic prediction in order to permit estimation of effects of markers that may be much larger in number than the genotypes tested (Meuwissen et al., 2001). The assumption of genotypes as either fixed or random can be considered as an intrinsic part of the single-stage model for plot data. A salient feature of the stage-wise approach advocated here, however, is that regardless of the status of genotypes as either fixed or random in the single-stage model, genotypes need to be formally taken as fixed through all stages of the analysis except the last, where genotypes are fixed or random depending on the status of genotypes in the single-stage model. It is shown in Piepho et al. (2012a) that this approach of stage-wise analysis leads to valid results that are identical to those of single-stage analysis when the same variance parameter values are used in single-stage and stage-wise analysis.

We frequently find in publications that a stage-wise analysis is conducted in which BLUP of genotype means or effects are used in the first stage. This practice is problematic and should be discouraged. For example, when BLUP is also used in the second stage, this entails a double-shrinkage of effects, the one occurring in the first stage and the other occurring in the second stage (Smith et al., 2001). To correct for this problem, BLUPs obtained in the first stage would need to be unshrunk, and it is not clear how. Also, the resulting analysis is not equivalent to single-stage analysis when the same variance values are used in both. For these reasons, we recommend not using BLUP in the first stage of two-stage analysis.

As a note of caution, we would like to point out that our view on the fixed versus random issue presented here is restricted entirely to the modeling of genotypic effects. We admit that the view is somewhat pragmatic. In particular, we do not think the random assumption requires that the tested genotypes literally have been randomly sampled from a larger population. In support of our view, we would like to cite from the (decidedly non-Bayesian) textbook of Lee et al. (2006, p147): "… even if the true model is the fixed-effects model, i.e., there is no random sampling involved, the use of random-effect estimation has been advocated as shrinkage estimation. […] Only when the number of random effects is small, for example three or four, will there be little gain from using the random-effect model (James and Stein, 1960)."

There are close ties between the analysis of series of trials and (network) meta-analysis (Vargas et al., 2012; Piepho et al., 2012c; Madden et al., 2016). The two-stage approach corresponds to what is standard practice in meta-analysis of clinical trials (Whitehead, 2002). The result of clinical trials is usually stored in the form of effect size estimates and associated standard errors. This information may be summarized across trials using a mixed model with random effects for heterogeneity, i.e. treatment-trial interaction, using the standard errors of effect estimates to compute suitable weights for the combination of effect estimates. The resulting mean treatment effect estimates are either equivalent or very similar to estimates obtained by a single-stage analysis of individual-patient data (Piepho et al., 2012c).

Instead of analyzing treatment differences as is common practice in meta-analysis, one may proceed as in two-stage analysis of MET data and summarize treatment means by a suitable model in the second stage. This analysis, which is particularly helpful in meta-analyses comprising more than two treatments and trials with different treatment designs, is fully equivalent to analysis based on treatment differences if the site (study) main effect is taken as fixed rather than random so that all information on treatment comparison comes from comparisons within sites (studies) only, i.e. no inter-site (study) information is recovered (Piepho et al., 2012c). This equivalence re-enforces our assertion that a two-stage analysis, if done properly, is appropriate with little difference from the corresponding single-stage analysis.

A key question in the analysis of series of trials is whether genotype effect estimates are to be obtained for the mean of a target population of environments (TPE) or for the individual environments where the trials were conducted. We would argue that in practical breeding programs the performance of genotypes in a specific test environment is hardly ever of any particular interest. This is because varieties are needed that perform well on average across all environments in a given TPE. If this is what is required, then it is useful to assess the performance of contending genotypes in a random sample of environments from the TPE (Yates and Cochran, 1938; Comstock and Moll, 1963). The main error term for inferences about the means in the TPE is the genotype-environment interaction variance (Talbot, 1997), meaning that the difference between different approximations for the variance-covariance matrices of adjusted genotype means in stage two is typically small (Möhring and Piepho, 2009). By contrast, when estimates for individual environments are of interest, which may be the case in research projects exploring the pattern and causes of genotype-environment interaction (corresponding to what is known as heterogeneity in meta-analysis), the only error term is that pertaining to the variance-covariance matrix of adjusted means. The genotype-environment interaction in this case is an effect to be predicted, not an error term. As a result, the difference between single-stage and two-stage results may be somewhat more relevant and the edge in efficiency in favor of a single-stage analysis may be more pronounced (Welham et al., 2010).

In most applications, however, it is not an individual site and year that is of interest, but either a new year and a new site at which no trial has been conducted, such as a specific farmer's field, or a larger TPE to which a new variety is to be released. If predictions for an individual farmer's field are to be made, one may be tempted to use predictions of the closest trial site. Valid standard errors for these predictions can not be obtained, however, because the interaction pattern between target site (farmer's field) and the nearest trial site, as well as the corresponding interactions with years, are unknown. If predictions are required for a whole TPE, however, valid inferences can be obtained, provided a random sample of sites and years from that TPE is available. From a breeder's perspective, prediction of the expected performance in a given TPE may be the most useful approach to analysis of MET because this helps identifying genotypes performing well on average (in the long run) in the TPE. By contrast, accurate predictions for an individual trial site and year are not usually of any intrinsic interest in themselves because the trial environment does not usually represent conditions identical to any other environment in the TPE (Piepho et al., 2012a).

What is often more informative than predictions for individual environments is to sub-divide a TPE into several agro-ecological zones, each represented by several environments and then obtain predictions per zone. Modelling genotype-zone effects as random allows borrowing strength across zones (Atlin et al., 2000; Kleinknecht et al., 2013; Piepho et al., 2016a). Realistic inferences are obtained at the zone level because several sites are used to assess the between-site sampling variation within zones. It needs to be borne in mind, however, that predictions are for zone means and not for individual sites within a zone. The random genotype-site interaction acts as the main error term for these zone-wise predictions, and as a result these predictions have a broader inference space than predictions for individual sites (McLean et al., 1991). Note that in this study the sites are assumed to be random samples, therefore all effects nested within sites, i.e., replicates and blocks (regardless of whether they are complete or incomplete), were considered as random (Piepho et al., 2012a).

The high correlation of random effects of the genotypes between the zones found in Example 4 indicates that based on this study the two zones are not very different agro-ecologically in terms of genotype means (Piepho and Möhring, 2005; Kleinknecht et al., 2013). At first sight, this result seems to contradict the Ethiopian maize breeder's perception, which is based on adjusted genotype means per zone rather than on estimates of the genotypic correlation between zones. It is a common finding in such studies, however, that the phenotypic correlations between zones are smaller than the corresponding genotypic correlations. Even if the current agro-ecology subdivision helps breeders in developing agro-ecologically adapted varieties, there still exists variability within zones which causes difficulties in selecting stable varieties. A further detailed examination of the agro-ecology within zones has been suggested by breeders (Worku et al., 2012). For a better delineation of zones, important agro-ecological factors should be taken into account (Gauch, 1992; Atlin et al., 2000).

An important example where appropriate weighing in the second stage can be crucial is in genomic prediction (Meuwissen et al., 2001) and genome-wide association mapping (George et al., 2015). For example, in genomic prediction it is common practice to compute genotype means across environments in one or several stages and then to submit these means to some standard routine for regularized regression on the markers such as GBLUP. The residual variance of such analyses comprises both true errors associated with the genotype mean estimates and residual genotypic effects. Typically, the residual variance component may occasionally take on extreme values, e.g. it may move to a very small value in iterations. Such numerical problems may be tackled by explicitly modeling the unexplained polygenic effect and the error-of-a-mean effect by separate variance components, fixing the residual variance at the variance of a mean from the first stage of the analysis (Piepho et al., 2012b).

Our LMM approach for MET analysis can be readily extended to generalized linear mixed models (GLMM) (Stroup, 2015). In a GLMM framework with other distributions and link functions, one can still obtain adjusted genotype means on the link scale, along with the variance-covariance matrix, in the first-stage analysis. With these results from the first stage, one can proceed exactly as described for LMM. So the only difference lies in the analysis of the first stage, where a different link and distribution function are used and for this purpose PROC GLIMMIX is used instead of PROC MIXED (Madden et al., 2016).

3.8 Appendix

We here briefly describe the two macros ***%get_one_big_omega*** and %***get_Smith_weights***, which are available in the Supplementary Material as get_one_big_omega.sas and get_Smith_weights.sas, respectively, at the journal's website along with the full SAS code for performing all analyses reported in this paper for the Ethiopian maize datasets.

3.8.1 The macro %get_one_big_omega

This macro processes a dataset containing the variance-covariance matrices of adjusted genotype means from several trials and generates a SAS dataset containing the block-diagonal variance-covariance matrix Ω in a form ready for use with the LDATA= option in a REPEATED statement specifying a LIN(1) variance-covariance structure using the TYPE= option (see sample code in Box 1). The input dataset for this macro must be ordered by trials (sites, site-year or zone-site-year combinations) and in a format as is generated when outputting adjusted means and associated variance-covariance matrices, computed with the MIXED procedure using the COV option on the LSMEANS statement and variables to identify trials (e.g. relevant combination of "site", "year" and "zone") as by-processing variables in a BY statement, via the output delivery system ODS. The macro generates a SAS dataset containing the following variables: PARM, a serial number for variance components of a specified linear variance-covariance structure (here PARM=1 for all rows in the dataset), ROW, a sequential number for rows in the dataset, and COL1-COL*n*, where *n* is the number of genotype-trial means in the dataset. These latter variables carry the block-diagonal variance-covariance matrix Ω.

3.8.2 The macro %get_Smith_weights

This macro processes the same kind of input dataset as the ***%get_one_big_omega*** macro. It uses a call of the MIXED procedure in order to compute the inverses Ω_j^{-1} from which the weights are then extracted and added as a column with variable name weight_smith in the output dataset. This weight variable can then be used in a stage-two analysis (see sample code in Box 2).

3.9 Supplemental information

Additional supporting information will be found in the online version of this article:

get_one_big_omega.sas contains the SAS macro ***%get_one_big_omega***.

get_Smith_weights.sas contains the SAS macro ***%get_Smith_weights.***

Examples get_one_big_omega.sas contains all the SAS codes used for performing fully-efficient two-stage and three-stage analysis for the four examples considered in this study.

Examples get_Smith_weights.sas contains all the SAS codes used for performing two-stage and three-stage analysis with diagonal weight matrix for the four examples considered in this study.

Supplemental tables: these include Table S1, Table S2, Table S3, Table S4, Table S5, and Table S6 which are cited in the paper; they are tables of genotype means and their variance-covariance matrix and serve to show the similarity of the results for single-stage and stage-wise analysis (fully efficient and diagonal weights of Smith et al. (2001)) for Examples 2, 3 and 4.

Chapter 4

Comparison of weighted and unweighted stage-wise analysis for genome-wide association studies and genomic selection

Tigist Mideksa Damesa[1], Jens Möhring[1], Manje Gowda[2], Yoseph Beyene[2], Biswanath Das[2], Kassa Semagn[2], Hans-Peter Piepho[1]*

[1] Biostatistics Unit, Institute of Crop Science, University of Hohenheim, Fruwirthstrasse 23, 70599 Stuttgart, Germany.

[2]International Maize and Wheat Improvement Center (CIMMYT) Nairobi, Kenya.

Received ___________.

*Corresponding author: hans-peter.piepho@uni-hohenheim.de

Supporting information web address:

NB: This paper is in major revision with Crop Science journal

4.1 Abstract

Both genome-wide association studies (GWAS) and genomic selection (GS) are done using phenotypic and genomic data. The phenotypic data are usually based on multi-environment trials (MET). For both GWAS and GS the analysis can be conducted using a single-stage or a stage-wise approach. Single-stage analysis is most efficient but it can also be computationally demanding. The computational demand increases compared to purely phenotypic analysis when marker information is added for doing the GWAS or the GS. Application of stage-wise analysis is a common alternative procedure to alleviate the computational burden in MET analysis, and it can also be used for GWAS and /or GS. If done properly, it can closely mimic single-stage analysis. The aim of this study is to compare weighted stage-wise analysis versus unweighted stage-wise analysis for GWAS and GS using phenotypic and genotypic maize data. For weighting we use a fully-efficient and a diagonal method. Our result show that weighting is to be preferred over unweighted analysis and that there is a modest advantage in using the fully-efficient weighting method over other weighting methods for GS. For GWAS the diagonal weighting method performs better, however, its difference from the fully efficient weighting is very small.

Abbreviations: BLUE, best linear unbiased estimators; CIMMYT, International Maize and Wheat Improvement Center; CV, cross validation; GBS, genotyping-by-sequencing; GS, genomic selection; GEBV, genomic estimated breeding values; GWAS, genome-wide association study; LD linkage disequilibrium; MET, multi-environment trials; MSD, mean squared difference; PCA, principal component analysis; PC, principal components

In conventional plant breeding programs, selection of the best genotypes is done based only on the phenotypic records of the traits of interest. However, the observed phenotypic effects of quantitative traits are determined by genetic effects. Identifying and mapping genes that confer resistance to constraints such as drought, disease, etc. is key to crop improvement. Genome-wide association studies (GWAS) are commonly used in breeding to scan the entire genome in order to identify genes that affect traits of interest. Population structure and familial relatedness can create a linkage disequilibrium (LD) between unlinked loci, which can result in false-positive marker-phenotype associations when ignored. For this reason, statistical methods have been developed for GWAS, which account for both the population structure and familial relatedness (Yu et al., 2006; Oraguzie et al., 2007; Stich et al., 2008).

Genomic selection (GS) is another technique for efficient selection of favourable genomic estimated breeding values (GEBV) in animal and plant breeding systems using all dense genome-wide markers, and phenotyped and non-phenotyped individuals (Meuwissen et al., 2001; Hayes and Goddard, 2010; Gowda et al., 2015). The main advantage of GS is an increase in genetic gain through prediction of genotypes that have not been phenotyped. A key challenge in GS is the identification of a suitably large training population to estimate the marker effects, requiring the combination of trial data across multiple environments and sets of genotypes (Auinger et al., 2016; Bernal-Vasquez et al., 2017).

Both GWAS and GS are done using phenotypic and genomic data. The phenotypic data is usually obtained from multi-environment trials (MET). For both GWAS and GS the analysis can be conducted using a single-stage or a stage-wise approach (Stich et al., 2008; Piepho et al., 2012a; Schulz-Streeck et al., 2013a; Damesa et al., 2017). The combined analysis of phenotype MET data using single-stage analysis is usually computationally demanding and computational demand increases further when marker information is added for doing the GWAS or the GS. Computation time is a crucial factor particularly for GWAS, where a large number of markers have to be screened, requiring a separate analysis for each marker. In addition, computation time is a determinant factor for both GWAS and GS when performing cross validation. Application of stage-wise analysis is a common procedure to alleviate the computation burden in MET analysis and for GWAS and /or GS. The first stage of stage-wise analysis is the calculation of adjusted genotype means across all environments, followed by GWAS or GS in the second stage using the adjusted means as a dependent variable and marker effects as an independent variable (Stich et al., 2008; Möhring and Piepho, 2009; Piepho et al., 2012a; Schulz-Streeck et al., 2013a). In order to further reduce the computation time, the phenotypic data can also be analysed in two stages (Stich et al., 2008; Möhring and Piepho, 2009; Piepho et al., 2012a; Schulz-Streeck et al., 2013a; Damesa et al., 2017). Data from MET typically display heterogeneity of variance between trials. If the data is analysed in stages, a weighting approach has been used as a remedy by different authors (Smith et al., 2001; Möhring and Piepho, 2009; Welham et al., 2010; Piepho et al., 2012a; Gogel et al.,

2018) to account for heterogeneity. The optimal choice of weighting method for the second stage is an important question. Generally, the weights are derived from the variances and covariances of adjusted means from the previous stage's analysis. There are different weighting methods, e.g. fully-efficient weighting, where the full variance-covariance matrix is carried forward (Piepho et al., 2012a; Damesa et al., 2017), and diagonal weighting, where the inverse of the diagonal element of the inverse variance-covariance matrix is used as a weight (Smith et al., 2001; Möhring and Piepho, 2009). The most efficient method is to use the full variance-covariance matrix of the adjusted means from the previous stage (fully-efficient weighting), because it usually produces quite similar results to single-stage analysis (Damesa et al., 2017). In most studies researchers compute genotype means from MET data in single-stage analysis or in stage-wise analysis (with or without weighting) and then feed these means into GWAS or GS analysis, often without any weighting. When using stage-wise analysis, the weighting method to be used for weighting of means from the first step is important for minimizing loss of information when forwarding results to the following step. Moreover, the weighting methods also help to obtain results which are close to the gold standard of single-stage analysis (Piepho et al., 2012a; Damesa et al., 2017). The objective of this study, therefore, is to compare weighted stage-wise analysis versus unweighted stage-wise analysis for GWAS and GS using phenotypic and genotypic maize data.

4.2 Materials and statistical methods

4.2.1 The Phenotypic and Genotypic Data

The data for this study consists of 418 improved maize genotypes from the African soils association mapping (IMS-AM) panel, obtained from the International Maize and Wheat Improvement Centre (CIMMYT) Global Maize Program. These genotypes are inbred lines, which represent tropical/subtropical maize germplasm, derived from breeding programs targeting tolerance to soil acidity, low N, resistance to insects and pathogens. Out of the 418 genotypes only 381 genotypes were genotyped using genotyping-by-sequencing (GBS), and all genotypes were phenotyped for various traits under water-stress and well-watered environments plus for Maize Lethal Necrosis disease (MLND) (Gowda et al., 2015). In this study we focus on the analysis of yield. Six field trials were conducted in five locations in

2011 and three field trials in three locations in 2012. In total six Kenyan sites were used. In all trials, an α-lattice design was used (see Table 1 for details). The original marker data for this study contains 955695 SNPs, however only 21966 SNPs were considered for our analysis after quality control. For quality control we excluded SNPs which were monomorphic, had missing values >1% (Sverrisdóttir et al., 2018) and a minor allele frequency (MAF) < 0.05. Missing values were imputed using a random imputation method. After missing value imputation, again SNP with MAF < 0.05 were removed and finally 21966 markers remained. The quality check and recoding of alleles were done using the 'Synbreed' R package (Wimmer et al., 2012).

Table 1. Description of the field experiments. Year of experiments, trial name, replicates, number of blocks, block size, number of genotypes, trial number, site name and site code.

Year	Trial name	Number of replicates	Number of blocks	Block size	Number of genotypes	Trial number	Site name	Site code
2011	KITOPT11A	2	64	6	384	5	Kitale	1
2011	KBKOPT11A	2	50	7	350	6	Kiboko	2
2011	KTLOPT11B	2	64	6	384	8	Kitale	1
2011	KKMOPT11A	2	64	6	384	9	Kakamega	3
2011	AGFOPT11A	2	50	7	350	11	AguaFria	4
2011	CDROPT11B	2	66	5	330	13	Cedara	5
2012	KBKOPT12A	2	44	7	308	17	Kiboko	2
2012	KBSOPT12A	2	44	7	308	19	Kibos	6
2012	KKGOPT12B	2	48	7	336	21	Kakamega	3

4.2.2 Statistical methods

Mixed model for genome-wide association studies and genomic selection

Both GS and GWAS analyses are usually conducted in two stages, where the first stage is the analysis of the phenotypic data and in the second stage GS or GWAS is performed using genetic marker data. The phenotypic analysis can also be done in two stages where in the first stage genotype means are computed per individual trial and in the second stage adjusted genotype means are computed across trials. If the phenotypic analysis is done in two stages, the whole analysis for GS or GWAS will take three stages. In this study we consider two-stage and three-stage analyses for GS and GWAS.

Two-stage analysis for GWAS or GS

First stage: Phenotypic analysis

The following single-stage linear mixed model was assumed for the plot data to perform the phenotypic analysis:

$$y_{ijhvkm} = \mu_i + t_{jhv} + r_{jhvk} + b_{jhvkm} + s_j + a_h + gs_{ij} + ga_{ih} + sa_{jh} + sag_{jhi} + e_{ijhvkm} \qquad (1)$$

where y_{ijhvkm} is the phenotypic observation (yield) for the *i*-th genotype in the *j*-th site, *h*-th year, *v*-th trial, *k*-th replicates, and *m*-th block, μ_i is the expected value of the *i*-th genotype and it is regarded as fixed effect, t_{jhv} is the random effect of the *v*-th trial nested within *j*-th site and *h*-th year with $\text{var}(t_{jhv}) = \sigma^2_{t(jhv)}$, r_{jhvk} is the random effect of the *k*-th replicate nested within the *j*-th site, *h*-th year and *v*-th trial with $\text{var}(r_{jhvk}) = \sigma^2_{r(jhv)}$, b_{jhvkm} is the random effect of the *m*-th block nested within the *j*-th site *h*-th year *v*-th trial and *k*-th replicate with $\text{var}(b_{jhvkm}) = \sigma^2_{b(jhv)}$, s_j is the random main effect of the *j*-th site with $\text{var}(s_j) = \sigma^2_s$, a_h is the random main effect of the *h*-th year with $\text{var}(a_h) = \sigma^2_a$, gs_{ij} is the random interaction effect of the *i*-th genotype and the *j*-th site, ga_{ih} is the random interaction effect of the *i*-th genotype and *h*-th year with $\text{var}(ga_{ih}) = \sigma^2_{ga}$, sa_{jh} is the random interaction effect of the *j*-th site and *h*-th year, sag_{jhi} is the random interaction effect of the *j*-th site, *h*-th year and *i*-th genotype with $\text{var}(sag_{jhi}) = \sigma^2_{sag}$, and e_{ijhvkm} is the residual plot error associated with y_{ijhvkm} with $\text{var}(e_{ijhvkm}) = \sigma^2_{e(jhv)}$. In model (1) the variances for replicate, block and error are assumed to be trial-specific. This assumption allows a stage-wise analysis to be fully equivalent to single-stage analysis (Piepho et al., 2012a), and it is also usually a more realistic assumption than homogeneity of variance across trials (So and Edwards, 2011). Fitting this model produces adjusted genotype means, representing estimates of μ_i. Since we have marker information for only 381 genotypes, after obtaining the adjusted mean of the 418 genotypes from the joint analysis we dropped out those 37 genotypes without marker information before continuing to the actual GWAS and GS analysis stage.

Second Stage: GWAS and GS Analysis

At the second stage, the adjusted means from the phenotypic analysis are used as the response variable in a model of the form

$$\hat{\mu}_i = \mu_i + e_i \tag{2}$$

where $\hat{\mu}_i$ is the adjusted genotype mean of the i-th genotype from the first stage and e_i the residual error associated with $\hat{\mu}_i$. We consider eq (2) as a representation of a general model for the trait-marker association analysis. Plugging in the following regression models (3a) and (3b) for the mean μ_i into eq (2) will give us the full model for GWAS and GS, respectively:

$$\mu_i = \phi + g_i + \beta_w x_{iw} + \sum_{u=1}^{z} Q_{iu} v_u \text{ and} \tag{3a}$$

$$\mu_i = \phi + g_i , \tag{3b}$$

where x_{iw} is the *w*-th SNP marker covariate for the i-th genotype and β_w is the fixed effect of the *w*-th marker, ϕ is a general intercept, g_i is the random genetic effect of i-th genotype. In eq. (3a), Q_{iu} is an element of the i-th row and u-th column of the population structure matrix $\boldsymbol{Q}_{(n \times z)}$, which is an n by z matrix, where n is the number of genotypes and z is the number of sub-populations, v_u is the fixed effect of the u-th column of the population structure matrix $\boldsymbol{Q}$. Note that $\boldsymbol{Q}$ can be calculated, e.g., from the marker data using either the STRUCTURE software (Pritchard 2000) or it can be derived from principal component analysis (PCA). After the PCA, the first z PCA axes are selected as the $\boldsymbol{Q}$ matrix (Zhao et al., 2007). Some studies confirm that GWAS based on the $\boldsymbol{Q}$ matrix calculated from STRUCTURE and using PCA are almost the same, however using STRUCTURE is computationally intensive particularly if the data is large. Therefore it can easily be substituted by the PCA method (Zhao et al., 2007; Stich et al., 2008). For this study, we use the PCA approach implemented in the TASSEL software (Bradbury et al., 2007). Based on the eigenvalues calculated by the PCA for the maize data, we used the first three principal components (PC) as a covariate to correct for the population structure. To choose the optimal number of PC, a scree plot can be used, which displays the eigenvalues or proportion of the sum of eigenvalues versus its order number or rank (Jollife, 2002, p.117). Approximately the point at which the plot levels off or the changes between consecutive eigenvalues becomes small is usually suggested as a cut-off to determine the number of PC to be included in the Q matrix (Fig. 1). It should also be mentioned that Gowda et al. (2015), who used this data for GWAS and GS analysis, also concluded three PC should be included.

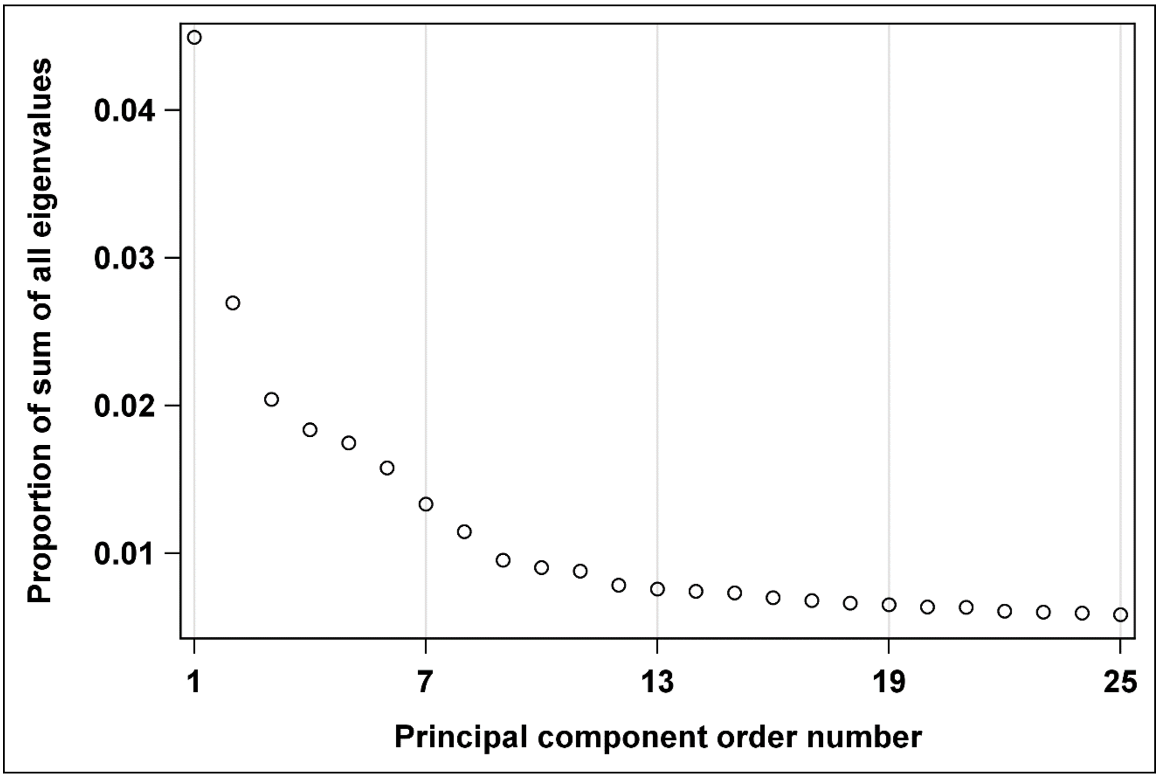


Fig. 1 Proportion of sum of all eigenvalues versus principal component (PC) number.

The random genotypic effect g_i can be collected into a vector $g = (g_1, g_2, \ldots, g_n)^T$, where n is the number of genotypes. The variance of g is given by

$$\mathrm{var}(g) = 2K\sigma_g^2, \tag{4}$$

where K is a kinship matrix which contains coefficients which provide information about the covariance between individuals. K can be determined from pedigree records or from marker-based information of the genotypes (Henderson, 1985; Yu et al., 2006). We estimate the realized marker based kinship matrix K using the following formula (Piepho, 2009; Piepho et al., 2012b)

$$K = ZZ^T \tag{5}$$

where Z is the marker matrix coded by 0 and 2 for homozygous markers and 1 for heterozygous markers. Collecting the errors e_i of eq. (2) into vector $e = (e_1, ..., e_n)^T$, then $Var(e) = \Omega$ where Ω is a $n \times n$ matrix which contains the error variances and covariances of the 381 genotypes which have both phenotypic and marker information. For the fully efficient weighting method, the variance-covariance matrix of the errors is set equal to $\Omega_1 = \Omega$, whereas for the diagonal weighting the variance-covariance matrix is approximated by $\Omega_2 = \left(diag\left(\Omega^{-1}\right)\right)^{-1}$, where $diag\left(\Omega^{-1}\right)$ is a square matrix which contains the diagonal element of Ω^{-1} (Smith et al., 2001; Damesa et al., 2017). In both fully efficient and diagonal weighting Ω_1 and Ω_2 are derived from the previous stage of the analysis. For the unweighted analysis the variance-covariance matrix has the form $I\sigma^2$, where I is an identity matrix and σ^2 is the variance.

Three-stage analysis for GWAS or GS

In three-stage analysis, the phenotypic analysis is done in two stages and at the third stage the trait-marker association is done by GWAS or GS (Piepho et al., 2012a).

First-stage analysis: In the first stage genotype means are computed per trial, site and year based on the model

$$y_{ijhvkm} = \mu_{ijhv} + r_{jhvk} + b_{jhvkm} + e_{ijhvkm} \tag{6}$$

where $\mu_{ijhv} = \mu_i + t_{jhv} + s_j + a_h + gs_{ij} + ga_{ih} + sa_{jh} + sag_{jhi}$ is the conditional expected value of the *i*-th genotype $(i = 1, ..., n)$ at the *j*-th site, *v*-th trial and *h*-th year, with effects as defined previously.

Second-stage analysis: In the second stage the adjusted means are computed across site-year-trial combination using the model

$$\hat{\mu}_{ijhv} = \mu_i + t_{jhv} + s_j + a_h + gs_{ij} + ga_{ih} + sa_{jh} + sag_{jhi} + e_{ijhv} \tag{7}$$

where e_{ijhv} is the residual of the *i*-th genotype in the *j*-th site, *h*-th year and *v*-th trial and $\text{var}(e_{jhv}) = \Omega_{jhv}$, with $e_{jhv} = (e_{1jhv}, e_{2jhv}, ..., e_{njhv})^T$, where Ω_{jhv} is replaced by its residual maximum likelihood (REML) estimate from the first stage, fully efficient Ω_1 or a diagonal approximation thereof Ω_2. To fit the model in the third stage, we need the variance-covariance matrix of the adjusted genotype means from the joint analysis, i.e., $e^T = (e_1^T, ..., e_M^T)$, where *M* represents the total number of environments (trial-site-year). Then the fully efficient and diagonal weights are computed from the variance-covariance matrix $Var(e)$. After estimating the adjusted means from the joint analysis using (7) at the third stage, the GWAS or GS analysis are done using the means from stage two.

Third-stage analysis: At the third stage the GWAS or GS are computed using the model (3a) or (3b).

In all stage-wise analyses we used the SAS macro ***%get_one_big_omega*** and ***%get_Smith_weights*** for fully-efficient and diagonal weight method, respectively. These macros assemble the adjusted means and the fully-efficient or diagonal weights from previous stage in a way suitable for use in the next-stage analysis (Damesa et al., 2017).

For ease of reference, we subsequently use the abbreviations in Table 2 to refer to the various single-stage and stage-wise approaches in the text.

Table 2. Abbreviations for the different stage-wise approaches to be used in the writing of the result and discussion sections.

Abbreviation	**Meaning**
1S	Single-stage analysis
2S	Two-stage analysis

3S	Three-stage analysis
$2S_{UNW}$	Two-stage analysis unweighted
$2S_{DIAGW}$	Two-stage analysis with diagonal weights
$2S_{FEW}$	Two-stage analysis with fully efficient weighting
$3S_{UNW}$	Three-stage analysis unweighted
$3S_{DIAGW}$	Three-stage analysis with diagonal weights
$3S_{FEW}$	Three-stage analysis with fully efficient weighting

We use superscripts P, GS, and GWAS in the abbreviations to represent specific types of analysis, i.e. P for purely phenotypic analysis, GS for genomic selection and GWAS for genome-wide association studies, respectively. For example to represent two-stage analysis with fully efficient weighting we used the abbreviation $2S_{FEW}^{P}$ for the phenotypic analysis, $2S_{FEW}^{GS}$ for genomic selection and etc.

Comparison of methods for the different stage-wise analysis methods for GWAS

The vast majority of markers under study are expected to be unlinked to the influential QTL, and thus under the null hypothesis of the test. For these markers, the P-values observed from the association study are expected to follow a uniform distribution. Even though not all markers will be under the null hypothesis, the empirical distribution of P-values across all markers is expected to be approximately uniform if the tests perform properly. For comparing the different weighting methods, we consider the empirical type I error rate. A method is considered best if it has empirical type I error rate in close agreement with the nominal type I error rate. To measure the type I error rate, we therefore use the mean squared difference (MSD) between observed and expected ordered P-values of all markers, assuming that expected P-values follow a uniform distribution. A high MSD implies a deviation of observed P-values from the expected uniform distribution. Therefore we consider a given method best for GWAS if it has the smallest MSD value (Stich et al., 2008). Originally, this type comparison was used to assess the degree of control of population structure by a GWAS method, noting that failure to properly control for such structure typically leads to spurious tests and hence to a departure from the expected uniform distribution of P-values. Here, we use the same rationale to assess the effect of not properly accounting for heterogeneity of variance and correlation among adjusted genotype means. For each observed P-values from the GWAS the expected P-values were calculated as $r(p_w)/W$, where $r(p_w)$ is the rank of the p-value p_w observed for the w-th marker and W represents the total number of markers (Stich et al., 2008). We compare the correlation of the P-values from the GWAS for the different approaches in order to quantify their degree of similarities. In addition we use pairwise plots of SNP effect estimate, P-value, standard error, and observed $-\log_{10}(P-value)$ to assess the similarity of the different approaches. If two methods are similar in the GWAS analysis, these plot should show most of the points near to the diagonal line; if they are different the points are expected to deviate from the diagonal line.

Comparison of methods for GS analysis

To compare the similarity of GS methods with weighting and without weighting we use the Spearman and Pearson correlation between the predicted GEBVs. We use cross validation (CV) to compare the predictive ability of the different methods. In k-fold CV, k=5 or k=10 is the commonly used approach with large sample sizes. The sizes of the training and validation sets have direct impacts on the accuracy of the estimated variance components and correlations. Moreover, high correlations can be obtained with a large reference set and a small validation set. However, if the objective is to compare different models for GS then a relatively large validation set is desirable (Erbe et al., 2010). Therefore, because of small sample size and our objective of model comparison, in this study we used a three-fold CV for all of the approaches considered (Estaghvirou et al., 2013). The procedure randomly divides the n genotypes, i.e., the n adjusted means obtained from the phenotypic analysis stage using weighted or unweighted methods, into three subsets. Two of the subsets serve as a training set and the remaining subset is used as a validation set. Each subset is used as a validation set once. The three-fold CV yielded a validation set with 127 and a training set with 254 observations. The process is repeated five times (Schulz-Streeck et al., 2013a; Estaghvirou et al., 2013; Song et al., 2017; Rice and Lipka, 2019; Palaiokostas et al., 2019). The prediction ability is calculated as the Pearson and Spearman correlation between observed and predicted values in the validation set for each replicate. The three-fold CV and five repetitions yield 15 sets of GS predictions. The correlations are averaged over the 15 sets. The method with the highest prediction ability is considered as the best method.

4.3 Results

4.3.1 Comparison of two-stage analysis with weighting and with-out weighting and single-stage phenotypic analysis

First, we analyzed the phenotype in two stages with fully-efficient weighting ($2S_{FEW}^{P}$), with diagonal weighting ($2S_{DIAGW}^{P}$) and without weighting ($2S_{UNW}^{P}$). For comparison we also analyzed the phenotypic data using single-stage analysis ($1S^{P}$). As in many other studies, the estimated adjusted genotype means show the similarity of 1S and 2S analysis in general. The adjusted genotype means based on the $2S_{FEW}^{P}$ were more highly correlated with 1S as compared to $2S_{DIAGW}^{P}$ and $2S_{UNW}^{P}$ (Table 3, Fig. S1).

	$2S_{UNW}^{P}$	$2S_{DIAGW}^{P}$	$2S_{FEW}^{P}$	$1S^{P}$
$2S_{UNW}^{P}$	1.0000	0.9917	0.9894	0.9863
$2S_{DIAGW}^{P}$	0.9881	1.0000	0.9978	0.9972
$2S_{FEW}^{P}$	0.9848	0.9968	1.0000	0.9994
$1S^{P}$	0.9803	0.9959	0.9988	1.0000

In order to assess the distribution of the weights, we used histograms. Since the fully-efficient weighting is derived from the full variance-covariance matrices, we investigate the heterogeneity using separate plots for the variances as well for the covariances of the adjusted means. The histogram of the covariances (Fig. S2a) is right skewed. The distribution of the variances is also skewed to the right (Fig. S2b). To inspect the distribution of the diagonal weights we plotted the diagonal elements of the inverse variance-covariance matrix of adjusted means (Fig. S3). This histogram is skewed to the left. All histograms display considerable heterogeneity.

4.3.2 Comparison among three-stage and two-stage genomic prediction analysis with weighting and without weighting

Correlation of GEBVs

The correlations of GEBVs obtained using $3S_{DIAGW}^{GS}$ and $3S_{FEW}^{GS}$ are relatively larger than the corresponding correlations with the GEBVs obtained from $3S_{UNW}^{GS}$ and $2S_{UNW}^{GS}$. The $2S_{UNW}^{GS}$ method is highly correlated with $3S_{UNW}^{GS}$. $2S_{DIAGW}^{GS}$ is almost perfectly correlated with $3S_{FEW}^{GS}$, followed by $3S_{DIAGW}^{GS}$. $2S_{FEW}^{GS}$ is also highly correlated with $3S_{FEW}^{GS}$ and $3S_{DIAGW}^{GS}$. The comparison of 2S methods among themselves has a similar trend as the 3S methods, i.e., $2S_{FEW}^{GS}$ is extremely highly correlated with $2S_{DIAGW}^{GS}$ and it is also highly correlated with $2S_{UNW}^{GS}$ (Table 4 and Fig. S4).

Table 4. Correlation between GEBVs (above the diagonal: Pearson's product-moment correlation; below the diagonal: Spearman's rank correlation) using $3S_{UNW}^{GS}$, $3S_{DIAGW}^{GS}$, $3S_{FEW}^{GS}$, $2S_{UNW}^{GS}$, $2S_{DIAGW}^{GS}$ and $2S_{FEW}^{GS}$.

Analysis Method	$3S_{UNW}^{GS}$	$3S_{DIAGW}^{GS}$	$3S_{FEW}^{GS}$	$2S_{UNW}^{GS}$	$2S_{DIAGW}^{GS}$	$2S_{FEW}^{GS}$
$3S_{UNW}^{GS}$	1.0000	0.9705	0.9681	0.9885	0.9668	0.9661
$3S_{DIAGW}^{GS}$	0.9796	1.0000	0.9984	0.9745	0.9980	0.9979
$3S_{FEW}^{GS}$	0.9781	0.9980	1.0000	0.9757	0.9994	0.9995
$2S_{UNW}^{GS}$	0.9859	0.9852	0.9872	1.0000	0.9768	0.9764
$2S_{DIAGW}^{GS}$	0.9762	0.9976	0.9992	0.9882	1.0000	0.9999
$2S_{FEW}^{GS}$	0.9753	0.9974	0.9993	0.9877	0.9998	1.0000

Correlation of the GEBVs and the observed values in the validation set

The mean Pearson and Spearman correlations indicate that all methods have similar results. However, even if the differences are minor, the mean Pearson correlations show that the $2S_{FEW}^{GS}$ method has the best predictive ability, followed by $3S_{FEW}^{GS}$, and $2S_{DIAGW}^{GS}$ is the third. On the other hand the mean Spearman correlation suggests that $3S_{FEW}$ is the best method, followed by $2S_{FEW}^{GS}$. $2S_{DIAGW}^{GS}$ and $3S_{UNW}^{GS}$ have the same rank, both in terms of the Pearson and Spearman rank correlation (Table 5).

Table 5. Mean Pearson's product-moment and Spearman's rank correlation coefficient between the GEBVs and the observed values in the validation set. Using $3S_{UNW}^{GS}$, $3S_{DIAGW}^{GS}$, $3S_{FEW}^{GS}$, $2S_{UNW}^{GS}$, $2S_{DIAGW}^{GS}$ and $2S_{FEW}^{GS}$.

Analysis method	Mean Pearson correlation	Rank of mean Pearson correlation	Mean Spearman correlation	Rank of mean Spearman correlation
$3S_{UNW}^{GS}$	0.3249	6	0.3361	6
$3S_{DIAGW}^{GS}$	0.3352	4	0.3366	5
$3S_{FEW}^{GS}$	0.3442	2	0.3478	1
$2S_{UNW}^{GS}$	0.3300	5	0.3404	4
$2S_{DIAGW}^{GS}$	0.3414	3	0.3437	3
$2S_{FEW}^{GS}$	0.3453	1	0.3468	2

4.3.3 GWAS

The MSD between observed and expected P-values are in general very small for all methods used in this study. However, there are some numerical differences among the MSD of the six methods. Among all methods the MSD of $3S_{DIAGW}^{GWAS}$ is the smallest, followed by the $3S_{FEW}^{GWAS}$

method. $2S_{UNW}^{GWAS}$ is the least performant as it has the largest MSD value, the second least performant is $3S_{UNW}^{GWAS}$. The $3S_{FEW}^{GWAS}$, $2S_{FEW}^{GWAS}$, and $2S_{DIAGW}^{GWAS}$ methods had quite similar MSD values, meaning that these methods had approximately equal performance (Table 6). The P-values of $3S_{UNW}^{GWAS}$ are highly correlated with those of $2S_{UNW}^{GWAS}$. $3S_{DIAGW}^{GWAS}$ has almost perfect correlation with $3S_{FEW}^{GWAS}$, however, $3S_{FEW}^{GWAS}$ is less correlated with $2S_{UNW}^{GWAS}$, and $3S_{FEW}^{GWAS}$ has very high correlation with the $2S_{FEW}^{GWAS}$ and $2S_{DIAGW}^{GWAS}$ methods. Comparison of the results for 2S and 3S shows that in both categories the diagonal and fully efficient weighting methods perform similarly as indicated by their high correlation of 0.9913 and 0.9996, respectively. These trends hold true for both the Pearson and Spearman correlations (Table 7).

Table 6. Mean square difference of P-values for GWAS: using $3S_{UNW}^{GWAS}$, $3S_{DIAGW}^{GWAS}$, $3S_{FEW}^{GWAS}$, $2S_{UNW}^{GWAS}$, $2S_{DIAGW}^{GWAS}$ and $2S_{FEW}^{GWAS}$.

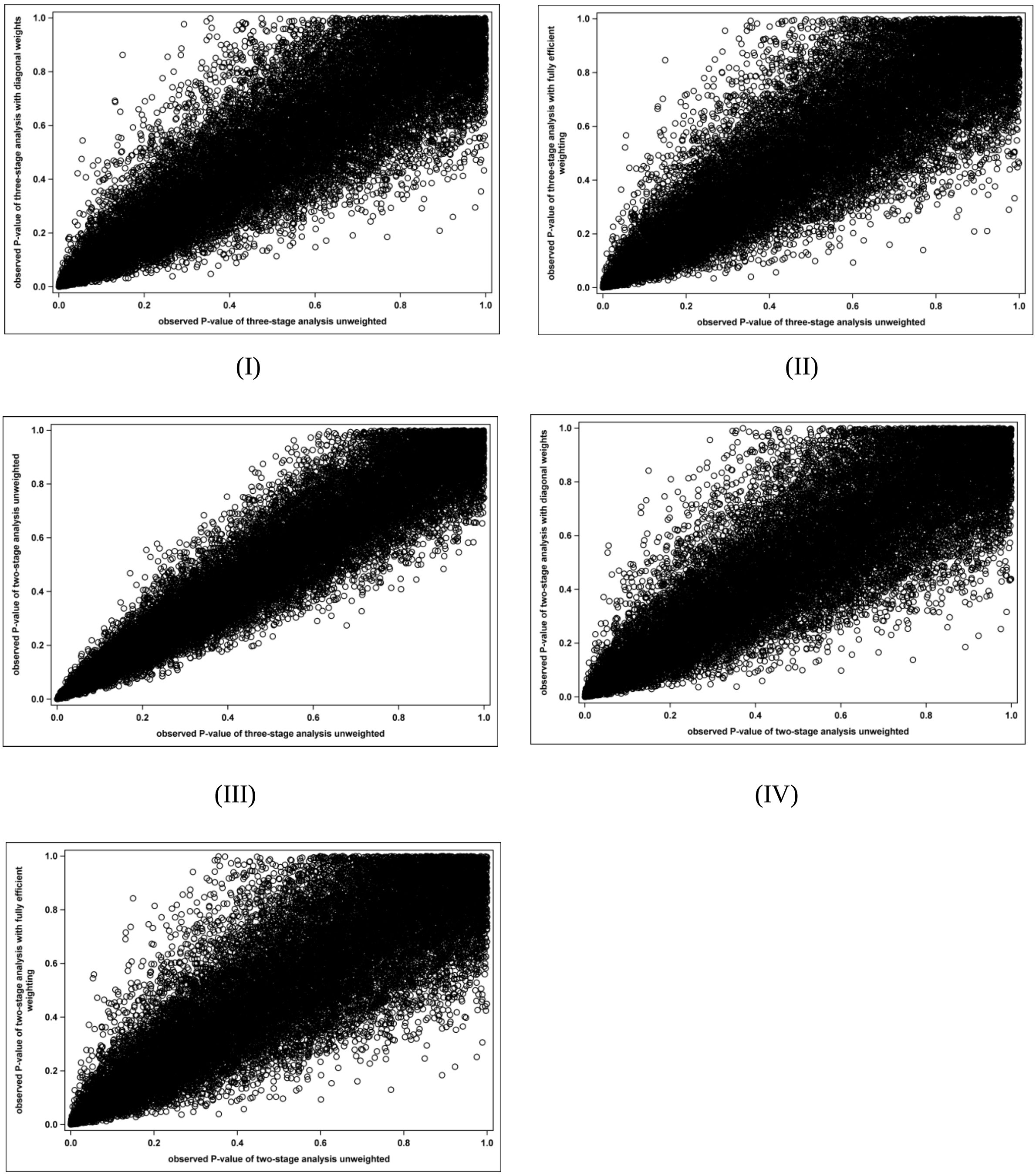
Analysis method	MSD value
$3S_{UNW}^{GWAS}$	13.3010000E-6
$3S_{DIAGW}^{GWAS}$	3.3508581E-6
$3S_{FEW}^{GWAS}$	4.1124605E-6
$2S_{UNW}^{GWAS}$	14.9080000E-6
$2S_{DIAGW}^{GWAS}$	4.6586410E-6
$2S_{FEW}^{GWAS}$	4.7140910E-6

The P-values of $3S_{UNW}^{GWAS}$ are highly correlated with those of $2S_{UNW}^{GWAS}$. $3S_{DIAGW}^{GWAS}$ has almost perfect correlation with the $3S_{FEW}^{GWAS}$, however, it is less correlated with the $2S_{UNW}^{GWAS}$ and . $3S_{FEW}^{GWAS}$ has very high correlation with the $2S_{FEW}^{GWAS}$ and $2S_{DIAGW}^{GWAS}$ methods. Comparison of the results among 2S and 3S shows that in both categories the diagonal and fully efficient weight perform similarly as indicated by their high correlation of 0.9913 and 0.9996, respectively. These trends hold true for both the Pearson and Spearman correlations (Table 7).

Table 7. Correlation between observed P values of GWAS (above the diagonal: Pearson's product-moment correlation; below the diagonal: Spearman's rank correlation) using $3S_{UNW}^{GWAS}$, $3S_{DIAGW}^{GWAS}$, $3S_{FEW}^{GWAS}$, $2S_{UNW}^{GWAS}$, $2S_{DIAGW}^{GWAS}$ and $2S_{FEW}^{GWAS}$.

Analysis method	$3S_{UNW}^{GWAS}$	$3S_{DIAGW}^{GWAS}$	$3S_{FEW}^{GWAS}$	$2S_{UNW}^{GWAS}$‡	$2S_{DIAGW}^{GWAS}$	$2S_{FEW}^{GWAS}$
$3S_{UNW}^{GWAS}$	1.00000	0.88833	0.87908	0.94724	0.87168	0.86989
$3S_{DIAGW}^{GWAS}$	0.88871	1.00000	0.99127	0.89753	0.98910	0.98857
$3S_{FEW}^{GWAS}$	0.87966	0.99132	1.00000	0.90191	0.99672	0.99721
$2S_{UNW}^{GWAS}$	0.94765	0.89785	0.90235	1.00000	0.90581	0.90416
$2S_{DIAGW}^{GWAS}$	0.87236	0.98918	0.99674	0.90632	1.00000	0.99957
$2S_{FEW}^{GWAS}$	0.87059	0.98865	0.99723	0.90469	0.99957	1.00000

All of the pairwise plots, i.e. the P-value plots, the minus $\log_{10}$ P-value plots, the pairwise plots of SNP effects, and the pairwise plots of standard errors have similar interpretation for the comparison of the methods for this study. These plots also show similar results to the correlation of P-values. The P-value plots of $3S_{UNW}^{GWAS}$ versus $2S_{UNW}^{GWAS}$ (Fig. 2a) have a rather diagonal shape, which suggests similarity of the two sets of P-values. The P-value plots show that the $3S_{DIAGW}^{GWAS}$ method has close similarities with the $3S_{FEW}^{GWAS}$, $2S_{DIAGW}^{GWAS}$ and $2S_{FEW}^{GWAS}$ methods (Fig. 2b), because points of these plots are close to the diagonal axis. Likewise $3S_{FEW}^{GWAS}$ has similar performance with $3S_{DIAGW}^{GWAS}$, $2S_{DIAGW}^{GWAS}$ and $2S_{FEW}^{GWAS}$ (Fig. 2.c). The supplemental figures also suggest similar interpretation to the P-value plots (Figs. S5, S6, and S7).



(I)

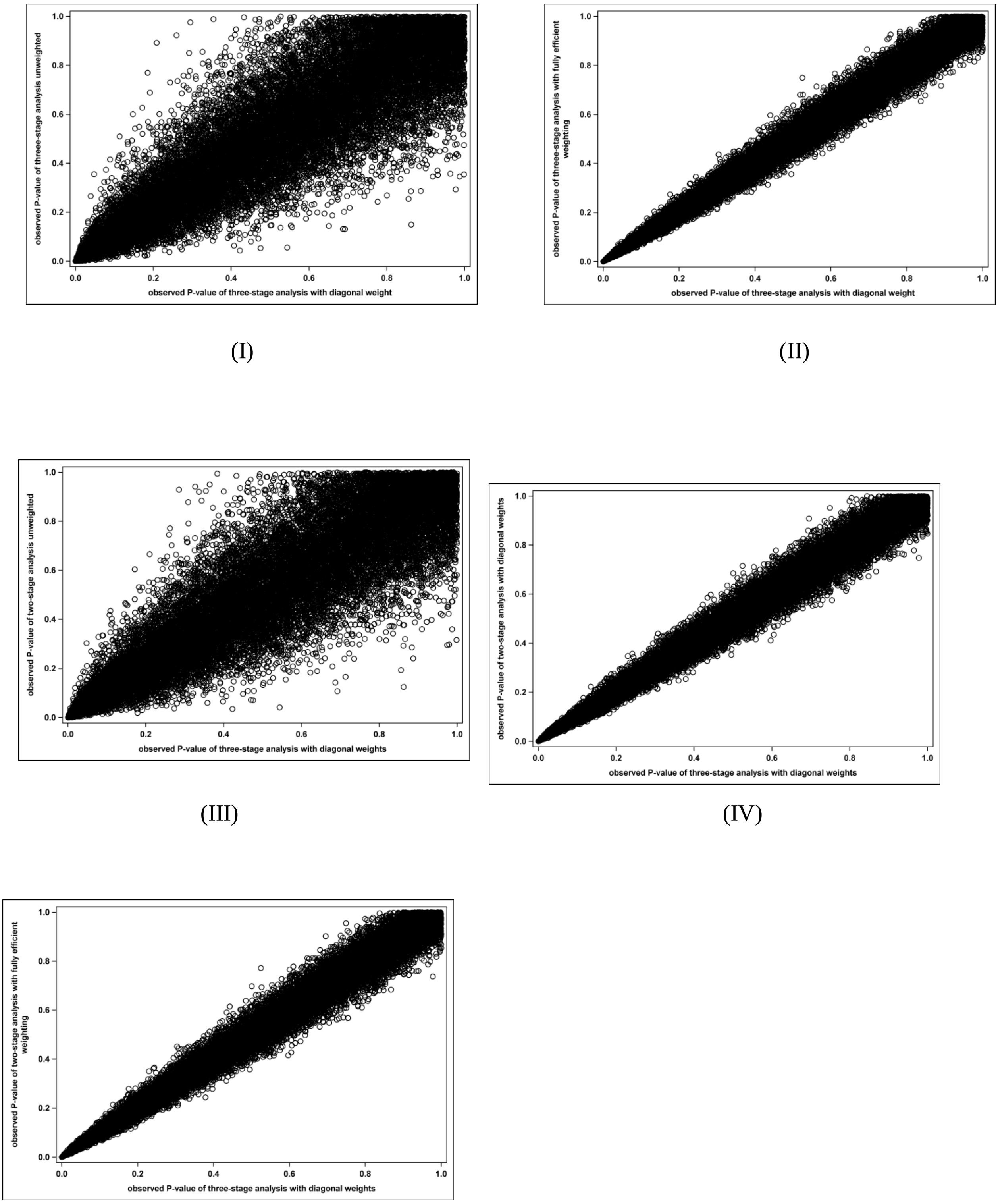
(II)

(III)

(IV)

(V)

Fig.2.a. Plots of P-value of $3S_{UNW}^{GWAS}$ versus (I) $3S_{DIAGW}^{GWAS}$, (II) $3S_{FEW}^{GWAS}$ (III) $2S_{UNW}^{GWAS}$ (IV) $2S_{DIAGW}^{GWAS}$ and (V) $2S_{FEW}^{GWAS}$



(I)

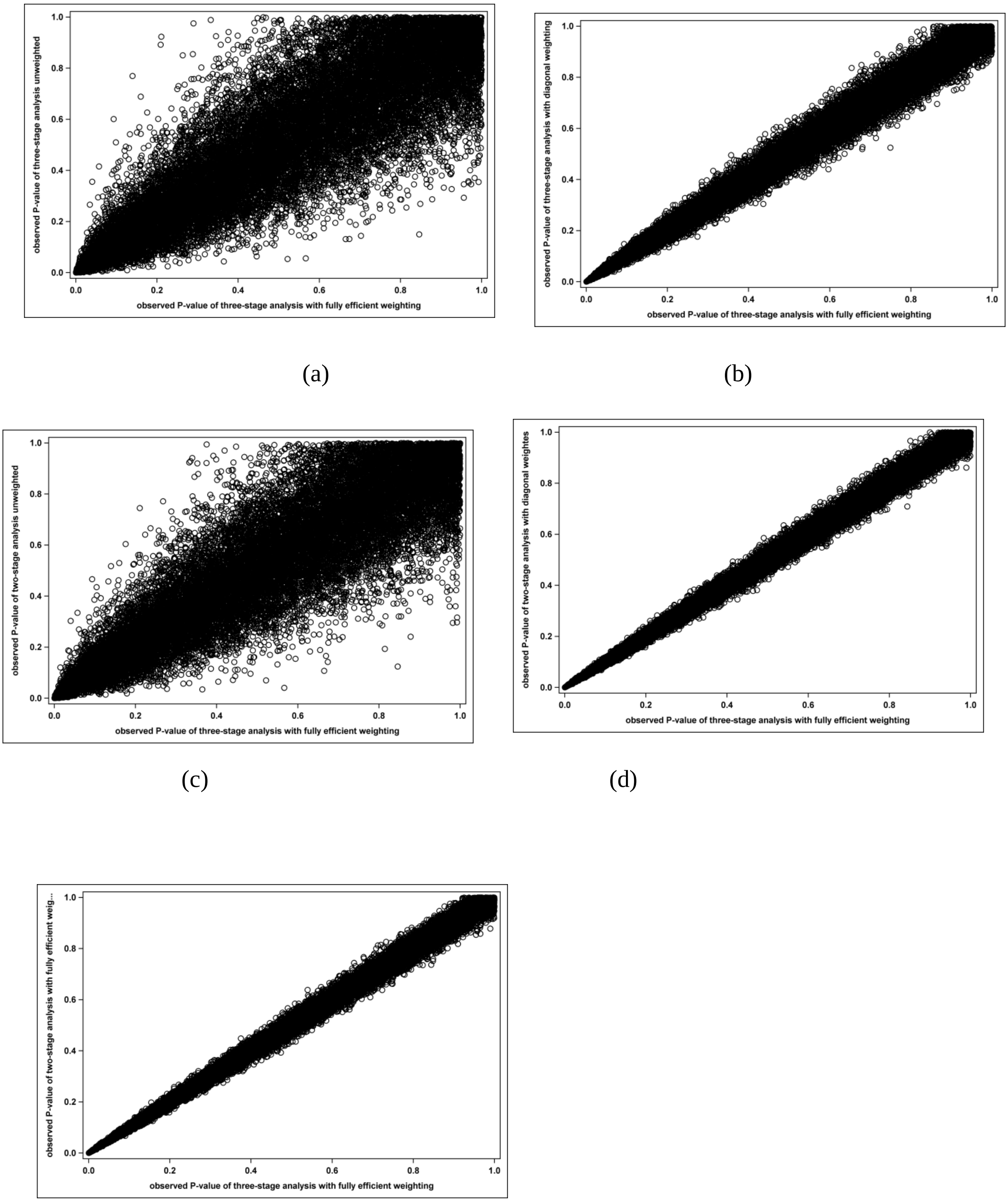
(II)

(III)

(IV)

(v)

Fig.2.b Plots of P-value of $3S_{DIAGW}^{GWAS}$ versus (I) $3S_{UNW}^{GWAS}$, (II) $3S_{FEW}^{GWAS}$ (III) $2S_{UNW}^{GWAS}$ (IV) $2S_{DIAGW}^{GWAS}$ and (V) $2S_{FEW}^{GWAS}$



(a)

(b)

(c)

(d)

(e)

Fig.2.c Plots of P-value of $3S_{FEW}^{GWAS}$ versus (a) $3S_{UNW}^{GWAS}$, (b) $3S_{DIAGW}^{GWAS}$ (c) $2S_{UNW}^{GWAS}$ (d) $2S_{DIAGW}^{GWAS}$ and (e) $2S_{FEW}^{GWAS}$

4.4. Discussion

4.4.1 Phenotypic data analysis

The comparison of the 1S and 2S methods for the phenotypic data shows that $2S_{FEW}$ is most closely correlated with 1S. Similar results were reported in other studies (Piepho et al., 2012a; Damesa et al., 2017). In MET, each trial may require a different design, possibly with high levels of spatial variability, calling for different modelling approaches for each individual trial. The use of spatial modeling techniques helps to account the existing spatial variability and increases the efficiency of analysis (Bernal et al., 2014; Damesa et al., 2018). However, using different modeling for different trials induces complex variance-covariance structures and thereby increases computational demand for analysis. In such cases, stage-wise analysis is a convenient option for practical data analysis. There are several stage-wise analysis methods based on the weighting methods used, from simplest weighting using, e.g., inverse squared standard errors, to the more efficient ones, e.g., fully-efficient weighting (Möhring and Piepho, 2009; Piepho et al., 2012a; and Damesa et al., 2017). Some studies revealed, however, that stage-wise analysis can produce similar results with different weighting methods, including unweighted analysis. This is usually true if the MET are balanced and or the covariance between the genotypes means are small or negligible (Möhring and Piepho, 2009; Piepho et al., 2012a; Damesa et al., 2017). However, in so far as convergence is attained and computation time is feasible, the use of the fully efficient weighting method is always recommended to avoid any loss of information (Welham et al., 2010; Damesa et al., 2017).

4.4.2 GWAS

The MSD between observed and expected P-values for all 3S methods considered in this study are small and have similar values, however the unweighted method is approximately four times larger in MSD than the $3S_{DIAGW}^{GWAS}$ and $3S_{FEW}^{GWAS}$ methods. The MSD of the $3S_{DIAGW}^{GWAS}$ method is smaller than that of the $3S_{FEW}^{GWAS}$, $2S_{DIAGW}^{GWAS}$ and $2S_{FEW}^{GWAS}$ methods. These results imply that for this study using the $3S_{DIAGW}^{GWAS}$ method is advantageous for GWAS. The pairwise correlation and pairwise plots have the same interpretation; these statistics indicate that the $3S_{DIAGW}^{GWAS}$, $3S_{FEW}^{GWAS}$, $2S_{DIAGW}^{GWAS}$ and $2S_{FEW}^{GWAS}$ methods have high pairwise correlation (> 0.98) and also more diagonal shaped pairwise plots, confirming the similar performance of the methods (Fig. 2, and supplement Figs. S5, S6, and S7; Table 7). The $2S_{UNW}^{GWAS}$ method has a moderate correlation (>0.90) with $3S_{FEW}^{GWAS}$, $2S_{DIAGW}^{GWAS}$ and $2S_{FEW}^{GWAS}$, suggesting that $2S_{UNW}^{GWAS}$ is the second best alternative method next to the diagonal and fully efficient weighting methods. However, the $3S_{UNW}^{GWAS}$ method has relatively smaller correlation (between 0.87 and 0.89) with $3S_{DIAGW}^{GWAS}$, $3S_{FEW}^{GWAS}$, $2S_{DIAGW}^{GWAS}$, and $2S_{FEW}^{GWAS}$. This suggests that weighting is worthwhile for an efficient GWAS analysis.

4.4.3 GS

In our study the accuracies obtained for the different methods are nearly the same and is also comparatively small. In this particular study, $2S_{FEW}^{GS}$ and $3S_{FEW}^{GS}$ were the best approaches based on the predictive ability according to Pearson and Spearman correlations, respectively (Table 5 and Fig. S4). This suggests that using all available information of the variance-covariance (i.e., fully-efficient weighting) is the best method to increase precision. On the other hand due to the loss of information for the unweighted method the genomic reliability decreases to some extent.

4.4.4 Statistical software limitations for using weighting in genomic selection and genome-wide association studies

Most currently available open source statistical packages for GS or GWAS do not have options for weighting techniques. All of our stage-wise analyses for the GS and GWAS were conducted using the commercial SAS software and the weights are assembled using a SAS macro called %get_one_big_omega and %get_smith_weight for the fully-efficient and diagonally weighted approximation, respectively (Damesa et al., 2017). ASREML-R is another commercial package that has an option for weighted analysis (Gogel et al., 2018). One exceptional case for GWAS that enables weighted mixed linear model (MLM) analysis is the open source software TASSEL. This package has an option for weighting, but the weighting works only if the weights can be given in one column, e.g. for the diagonal weighting method, however it is not possible to use the fully efficient weighting. For the future it would be desirable to enable all types of weighting in open source software such as TASSEL or R-packages for GWAS and GS analysis.

4.4.5 Possible extensions of the model used in this study

Spatial analysis is one approach to increase precision of phenotypic MET analysis and thereby to increase accuracy of GS and GWAS (Gilmour et al., 1997; Stefanova et al., 2009; Bernal-Vasquez et al., 2014; Damesa et al., 2018). However, in this study spatial modelling was not applied because the field trial data did not have spatial information on the plots (row and column numbers).

In mixed modeling of field trials, the partitioning of the total genetic effect into additive and non-additive effects using pedigree and/or molecular markers is possible and can be superior to analyses ignoring this partitioning. Oakey et al. (2006; 2007) showed that in a single trial and METs, modeling of epistasis and dominance effects can be better than modelling of additive effects only in terms of estimated prediction error and accuracy of selection of the best performing genotypes. In the same vein, value may be added in both GWAS and GS when exploiting of dominance and epistasis effects, e.g. in improving prediction accuracy and obtaining more precise estimates of marker effects (Jiang et al., 2015; Oakey et al., 2016; Bonnafous et al., 2018). However, in this study we did not consider non-additive modeling because our main focus was to demonstrate the benefit of weighting and because analyses based on additive genetic effects only are still the standard approach in both GWAS and GS.

In MET data analysis researchers are often interested in quantifying the genotype by environment interaction (GEI). The GEI variance is often considered to be heterogeneous (Patterson and Nabugoomu, 1992; Frensham et al., 1997; Cullis et al., 1998). To fit heterogeneous GEI variances, different approaches were proposed (Gogel et al., 1995; Piepho, 1997; Smith et al., 2015; Smith et al., 2018). Among others, the factor-analytic (FA) variance structure is considered to be convenient and appropriate for modelling the GEI because this structure accommodates heterogeneity of both variance and covariance. For this study the simple variance components model structure was considered for the GEI. We tried to fit FA structures for the GEI, considering the genotype main effect and the three-way interaction of trial, year and site as fixed effects. The first scenario fits a one-dimensional FA variance structure (FA1) for all of the GEI effects (genotype by site, genotype by year, and genotype by site by year) with independent genotypes and correlated environments at the second stage of a two-stage analyses with the diagonal weight. This first scenario was fitted in SAS, the Akaike information criterion (AIC) indicates that the FA1 was better than the simple variance structure with the same model effects, however, the Hessian matrix was not positive definite for the FA1 model. The second scenario FA1 was fitted only for the three-way interaction (genotype by site by year) with independent genotypes and correlated environments. A single-stage analysis of scenario two was performed in ASReml-R (Butler et al., 2009) and Asreml standalone. But with both there were convergence problems for our dataset, i.e. the average information matrix was found to be singular. Conducting the same single-stage analysis for a simple variance components model using the same model effects as with FA1 above, the model did not have convergence problems. Even though FA1 had a better performance than the simple variance component model in terms of AIC, the fit was not reliable because of the singularity in the likelihood. Also, the correlation between the genotype mean estimates based on FA1 versus the simple variance components model was only 0.71. To study the cause of the singularity we fitted scenario two without the fixed genotype main effect, obtaining proper convergence (in both ASReml-R and ASReml standalone). This suggests that the fixed genotype main effect is interfering with the covariance across environments as it is fitted in the FA structure. Also, the singularity problem did not occur when we fitted scenario two with genotype main effect as random (Gogel et al. 2018). We note, however, that dropping the fixed main effect or fitting it as random is not an option here because BLUE of genotype means are needed for the final GS and GWAS analysis. Therefore we decided not to use the FA1 model for GS and GWAS analysis.

Different markers might have different effects in different environments. Therefore including of a marker-by-environment interaction (MEI) effect in QTL mapping can enhance the accuracy of the QTL mapping (van Eeuwijk et al., 2002; Piepho, 2005). Likewise the modelling of the marker-by-environment effect for GS can possibly increase the predictive accuracy (Piepho, 2009; Crossa et al., 2010). We did not proceed further with this idea, however, because it is computationally more demanding and not crucial for our main objective, which is to compare different methods of weighting using standard procedures.

Moreover, Schulz-Streeck et al. (2013b) showed that there is no gain in prediction accuracy when using the marker information to model the genotype-environment interaction.

A recent study which compares single-stage versus two-stage analysis of crop trials has shown that due to the improvement of computing power in ASREML-R (Butler et al. 2017), single-stage analysis of large-sized MET data is possible (Gogel et al., 2018). The authors suggested that a single-stage analysis of large-sized MET data is plausible, particularly when the number of trials is not too many. However, for many users, computation time and effort are still a concern when trials are large. In such cases the use of stage-wise analysis is a viable alternative (Möhring and Piepho, 2009; Piepho et al., 2012a; Morris et al., 2018).

In GWAS, optimizing power and minimizing type II error rate are useful objectives. However, a full power analysis would be beyond the scope of our paper, as this would require a comprehensive simulation study. It is clear, however, that single-stage analysis is expected to have the best power, provided the model is correctly specified, because it uses empirical best linear unbiased estimators (BLUE) of the marker effects. The two-stage approximations are expected to be slightly lower in power, as they only approximate the single-stage analysis.

4.5 Conclusion

Our result indicates that fully-efficient and diagonal weights have quit similar performance in the phenotypic analysis (> 0.99), on the other hand, the correlation of $2\mathrm{S}_{\mathrm{UNW}}^{\mathrm{P}}$ with the weighted 2S analyses is moderately smaller. Likewise the weighted and unweighted 2S and 3S analyses of GS and GWAS have a similar trend as the phenotypic analysis. The GEBVs obtained using the weighted methods namely $3\mathrm{S}_{\mathrm{DIAGW}}^{\mathrm{GS}}$, $3\mathrm{S}_{\mathrm{FEW}}^{\mathrm{GS}}$, $2\mathrm{S}_{\mathrm{DIAGW}}^{\mathrm{GS}}$ and $2\mathrm{S}_{\mathrm{FEW}}^{\mathrm{GS}}$, have high correlations (>0.99). However, correlations of the GEBVs obtained using $3\mathrm{S}_{\mathrm{UNW}}^{\mathrm{GS}}$ and $2\mathrm{S}_{\mathrm{UNW}}^{\mathrm{GS}}$ are comparatively smaller with the weighted GS. The correlations of the P-values obtained from GWAS using the different approaches have a similar implication as the GS analyses.

To sum up, this study concerned the evaluation of weighted and unweighted stage-wise analysis for GS and GWAS. There are several different statistical methods for GS and GWAS analyses, and each of the different methods are usually applied in two stages, i.e., phenotypic analysis stage and, GWAS and/or GS analysis stage (Stich et al., 2008; Ogutu et al., 2011; Ogutu et al., 2012). In both stages of analyses, researchers can use either weighted or unweighted analysis. Different researchers have different experiences regarding the use of weighting methods. Some researchers use weighting methods for the analysis of phenotypic data, but the unweighted method in the GWAS and GS stage and other researchers use unweighted analysis for both phenotypic and actual GWAS and GS analyses. The correlation plot of the adjusted genotype means from phenotypic analysis (Fig. S1) indicates that all the different methods have similar performance. From these results one may expect to obtain similar performance in GWAS and GS with the different weighted and unweighted approaches like the phenotypic results. However, this assumption needs to be checked with empirical examples in order to quantify the degree of similarity of the different methods. Even though the differences are relatively small, in this particular study the weighting approaches performed better than the unweighted analysis for both GS and GWAS in terms of predictive ability for GS and in terms of MSD of observed and expected P-values for GWAS. The best weighting method found for GS and GWAS are not the same. For our dataset the fully-efficient weighting method performed better than the diagonal weights for GS. By contrast, for GWAS the diagonal weighting method performed better but with very small difference compared to the fully-efficient method. In general this result suggests that there are minor differences between the different approaches and the unweighted method is acceptable for most practical purposes, but there is a slight edge in favor of weighted methods. The fully-efficient weighting method can be recommended provided convergence criteria are met and computation time is feasible. Otherwise, diagonal weights are adequate and by comparison they are advantageous computationally.

4.6 Supplementary material

Supplementary Figures these includes: Fig. S1.a, Fig. S1.b, Fig. S2, Fig. S3, Fig. S4, Fig. S5.a, Fig. S5.b, Fig.S5.c, Fig. S6.a, Fig. S6.b, Fig.S6.c, Fig.S7.a, Fig. S7.b, and Fig.S7.c are made available online.

Chapter 5

General discussion

Crop breeding and improvement involves the evaluation of trials at multiple sites and years (MET). MET have been contributing to the identification of superior and stable genotypes using observed phenotype data. In MET there are many different sources of variations. Important are within-trial error and within-trial spatial variation, and there typically is both within-trial and between-trial error variance heterogeneity. Reliable estimates of genotype effects require proper accommodating of these sources of variation. This thesis demonstrates the handling of spatial variation and the use of weighting methods to account for within-trial and between-trial error variance heterogeneity in MET. Beside this the impact of weighting methods when applied to GS and GWAS is also evaluated.

5.1 Spatial modeling is an add-on

Classically, data from field experiments have been analyzed based on the randomization design of the experiments. In addition, different spatial methods have been proposed for adjusting means for any spatial trend that may exist (Gilmour et al., 1997; Piepho et al., 2008; Piepho and Williams 2010). All of these spatial methods are based on the assumption of correlated neighboring plots. While spatial models might be useful to increase precision and efficiency compared to the baseline model, there are cases for which the baseline model outperforms the spatial models fitted for a given trial. This shows that spatial modelling is not necessarily a substitute for a randomization-based model. But rather it is an add-on to the randomization-based or baseline model. This idea coincides with results in Chapter 2, where in one out of the three empirical examples the baseline model without the spatial covariance structure outperformed the spatial model. Appropriate experimental design is always mandatory for obtaining reliable results. The best approach of spatial modeling is to begin with the randomization-based baseline model and then add spatial components and compare these extended models with the baseline model to check if the fit is improved or not (Williams, 1986; Williams et al., 2006; Piepho and Williams, 2010; Müller et al., 2010). Borges et al. (2018) have evaluated the performance of design based and spatial modeling to answer the question whether spatial modeling can substitute design-based models or not. They make the comparison with different sizes and types of experimental when performed with different degree of variability between experiment sites. Their result show that for small-sized experiments the design based and spatial models have similar performance even when the degree of variability differs. When the number of experiments and variability among sites was high, using spatial models with completely randomized design (CRD) and randomized complete block design (CRBD) did not outperform the more efficient design they considered, i.e., alpha-lattice design. However, spatial models performed best when combined with the best experimental design (alpha-lattice design). Therefore in general spatial modeling is no substitute for experimental design but should be added at the analysis stage to control any spatial correlation that may exist in the field.

The use of spatial information in the design phase is another interesting application of spatial techniques in agriculture. Several studies have been conducted on this topic (Williams et al., 2006; Piepho and Williams, 2010; Williams and Piepho, 2013; Piepho et al., 2016b). Spatial design has a limitation, however, due to the presence of error variance bias (Williams and Piepho, 2018).

5.2 Spatial and variance modeling

In the analysis of individual trials from MET, there may exist within-trial variance heterogeneity. Accounting this heterogeneity is a plus to increase efficiency of estimation and it can also improve selection response (Edwards and Orellana, 2015). Appropriate variance modelling and data transformation are two possible choices when the assumption of within-trial homogeneneity of variance is violated. In Chapter 2, it was shown that Box-Cox transformation is suitable to stabilize variance but has the main disadvantage that results are difficult to report on the original scale. This is a major issue when joint analysis of MET is required, since means and variance-covariance estimates are required on the original scale. For a single trial often the inverse of the Box-Cox transformation is computed and reported as an estimate of the median on the original scale (Piepho, 2009a). An alternative way to report results based on the Box-Cox transformation on the original scale is using an expression derived by Freeman and Modarres (2006), provided that the transformation parameter is in between the interval 0 and 1. However, this expression is not applicable for this study because the transformtion parameter estimates are not in this interval. In contrast, variance modeling is a good alternative option for correcting the variance heterogeneity and to proceed with a joint analysis of MET, because it does not require a data transformation.

In Chapter 2, an application of variance modelling is illustrated. POM and exponential variance model are the particular variance models we considered (Carroll and Ruppert, 1988). The three examples illustrated that the variance modeling remedy was successful for dealing with the variance heterogeneity problem, meaning that the use of POM or exponential variance model resulted in better model fits than models without using the variance model.

Spatial variation and error variance heterogeneity are common in field trials. In Chapter 2, the performance of spatial models is compared with design-based models when ignoring the variance heterogeneity. The result shows spatial model performance to be better for the design-based model for all example dataset in this study. Moreover, spatial models along with POM and exponential variance models are compared with the same spatial model assuming homogeneous variance. The result indicates that a spatial model with both POM and exponential variance model performs better than the spatial homogeneous variance model. This study shows that both spatial model and variance heterogeneity can be accounted for simultaneously. Which variance model needs to be chosen depends on the specific data. Therefore evaluating different spatial and different variance models and then choosing the best fited model using model selection criteria will be the proper approach.

While variance modeling is advantageous applying proper weighting, it has limitations if the number of observations is small. This is because variance estimation based on a small number of degrees of freedom is typically unstable. In such cases estimation of variance induces extra variability (Carroll and Ruppert, 1988). Edwards and Jannink (2006) proposed to use the Bayesian approach as a solution for modeling heterogeneous error and genotype-environment interactions (GEI) when the number of observations is small. This type of approach to modelling heterogeneity is worth exploring in future work.

5.3 Stage-wise analysis for MET, GS and GWAS

Stage-wise analysis is often the approach for practical analysis of large scale MET data, GS and GWAS. Most stage-wise analyses do not fully reproduce results of single-stage analysis. The reason is that the variance-covariance matrix of adjusted means estimated in the previous stage is not fully forwarded to the next stage and as a result information is lost. In MET analysis the error variance between trials are usually heterogeneous, which requires remedial measures. Moreover, each trial may require different randomization and modeling approaches. Failure to properly accommodate all these sources of variability may induce unequal results of single-stage and stage-wise analysis.

Different types of weighting strategies are recommended by different authors to account for heterogeneity and thereby to increase efficiency (Smith et al., 2001; Möhring and Piepho, 2009; Piepho et al., 2012a). Among them the most efficient one is the fully efficient weighting method, which forwards the full variance-covariance matrix from the first stage to the second stage. The diagonal weighting approach which proposed by Smith et al. (2001) is the most popular one. In this study diagonal and fully-efficient weights have been used. Results from the pure phenotypic analysis (Chapter 3) showed that the difference between results from the two different weighting methods are very small. The choice of weighting method depends on the extent and on the complexity of the within-trial and between-trial variability. Likewise, the results of GS and GWAS analyses indicates that the choice of proper weighting method depends on the dataset and on the objective of the study. For example, in Chapter 4, the fully efficient method performed best for GS, while for GWAS diagonal weighting was the best approach. Kaio et al. (2019) concluded that weighted genomic prediction outperformed unweighted analysis. Generally, stage-wise analysis reproduces the same results of single-stage analysis if fully efficient weighting is used and the non-genetic variances are replaced by their REML estimates from the first stage.

In the analysis of field experiment data, genotype effects can be fitted as a random effect or as a fixed effect. The choice of fixed or random genotype effect usually depends on the objective of the study. Genotype should be fitted as random when the objective of the research is selection; however, genotype can be fitted as fixed when the objective is comparison (Smith et al., 2005). Genotype is fitted as random in GS and GWAS (Chapter 4). While the researcher should decide which type of effect to fit for genotype, in stage-wise analysis genotypes should be fitted as fixed effect in all stages except the last where genotype can be fitted as random or fixed depending on the objective (Piepho et al., 2012a).

Stage-wise analysis is suitable for practical analysis of MET because it is relatively simple to analyse individual trials accounting for all sources of variation, e.g., spatial correlation and error variance heterogeneity (Chapter 3, Example 4). Computational efficiency is another main reason for using stage-wise analysis. For example in this study it has been shown using an example that the time taken for single-stage analysis was about 23 hours and for two-stage and three-stage analysis it was only 2 and 3 minutes, respectively (Chapter 3). In so far as convergence is attained, in stage-wise analysis the use of fully efficient weighting is always suggested to minimize the loss of information which may occur when forwarding results from a previous to the next stage analysis (Piepho et al., 2012; Schulz-Streek et al., 2013). However, if the number of genotypes is large and a correlated covariance structure is used for the genetic effects, this may require large computation time. In such cases it is often advantageous to use a weighting method which is feasible with regard computation time. A study by Gogel et al. (2018), however, proves the possibility of single-stage analysis of MET, where there example dataset consists of more than 100 trials using ASREML-R (Butler et al. 2017). From our exprience for users of other statistical packages stage-wise analysis is a viable alternative (Möhring and Piepho, 2009; Piepho et al., 2012a; Morris et al., 2018).

Except for the single-stage analysis of the phenotypic data of Chapter 3, where we used ASREML-R package, all of the stage-wise analyses which involve weighting were computed in SAS (Chapter 2 and 4). SAS macros that can help to get the fully-efficient and diagonal weights were provided in this study (Chapter 2). In addition to these two weighting methods, the macro can be extended for other types of weights that can be used in stage-wise analysis of MET data, e.g. for the weights proposed for Möhring and Piepho (2009). Most statistical software packages for GS and GWAS analysis, e.g. TASSEL and R, do not have an option for applying of weights. Based on the results presented in this thesis, it is suggested that weighting methods also be implemented for GS and GWAS approaches.

5.4 Modeling of additive and non-additive effects and genotype by environment interaction

Partitioning of total genetic effects into additive and non-additive effects is often considered to be superior in performance to only fitting additive effects (Oakey et al., 2006, 2007). This partitioning can minimize prediction error and maximize accuracy of selection of genotypes in MET, enables accurate identification of significant marker in GWAS and increase prediction accuracy in GS (Jiang et al., 2015; Oakey et al., 2016; Bonnafous et al., 2018). In this thesis, however, this partitioning was not considered, assuming that it is not crucial for judging the relative merit of alternative weighting methods.

Proper accounting of the GEI variance and covariance heterogeneity is often of interest to researchers. The factor analytic (FA) variance structure is known for its ability to account for possible heterogeneity (Gogel et al., 1995; Piepho, 1997; Smith et al., 2015; Smith et al., 2018). In Chapter 4, we were tried the FA1 structure, but we couldn't proceed because first of all there were convergence problems and secondly estimated genotype means were not reliable when compared to the genotype mean estimates obtained with the simple variance components model.

5.5 MET analysis versus meta-analysis

In clinical trials in order to precisely estimate treatment effects, similar trials are conducted in two or more sites. Such studies are analyzed using a statistical technique known as meta-analysis. Two-stage meta-analysis of clinical trials and MET data from field trials are similar in sprit (Whitehead, 2002, Piepho et al., 2012a; Piepho et al., 2012c; Vargas et al., 2013; Madden et al., 2016). Mathew and Nordström (2010) explore an approach under which single-stage and two-stage meta-analysis gives identical results when optimal weighting is used at the second stage. Their scheme is similar to the two-stage analysis of MET with fully efficient weighting. However, the main difference between these methods is that in two-stage meta-analysis the first-stage analysis computes treatment differences and their corresponding standard errors separately for each trial. By contrast, MET analysis computes treatment means and associated standard errors in the first stage. In the second stage, joint analysis is done using inverse squared standard errors as weights. In addition in meta-analysis two-stage analysis is fully equivalent to single-stage analysis if the site (study) main effect is taken as fixed rather than random so no inter-site (study) information is recovered (Piepho et al., 2012b). Morris et al. (2018) compared one-stage versus two-stage meta-analysis. According to their results meta-analysis should use the same model effects assumption for both single-stage and two-stage analysis, otherwise meta-analysists are free to use whichever procedure.

Chapter 6

Conclusion

The objective of this thesis has been to develop methods accounting for within trial error variance heterogeneity and also to evaluate the performance of fully efficient weighting to control between-trial error variance heterogeneity. As the three examples presented in Chapter 2 showed, the Box-Cox transformation is a potential approach to stabilize variance, but it has a drawback because of the difficulty to report genotype means and their standard error in the original scale particularly for joint analysis of MET. Another appealing alternative to the Box-Cox transformation is to include variance heterogeneity in the model. Moreover, in field trials it is common to find correlated error between neighboring plots; this is a contradiction to the independent error assumption. To correct for this assumption failure, spatial modeling can be performed. As illustrated in this thesis, variance and spatial modeling can be implemented simultaneously for data from field trials.

Due to complex data structure MET data is usually analyzed using stage-wise analysis. To accommodate the variance heterogeneity between trials, weighting methods are usually applied in the joint analysis. For the datasets used in this study, it is shown that results from stage-wise analysis (with weight and without weight) agrees reasonably well with single-stage analysis. However, gain in efficiency of stage-wise analysis can be increased by using a fully-efficient weighting approach.

The evaluation of weighting methods in GS and GWAS analysis stage indicates no significant difference. For our dataset the difference between weighted and unweighted analysis of GS and GWAS were relatively small. However, we recommened the use of fully efficient weighting to maximize efficiency.

Chapter 7

Summary

In plant breeding programmes MET form the backbone for phenotypic selection, GS and GWAS. Efficient analysis of MET is fundamental to get accurate results from phenotypic selection, GS and GWAS. On the other hand inefficient analysis of MET data may have consequences such as biased ranking of genotype means in phenotypic data analysis, small accuracy of GS and wrong identification of QTL in GWAS analysis. A combined analysis of MET is performed using either single-stage or stage-wise (two-stage) approaches based on the linear mixed model framework. While single-stage analysis is a fully efficient approach, MET data is suitably analyzed using stage-wise methods. MET data often show within-trial and between-trial variance heterogeneities, which is in contradiction with the homogeneity of variance assumption of linear models, and these heterogeneities require corrections. In addition it is well documented that spatial correlations are inherent to most field trials. Appropriate remedial techniques for variance heterogeneities and proper accounting of spatial correlation are useful to improve accuracy and efficiency of MET analysis.

Chapter 2 studies methods for simultaneous handling of within-trial variance heterogeneity and within-trial spatial correlation. This study is conducted based on three maize trials from Ethiopia. To stabilize variance Box-Cox transformation was considered. The result shows that, while the Box-Cox transformation was suitable for stabilizing the variance, it is difficult to report results on the original scale. As alternative variance models, i.e. power-of-the-mean (POM) and exponential models, were used to fix the variance heterogeneity problem. Unlike the Box-Cox method, the variance models considered in this study were successful to deal simultaneously with both spatial correlation and heterogeneity of variance.

For analysis of MET data, two-stage analysis is often favored in practice over single-stage analysis because of its suitability in terms of computation time, and its ability to easily account for any specifics of each trial (variance heterogeneity, spatial correlation, etc). Stage-wise analyses are approximate in that they cannot fully reproduce a single-stage analysis because the variance–covariance matrix of adjusted means from the first-stage analysis is sometimes ignored or sometimes approximated and the approximation may not be efficient. Discrepancy of results between single-stage and two-stage analysis increases when the variance between trials is heterogeneous. In stage-wise analysis one of the major challenges is how to account for heterogeneous variance between trials at the second stage. To account for heterogeneous variance between trials, a weighted mixed model approach is used for the second-stage analysis. The weights are derived from the variances and covariances of adjusted means from the first-stage analysis. In Chapter 3 we compared single-stage analysis and two-stage analysis. A new fully efficient and a diagonal weighting matrix are used for weighting in the second stage. The methods are explored using two different types of maize datasets. The result indicates that single-stage analysis and two-stage analysis give nearly identical results provided that the full information on all effect estimates and their associated estimated variances and covariances is carried forward from the first to the second stage.

GWAS and GS analysis can be conducted using a single-stage or a stage-wise approach. The computational demand for GWAS and GS increases compared to purely phenotypic analysis because of the addition of marker data. Usually researchers compute genotype means from phenotypic MET data in stage-wise analysis (with or without weighting) and then forward these means to GWAS or GS analysis, often without any weighting. In Chapter 4 weighted stage-wise analysis versus unweighted stage-wise analysis are compared for GWAS and GS using phenotypic and genotypic maize data. Fully-efficient and a diagonal weighting are used. Results show that weighting is preferred over unweighted analysis for both GS and GWAS.

In conclusion, stage-wise analysis is a suitable approach for practical analysis of MET, GS and GWAS analysis. Single-stage and two-stage analysis of MET yield very similar results. Stage-wise analysis can be nearly as efficient as single-stage analysis when using optimal weighting, i.e., fully-efficient weighting. Spatial variation and within-trial variance heterogeneity are common in MET data. This study illustrated that both can be resolved simultaneously using a weighting approach for the variance heterogeneity and spatial modeling for the spatial variation. Finally beside application of weighting in the analysis of phenotypic MET data, it is recommended to use weighting in the actual GS and GWAS analysis stage.

Chapter 8

Zusammenfassung

In Pflanzenzüchtungsprogrammen bilden Versuchsserien die Grundlage für die phänotypische Selektion, genomische Selektion (GS) und genomweite Assoziationsstudien (GWAS). Eine effiziente Analyse der Versuchsserien ist grundlegend, um genaue Ergebnisse der phänotypischen Auswahl von GS und GWAS zu erhalten. Andererseits kann eine ineffiziente Analyse von Versuchsserien-Daten zu einer verzerrten Bewertung von Genotyp-Mitteln bei der Analyse phänotypischer Daten, einer geringen Genauigkeit der GS und einer falschen Identifizierung von QTL in der GWAS-Analyse führen. Eine kombinierte Analyse der Versuchsserien wird auf der Grundlage von linearen gemischten Modellen entweder einstufig oder stufenweise (zweistufig) durchgeführt. Während die einstufige Analyse ein vollständig effizienter Ansatz ist, werden die Versuchsserien-Daten in geeigneter Weise mit stufenweisen Methoden analysiert. Versuchsserien-Daten zeigen häufig Varianzheterogenitäten innerhalb von und zwischen Versuchen, die der Annahme der Varianzhomogenität für linearer Modelle widersprechen und Korrekturen erfordern. Darüber hinaus ist gut dokumentiert, dass räumliche Korrelationen in den meisten Feldversuchen vorhanden sind. Geeignete Abhilfemethoden für Varianzheterogenitäten und eine korrekte Berücksichtigung der räumlichen Korrelation sind hilfreich, um die Genauigkeit und Effizienz der versuchsserien-Analyse zu verbessern.

In Kapitel 2 werden Methoden zum gleichzeitigen Umgang mit Varianzheterogenitat zwischen und räumlicher Korrelation innerhalb der Versuche untersucht. Diese Studie basiert auf drei Maisversuchen aus Äthiopien. Um die Varianz zu stabilisieren, wurde die Box-Cox-Transformation in Betracht gezogen. Das Ergebnis zeigt, dass, obwohl die Box-Cox-Transformation zur Stabilisierung der Varianz geeignet war, es schwierig ist, Ergebnisse auf der ursprünglichen Skala darzustellen. Als alternative Varianzmodelle wurden Power-of-the-mean (POM) und Exponentialmodelle verwendet, um das Varianzheterogenitätsproblem zu beheben. Im Gegensatz zur Box-Cox-Methode gelang es den in dieser Studie betrachteten Varianzmodellen, sowohl räumliche Korrelation als auch Heterogenität der Varianz gleichzeitig zu berücksichtigen.

Bei der Analyse von MET-Daten wird die zweistufige Analyse in der Praxis häufig gegenüber der einstufigen Analyse bevorzugt, da sie die Berechnungszeit kürzer ist und die Besonderheiten der einzelnen Versuche (Varianzheterogenität, räumliche Korrelation usw.) leicht berücksichtigt werden können. Stufenweise Analysen sind insofern approximierend, als sie eine einstufige Analyse nicht vollständig reproduzieren können, da die Varianz-Kovarianz-Matrix der angepassten Mittelwerte aus der ersten Analyse-Phase manchmal ignoriert oder manchmal approximiert wird und die Approximation möglicherweise nicht effizient ist. Die Diskrepanz der Ergebnisse zwischen einstufiger und zweistufiger Analyse nimmt zu, wenn die Varianzen zwischen den Studien heterogen sind. Bei der stufenweisen Analyse besteht eine der größten Herausforderungen darin, die heterogene Varianz zwischen den Versuchen auf der zweiten Stufe zu berücksichtigen. Um die heterogene Varianz zwischen den Studien zu berücksichtigen, wird für die Analyse der zweiten Stufe ein gewichteter gemischter Modellansatz verwendet. Die Gewichtungen werden aus den Varianzen und den Kovarianzen der angepassten Mittel aus der Analyse der ersten Stufe abgeleitet. In Kapitel 3 haben wir die einstufige Analyse und die zweistufige Analyse verglichen. In der zweiten Stufe wird eine neue voll effiziente und eine diagonale Gewichtungsmatrix für die Gewichtung verwendet. Die Studien werden anhand zweier verschiedener Arten von Mais-Datasätze untersucht. Das Ergebnisse zeigen, dass die einstufige Analyse und die zweistufige Analyse nahezu identische Ergebnisse liefern, vorausgesetzt, die vollständigen Informationen zu allen Effektschätzungen und den damit verbundenen geschätzten Varianzen und Kovarianzen werden von der ersten zur zweiten Stufe übertragen.

Die GWAS- und GS-Analyse kann nach einem einstufigen oder einem stufenweisen Ansatz durchgeführt werden. Der rechnerische Bedarf an GWAS und GS steigt im Vergleich zur rein phänotypischen Analyse aufgrund der Hinzufügung von Markerdaten. In der Regel berechnen Forscher Genotyp-Mittel aus phänotypischen Versuchsserien-Daten in stufenweisen Analysen (mit oder ohne Gewichtung) und leiten diese dann in die GWAS- oder GS-Analyse weiter, oft ohne Gewichtung. In Kapitel 4 wird die gewichtete stufenweise Analyse gegen die ungewichtete stufenweise Analyse für GWAS und GS anhand von phänotypischen und genotypischen Maisdaten verglichen. Es werden volleffiziente und diagonale Gewichtungen verwendet. Die Ergebnisse zeigen, dass die gewichtete gegenüber der nicht gewichteten Analyse sowohl für GS als auch für GWAS besser ist.

Zusammenfassend ist die stufenweise Analyse ein geeigneter Ansatz für die praktische Versuchsserien-, GS- und GWAS-Analyse. Einstufige und zweistufige Versuchsserien-Analysen führen zu sehr ähnlichen Ergebnissen. Eine stufenweise Analyse kann wie eine einstufige Analyse effizient sein, indem eine optimale Gewichtung verwendet wird, d. h. eine vollständig effiziente Gewichtung. In Versuchsserien-Daten sind räumliche Variation und Varianzheterogenität innerhalb der Versuche üblich. Diese Studie zeigte, dass beide gleichzeitig unter Verwendung eines Gewichtungsansatzes die Varianzheterogenität und räumliche Korrelation berücksichtigen können. Neben der Anwendung der Gewichtung bei der Analyse phänotypischer MET-Daten wird empfohlen, die Gewichtung in der eigentlichen GS- und GWAS-Analysestufe zu verwenden.

References

Akaike, H. 1974. A new look at the statistical model identification, IEEE Transactions on Autom. Control 19: 716-723.

Atkinson, A.C. 1985. Plots, transformations and regression: An introduction to graphical methods of diagnostics regression analysis. Oxford University Press, Oxford.

Atlin, G.N., R.J. Baker, K.B. McRae, and X. Lu. 2000. Selection response in subdivided target regions. Crop Sci. 40:7–13.

Auinger, H.J., M. Schönleben, C. Lehermeier, M. Schmidt, V. Korzun, H.H. Geiger, H.P. Piepho, A. Gordillo, P. Wilde, E. Bauer, C.C. Schön. 2016. Model training across multiple breeding cycles significantly improves genomic prediction accuracy in rye (Secale cereale L.). Theor. Appl. Genet. 129:2043–2053.

Bailey, R.A., and C.J. Brien. 2016. Randomization-based models for multitiered experiments: I. A chain of randomizations. Annals of Statistics 44:1131–1164.

Bernal-Vasquez, A.M., J. Möhring, M. Schmidt, M. Schönleben, C.C. Schön, and H.P. Piepho. 2014. The importance of phenotypic data analysis for genomic prediction-a case study comparing different spatial models in rye. BMC Genomics 15:646.

Bernal-Vasquez, A.M., A. Gordillo, M. Schmidt, and H.P. Piepho. 2017. Genomic prediction in early selection stages using multi-year data in a hybrid rye breeding program. BMC Genetics 18:51.

Bonnafous, F., G. Fievet, N. Blanchet, M.C. Boniface, S. Carrère, J. Gouzy, L. Legrand, G. Marage, E. Bret-Mestries, S. Munos, et al. 2018. Comparison of GWAS models to identify non-additive genetic control of flowering time in sunflower hybrids. Theor. Appl. Genet. 131:319–32.

Borges, A., A. González-Reymundez, O. Ernst, M. Cadenazzi, J. Terra, and L. Gutiérrez. 2019. Can spatial modeling substitute for experimental design in agricultural experiments? Crop Sci. 59:44–53.

Box, G.E.P., and D.R. Cox. 1964. An analysis of transformations. J. R. Stat. Soc. Series B. 26:211–252.

Bradbury, P.J., Z. Zhang, D.E. Kroon, T.M. Casstevens, Y. Ramdoss, and E.S. Buckler. 2007.

TASSEL: Software for association mapping of complex traits in diverse samples. Bioinformatics 23:2633–2635.

Burnham, K.P., and D.R. Anderson. 1998. Model selection and multimodel inference: a practical information-theoretic approach. Springer, New York.

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

Caliński, T., S. Czajka, Z. Kaczmarek, P. Krajewski, and W. Pilarczyk. 2005. Analyzing multi-environment variety trials using randomization-derived mixed models. Biometrics 61:448–455.

Carroll, R.J., and D. Ruppert. 1988. Transformation and weighting in regression. Chapman and Hall, New York.

Comstock, R.E., and R.H. Moll. 1963. Genotype-environment interaction. In Statistical genetics and plant breeding. National Academy of Sciences-National Research Council No. 982:164–194.

Cochran, W.G. 1937. Problems arising in the analysis of a series of similar experiments. Journal of the Royal Statistical Society 4:102–118.

Cressie, N. 1991. Statistics for spatial data. John Wiley & Sons, New York.

Crossa, J. 1990. Statistical Analyses of Multilocation Trials. Advances in Agronomy 44:55–85.

Crossa, J., G.L. Campos, P. Pérez, D. Gianola, J. Burgueno, J.L. Araus, D. Makumbi, R.P. Singh, S. Dreisigacker, and J. Yan. 2010. Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics 186:713–724.

Cullis, B.R., and A.C. Gleeson. 1991. Spatial analysis of field experiments: An extension to two dimensions. Biometrics 47:1449–1460.

Cullis, B.R., B.G. Gogel, A.P. Verbyla, and R. Thompson. 1998. Spatial analysis of multi-environment early generation trials. Biometrics 54:1–18.

Damesa, T. M., J. Möhring, M. Worku, and H.P. Piepho. 2017. One step at a time: Stage-wise analysis of a series of experiments. Agronomy Journal 109:845–857.

Damesa, T.M., J. Möhring, J. Forkman, and H.P. Piepho. 2018. Modeling spatially correlated and heteroscedastic errors in Ethiopian maize trials. Crop Sci. 58:1575–1586.

Edmondson, R.N. 2005. Past developments and future opportunities in the design and analysis of crop experiments. J. Agric. Sci. 143:27–33.

Edriss, V., Y. Gao, X. Zhang, M.B. Jumbo, D. Makumbi, M.S. Olsen, … J.L. Jannink. 2017. Genomic prediction in a large African maize population. Crop Sci. 57:2361–2371.

Edwards, J.W., and J.L. Jannink. 2006. Bayesian modeling of heterogeneous error and genotype by environment interaction variances. Crop Sci. 46: 820–833.

Edwards, J.W., and M. Orellana. 2015. Increasing selection response by Bayesian modeling of heterogeneous environmental variances. Crop Sci. 55:556–563.

Estaghvirou, S. B. O., J.O. Ogutu, T. Schulz-Streeck, C. Knaak, M. Ouzunova, A. Gordillo, and H.P Piepho. 2013. Evaluation of approaches for estimating the accuracy of genomic prediction in plant breeding. BMC genomics 14:860.

Erbe, M., E.C.G. Pimentel, A.R. Sharifi, and H. Simianer. 2010. Assessment of cross-validation strategies for genomic prediction in cattle. In Proceedings of the 9th World Congress on Genetics Applied to Livestock Production. Leipzig, Germany.

FAOSTAT, FAO, Statistics, Food and agricultural organization of the United Nations. 2016. http://www.fao.org/faostat/en/#data/QC/visualize.

Frensham, A.B., B.R. Cullis, and A.P. Verbyla. 1997. Genotype by environment variance heterogeneity in atwo-stage analysis. Biometrics 53:1373–1383.

Finlay, K.W., and G.N. Wilkinson. 1963. The analysis of adaptation in a plant-breeding programme. Australian Journal of Agricultural Research 14:742–754.

Fisher, R.A. 1935. The design of experiments. Oliver & Boyd, Oxford, UK.

Freeman, J., and R. Modarres. 2006. Inverse Box–Cox: The power-normal distribution. Stat Probabil. Lett 76:764–772.

Gauch, H.G. 1992. Statistical analysis of regional yield trials. Elsevier, Amsterdam.

Gilmour, A.R., B.R. Cullis, and A.P. Verbyla. 1997. Accounting for natural and extraneous variation in the analysis of field experiments. J. Agric. Biol. Environ. Stat. 2:269–293.

Gleeson, A.C., and B.R. Cullis. 1987. Residual maximum likelihood (REML) estimation of a neighbor model for field experiments. Biometrics 43:277–288.

Gogel, B.J., B.R. Cullis, and A.P. Verbyla. 1995. REML estimation of multiplicative effects in multi-environmentvariety trials. Biometrics 51:744–749.

Gogel, B.J., A.B. Smith, B.R. Cullis. 2018. Comparison of a one- and two-stage mixed model analysis of Australia's national variety trial Southern region wheat data. Euphytica, 214:1–21.

Gomez, K.A., and A.A. Gomez. 1984. Statistical procedures for agricultural research (2 ed.). John Wiley and Sons, NewYork.

Gowda, M., B. Das, D. Makumbi, R. Babu, K. Semagn, G. Mahuku, … B.M. Prasanna. 2015. Genome-wide association and genomic prediction of resistance to maize lethal necrosis disease in tropical maize germplasm. Theor. Appl. Genet. 128:1957–1968.

Hayes, B., and M. Goddard. 2010. Genome-wide association and genomic selection in animal breeding. Genome 53:876–883.

Henderson, C.R. 1985. Best linear unbiased prediction of nonadditive genetic merits in noninbred populations. Journal of Animal Science 60:111–117.

James, W., and C. Stein. 1960. Estimation with quadratic loss. In Proc. Fourth Berkeley Symp. Math. Statist. Probability 1:361–380. University of California Press, Berkeley.

Jiang, Y., and J.C. Reif. 2015. Modeling epistasis in genomic selection. Genetics 201:759–768.

Jørgensen, B. 1987. Exponential dispersion models (with discussion). J. R. Stat. Soc. Series B Stat. Methodol. 49:127–162.

Jolliffe, I.T. 2002. Principal component analysis. 2nd edition. Springer, NewYork.

Kempton, R.A.1984. The use of biplots in interpreting variety by environment interactions. The Journal of Agricultural Science 103:123–138.

Kleinknecht, K., J. Möhring, K.P. Singh, P.H. Zaidi, G.N. Atlin, and H.P. Piepho. 2013. Comparison of the performance of BLUE and BLUP for zoned Indian maize data. Crop Sci. 53:1384–1391.

Kozak, M., and H.P. Piepho. 2018. What's normal anyway? Residual plots are more telling than significance tests when checking ANOVA assumptions. Journal of Agronomy and Crop Science 204: 86–98.

Lee, Y., J.A. Nelder, and Y. Pawitan. 2006. Generalized linear models with random effects. Unified analysis via H-likelihood. Chapman & Hall/CRC, Boca Raton.

Lee, C.J., M. O’Donnell, and M. O’Neill. 2008. Statistical analysis of field trials with changing treatment variance. Agron. J. 100:484–489.

Littell, R.C., G.A. Milliken, W.W. Stroup, R.D. Wolfinger, and O. Schabenberger. 2006. SAS for mixed models. Second edition. Cary, NC: SAS Institute Inc.

Madden, L.V., H.P. Piepho, and P.A. Paul. 2016. Statistical models and methods for network meta-analysis. Phytopathology 106:792–806.

Mathew, T., and K. Nordström. 2010. Comparison of one-step and two-step meta-analysis models using individual patient data. Biometrical Journal 52:271–287.

McLean, R.A., W.L. Sanders, and W.W. Stroup. 1991. A unified approach to mixed linear models. The American Statistician 45:54–64.

Meuwissen, T. H. E., B.J. Hayes, and M.E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829.

Morris, T.P., D.J. Fisher, M.G. Kenward, J.R. Carpenter. 2018. Meta-analysis of gaussian individual patient data: two-stage or not two-stage? Statistics in Medicine 37:1419–1438.

Möhring, J., and H.P. Piepho. 2009. Comparison of weighting in two-stage analyses of plant breeding trials. Crop Sci. 49:1977-1988.

Müller, B.U., K. Kleinknecht, J. Möhring, and H.P. Piepho. 2010. Comparison of spatial models for sugar beet and barley trials. Crop Sci. 50:794–802.

Oakey, H., A.P. Verbyla, W. Pitchford, B.R. Cullis, and H. Kuchel. 2006. Joint modeling of additive and non-additive genetic line effects in single field trials. Theor. Appl. Genet. 113:809–819.

Oakey, H., A.P. Verbyla, B.R. Cullis, X. Wei, and W. Pitchford. 2007. Joint modelling of additive and non-additive (genetic line) effects in multi-environment trials. Theor. Appl. Genet. 114:1319–1332.

Oakey, H., B.R. Cullis, R. Thompson, J. Comadran, C. Halpin, and R. Waugh. 2016. Genomic Selection in multi- environment crop trials. G3: Genes, Genomes, Genetics 6:1313–1326.

Ogutu, J.O., H.P. Piepho and T. Schulz-Streeck. 2011. A comparison of random forests, boosting and support vector machines for genomic selection. In BMC Proceedings 5(Suppl 3):S11.

Ogutu, J.O., T. Schulz-Streeck, and H.P. Piepho. 2012. Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions. BMC Proceedings 6(Suppl 2):S10.

Oraguzie, N.C., E.H.A. Rikkerink, S.E. Gardiner and H.N. de Silva. 2007. Association mapping in plants. Springer, New York.

Palaiokostas, C., T. Vesely, M. Kocour, M. Prchal, D. Pokorová, V. Piackova, L. Pojezdal, and R.D. Houston. 2019. Optimizing genomic prediction of host resistance to koi herpesvirus disease in carp. Frontiers in genetics 10:543. doi:10.3389/fgene.2019.00543.

Patterson, H.D., and R. Thompson. 1971. Recovery of interblock information when block sizes are unequal. Biometrika 31:100–109.

Peel, D., M.V. Bravington, N. Kelly, S.N. Wood, and I. Knuckey. 2012. A model-based approach to designing a fishery-independent survey. J. Agric. Biol. Environ. Stat. 18:1–21.

Piepho, H.P. 1997. Analyzing genotype-environment databy mixed models with multiplicative terms. Biometrics 53:761–767.

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

Piepho, H.P. 2005. Statistical tests for QTL and QTL-by-environment effects in segregating populations derived from line crosses. Theor. Appl. Genet. 110:561–566.

Piepho, H.P. 2009a. Data Transformation in statistical analysis of field trials with changing treatment variance. Agron. J. 101:865–869.

Piepho, H. P. 2009b. Ridge regression and extensions for genomewide selection in maize. Crop Sci. 49:1165–1176.

Piepho, H.P., C. Richter, and E.R. Williams. 2008. Nearest neighbor adjustment and linear variance models in plant breeding trials. Biometrics 50:164–189.

Piepho, H.P., and J. Möhring. 2005. Best linear unbiased prediction for subdivided target regions. Crop Sci. 45:1151–1159.

Piepho, H.P., and E.R. Williams. 2010. Linear variance models for plant breeding trials. Plant Breed. 129:1–8.

Piepho, H.P., J. Möhring, T. Schulz-Streeck, and J.O. Ogutu. 2012a. A stage-wise approach for the analysis of multi-environment trials. Biometrical Journal 54:844–860.

Piepho, H.P., J.O. Ogutu, T. Schulz-Streeck, B. Estaghvirou, A. Gordillo, and F. Technow. 2012b. Efficient computation of ridge-regression BLUP in genomic selection in plant breeding. Crop Sci. 52:1093–1104.

Piepho, H.P., E.R. Williams, and L.V. Madden. 2012c. The use of two-way mixed models in multi-treatment meta-analysis. Biometrics 68:1269–1277.

Piepho, H.P., J. Möhring, and E.R. Williams. 2013. Why randomize agricultural experiments? Journal of Agronomy and Crop Science 199:374–383.

Piepho, H.P., and T. Eckl. 2014. Analysis of series of variety trials with perennial grasses. Grass Forage Science 69:431–440.

Piepho, H.P., M.F. Nazir, M. Qamar, A.ur.R. Rattu, Riaz-du-Din, M. Hussain, G. Ahmad, Fazal-e-Subhan, J. Ahmed, Abdullah, K.B. Laghari, I.A. Vistro, M.S. Kakar, M.A. Sial, and M. Imtiaz. 2016a. Stability analysis for a country-wide series of wheat trials in Pakistan. Crop Sci. 56:2465–2475.

Piepho, H.P., E.R. Williams, and V. Michel. 2016b. Nonresolvable Row–Column Designs with an Even Distribution of Treatment Replications. Journal of Agricultural, Biological, and Environmental Statistics 21:227–242.

Pritchard, J.K., M. Stephens, and P. Donnelly. 2000. Inference of population structure using multilocus genotype data. Genetics 155:945–959.

Ranum, P., J.P. Peña-Rosas, and M.N. Garcia-Casal. 2014. Global maize production, utilization, and consumption. Annals of the New York Academy of Sciences.

Rice, B., and A.E. Lipka. 2019. Evaluation of RR-BLUP Genomic Selection Models that Incorporate Peak Genome-Wide Association Study Signals in Maize and Sorghum. The Plant Genome, 12:1. https://doi.org/10.3835/plantgenome2018.07.0052

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

Sakia, R.M. 1992. The Box-Cox transformation technique: a review. The Statistician 41:169–178.

Schabenberger, O., and F.J. Pierce. 2002. Contemporary statistical models for the plant and soil sciences, CRC Press, Boca Raton, FL, USA.

Schabenberger, O., and C.A. Gotway. 2005. Statistical methods for spatial data analysis. Chapman & Hall/ CRC, press.

Schulz-Streeck, T., J.O. Ogutu, and H.P. Piepho. 2013a. Comparisons of single-stage and two-stage approaches to genomic selection. Theor. Appl. Genet. 126:69–82.

Schulz-Streeck T., J.O. Ogutu, A. Gordillo, Z. Karaman, C. Knaak, H.P. Piepho. 2013b. Genomic selection allowing for marker-by-environment interaction. Plant Breed. 132:532–538.

Searle, S.R., G. Casella, and C.E. McCulloch. 1992. Variance components. Wiley, New York.

Shikha, M., A. Kanika, A.R. Rao, M.g. Mallikarjina, H.S. Gupta, and T. Nepolean. 2017. Genomic selection for drought tolerance using genome-wide SNPs in maize. Frontiers in Plant Science 8:550.

Smith, A., B.R. Cullis, and A. Gilmour. 2001. The analysis of crop variety evaluation data in Australia. Australian and New Zealand Journal of Statistics 43:129–145.

Smith, A.B., B.R. Cullis, and R. Thompson. 2005. The analysis of crop cultivar breeding and evaluation trials. An overview of current mixed model approaches. J. Agric. Sci. (Cambridge) 143:449–462.

Smith, A.B., A. Ganesalingam, H. Kuchel, and B.R. Cullis. 2015. Factoranalytic mixed models for the provision of grower information from national crop variety testing programmes. Theor. Appl. Genet 128:55–72.

Smith, A.B., and B.R. Cullis. 2018. Plant breeding selection tools built on factor analytic mixed models for multi-environment trial data. Euphytica 214:143.

So, Y.S., and J. Edwards. 2011. Predictive ability assessment of linear mixed models in multi-environment trials in corn. Crop Sci. 51:542–552.

Song, J., B.F. Carver, C. Powers, L. Yan, J. Klapste, Y.A. El-Kassaby, C. Chen. 2017. Practical application of genomic selection in a doubled-haploid winter wheat breeding program. Mol. Breed. 37:117.

Sripathi, R., P. Conaghan, D. Grogan, and M.D. Casler. 2017. Spatial variability effects on precision and power of forage yield estimation. Crop Sci. 57:1–11.

Stich, B., J. Möhring, H.P. Piepho, M. Heckenberger, E.S. Buckler, and A.E. Melchinger. 2008. Comparison of mixed-model approaches for association mapping. Genetics 178:1745–1754.

Stroup, W.W. 2015. Rethinking the analysis of non-normal data in plant and soil science. Agronomy Journal 107:810–827.

Stefanova, K.T., A.B. Smith, and B.R. Cullis. 2009. Enhanced diagnostics for the spatial analysis of field trials. Journal of agricultural, biological, and environmental statistics 14:392–410.

Sverrisdóttir, E., Sundmark, E., Johnsen, H. Ø., Kirk, H. G., Asp, T., Janss, L., … Nielsen, K. L. (2018). The Value of Expanding the Training Population to Improve Genomic Selection Models in Tetraploid Potato. Frontiers in plant science, 9, 1118. doi:10.3389/fpls.2018.01118

Talbot, M. 1997. Resource allocation for selection systems. In R.A. Kempton and P.N. Fox (Eds.). Statistical methods for plant variety evaluation. Chapman and Hall, London, 162–174.

Tweedie, M.C.K. 1947. Functions of a statistical variate with given means, with special reference to Laplacian distributions, Math. Proc. Cambridge 43:41–49.

Tweedie, M.C.K. 1984. An index which distinguishes between some important exponential families. *In* Statistics: Applications and New Directions. Proceedings of the Indian Statistical Institute Golden Jubilee International Conference, Indian Statistical Institute, Calcutta, pp. 579–604.

van Eeuwijk, F.A., J. Crossa, M. Vargas, and J.M. Ribaut. 2002. Analysing QTL by environment interaction by factorial regression, with an application to the CIMMYT drought and low nitrogen stress programme in maize. In: Kang, M.S. (Ed.) 'Quantitative Genetics, Genomics and Plant Breeding'. pp. 245–256. CAB International, Wallingford, UK.

Vargas, M., E. Combs, G. Alvarado, G. Atlin, K. Mathews, and J. Crossa. 2012. META: A suite of SAS programs to analyze multi-environment breeding trials. Agronomy Journal 105:11-19.

Wangai, A.W., M.G. Redinbaugh, Z.M. Kinyua, D.W. Miano, P.K. Leley, M. Kasina, G. Mahuku, K. Scheets, and D. Jeffers. 2012. First report of maize chlorotic mottle virus and maize lethal necrosis in Kenya. Plant Dis 96:1582.

Welham, S., B.J. Gogel, A.B. Smith, R. Thompson and B.R. Cullis. 2010. A comparison of analysis methods for late-stage variety evaluation triaLS. Australian and New Zealand Journal of Statistics 52: 125–149.

Whitehead, A. 2002. Meta-analysis of controlled clinical trials. Wiley, New York.

Williams, E.R. 1986. A neighbour model for field experiments. Biometrika 73:279–287.

Williams, E.R., J.A. John, and D. Whitaker. 2006. Construction of resolvable spatial row-column designs. Biometrics 62:103–108.

Williams, E.R., and H.P. Piepho. 2018. An evaluation of error variance bias in spatial designs. Journal of Agricultural, Biological, and Environmental Statistics 23:83–91.

Wimmer, V., T., Albrecht, H.J. Auinger, and C.C. Schön. 2012. synbreed: a framework for the analysis of genomic prediction data using R. Bioinformatics 28:2086–2087.

Wood, S.N., and M. Fasiolo. 2017. A generalized Fellner-Schall method for smoothing parameter optimization with application to Tweedie location, scale and shape models. Biometrics 73:1071–1081.

Worku, M., L. Wolde, B. Tadesse, D. Girma, A. Girum, W. Abera, et al. 2012. Status and future direction of maize research and production in Ethiopia. *In: Worku, M., S. Twumasi-Afriyie, L. Wolde, B. Tadesse, G. Demisie, G. Bogale, et al. (Eds.).* Meeting the challenges of global climate change and food security through innovative maize research. *Proceedings of the 3rd national maize workshop of Ethiopia April 17-20,* *2011; Addis Ababa, Ethiopia, 17-23.* The International Maize and Wheat Improvement Center.

Vargas, M., E. Combs, G. Alvarado, G. Atlin, K. Mathews, and J. Crossa. 2012. META: A suite of SAS programs to analyze multi-environment breeding trials. Agronomy Journal 105:11–19.

Yates, F., and W.G. Cochran. 1938. The analysis of groups of experiments. Journal of Agricultural Science 28:556–580.

Yu, J., G. Pressoir, W.H. Briggs, I.V. Bi, M. Yamasaki, J.F. Doebley, … E.S. Buckler. 2006. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nature Genetics 38:203-208.

Zhao, K., M.J. Aranzana, S. Kim, C. Lister, C. Shindo, C. Tang, …, M. Nordborg. 2007. An Arabidopsis example of association mapping in structured samples. PLoS Genetics 3:e4.

Zimmerman, D.L., and D.A. Harville. 1991. A random field approach to the analysis of field-plot experiments and other spatial experiments. Biometrics 47:223–239.