FAKULTÄT AGRARWISSENSCHAFTEN

Aus dem Institut für Kulturpflanzenwissenschaften

Universität Hohenheim

Fachgebiet: Biostatistik

Prof. Dr. H.-P. Piepho

Model selection by cross-validation in multi-environment trials

Dissertation
zur Erlangung des Grades eines Doktors
der Agrarwissenschaften

vorgelegt

der Fakultät Agrarwissenschaften

von

Steffen Hadasch

aus Heilbronn

2017

Die vorliegende Arbeit wurde am 13.12.2017 von der Fakultät Agrarwissenschaften der Universität Hohenheim als "Dissertation zur Erlangung des Grades eines Doktors der Agrarwissenschaften" angenommen

Tag der mündlichen Prüfung: 13.06.2018

Leiter der Prüfung:	Prof. Dr. Thilo Streck
Berichterstatter, 1. Prüfer:	Prof. Dr. H.-P. Piepho
Mitberichterstatter, 2. Prüfer:	Dr. Johannes Forkman
3. Prüfer:	Dr. Tobias Würschum

Table of contents

1. General Introduction

The aim of plant breeding is to breed varieties that are superior to other varieties with regard to, typically quantifiable, phenotypic traits like yield, quality traits, disease resistance, etc. In cropping it is well known that the trait of a genotype may depend on the environment in which it is grown. Such genotype-environment interaction allows breeding varieties that are well adapted to a certain target population of environments (TPE) which represents a geographic region that requires different aspects of specific adaptation (Cooper et al., 1997; Cooper, 1999). To quantify the traits, breeders typically conduct multi-environment trials (MET) in which a set of genotypes is tested in a set of environments which can be seen as a sample from the TPE. Based on the data of such trials, the breeder may aim to select genotypes that show desirable trait values either across all tested environments or in a certain set of the tested environments. Statistical models are frequently used to analyse observed traits (phenotypic data) from such experiments (Lynch and Walsh, 1998). These models use the experimental data to estimate the trait means of the tested genotype (-environment combinations) and a measure of uncertainty associated with the estimated trait means. In practice the estimated trait means are used by farmers as a guide to what they can expect 'on average' from growing this genotype. For breeders the use of statistical models further allows determining genetic markers that are associated with the trait and therefore plays an important role in order to generate crosses with desirable traits. Furthermore, statistical models allow evaluating the stability of the genotypes across the tested environments, to group genotypes that show similar trait values, or form groups of similar environments (Gauch, 1992; Yan et al., 2007).

1.1. Statistical models in plant breeding

Statistical models aim to describe observed data by a mathematical function of several variables. The success in the application of statistical models depends on how well the model approximates the data generating process which is mostly unknown in practice. Knowledge about the true data generating process, however, usually is of minor importance as the scope in the application of statistical models is to describe the observed data (James et al., 2013). Data from MET are typically described by linear mixed models as such models allow to model correlations between observations, representing the fact that related genotypes show similar trait values. Within the environments of a MET the experiments are typically laid out according

to an experimental design in order to control for the heterogeneity in growing conditions within environments. For a single observation of a MET laid out as a resolvable incomplete block design (rIBD) a linear mixed model can be written as

$$y_{ijkl} = \mu + g_i + e_j + w_{ij} + r_{jk} + b_{jkl} + \varepsilon_{ijkl}$$

where μ is the intercept, g_i and e_j are the main effects of the i-th genotype and the j-th environment, w_{ij} is the interaction of the i-th genotype and the j-th environment, r_{jk} is the effect of the k-th replicate in the j-th environment, b_{jkl} is the effect of the l-th incomplete block within the k-th replicate in the j-th environment, ε_{ijkl} is the error, and y_{ijkl} is the observed trait value. The effects of b_{jkl}, r_{jk}, and ε_{ijkl} are assumed to be random effects, whereas g_i, e_j, and w_{ij} can be modelled as fixed or random effect, depending on the objective of the breeder. With data from plant breeding trials, the random effects are often assumed to be normally distributed. With such a model, the selection of genotypes with desirable trait values across all environments can be done using the genotype means $\mu + g_i$ assuming that e_j and w_{ij} are random effects with zero mean, while selection of genotypes with a high environment-specific mean is done by the genotype-environment means $\mu + g_i + e_j + w_{ij}$.

When marker information of the genotypes is available, the genotype effects can be modelled by a sum of marker effects, $g_i = \sum_{m=1}^{M} b_m x_{im}$, where x_{im} represents a dummy coded version of the allelic information of the i-th genotype at the m-th marker, and b_m is a regression slope associated with the m-th marker. When genetic marker information is used, the genotype effects can be regarded as a multiple regression of the genotype effects on the marker covariates. Similarly, the genotype-environment interaction effects can be replaced by a sum of environment specific marker effects $w_{ij} = \sum_{m=1}^{M} b_{mj} x_{im}$. Models that use marker information are used in association mapping (Stich et al., 2008), marker assisted selection (MAS; Heffner et al., 2009) and genomic prediction (GP; Meuwissen et al., 2001). In Chapter 2, the marker information is used for MAS and GP in order to predict the means of genotypes that were not used to estimate the model.

Instead of a regression of the interaction effects on marker covariates, a regression may also be done using a latent covariate. In the simplest case, the interaction effects may be described by a regression of the interaction on the environment effects, i.e. $w_{ij} = b_i e_j$, where b_i is a regression slope of the i-th genotype (Yates and Cochran, 1938) which often is denoted as Finlay-Wilkinson regression model (Finlay and Wilkinson, 1963). In this model, both b_i and

e_j can be seen as latent variables as they are not observable, unlike genetic marker information or environmental covariates. For an interpretation of the effects, either e_j or b_i could be interpreted as covariate while the other one represents a regression slope. Because of the multiplicative structure of the interaction, the effects need to be restricted such that the model is identifiable and furthermore, models containing multiplicative effects are non-linear in the parameters. In the context of regression on latent covariates, the additive main effects and multiplicative interaction (AMMI) model (Gauch, 1992) can be regarded as an extension of the Finlay-Wilkinson model. In the AMMI model, the interaction effects are described by a sum of multiplicative terms, i.e. $w_{ij} = \sum_{m=1}^{N} \lambda_m u_{im} v_{jm}$, where u_{im} and v_{jm} represent latent genotypic, and environmental variables, λ_m represents a scaling factor, and N represent the number of multiplicative terms which is limited by the dimension of genotypes and environments. In this model the parameters that constitute the multiplicative terms are subject to the constraints of the singular value decomposition (SVD). Similar to the AMMI model, the genotype and genotype-environment interaction (GGE) model describes the sum of genotype and interaction effects by multiplicative terms, i.e. $g_i + w_{ij} = \sum_{m=1}^{N} \lambda_m u_{im} v_{jm}$. The multiplicative terms of these models are typically used to visualize the interaction using biplots which allow to group genotypes and/or environments that perform similarly (Yan et al., 2000; Yan et al., 2007; Gauch, 2013). AMMI and GGE models are focused in Chapters 3 and 4. In Chapter 3, the determination of the optimal number of multiplicative terms is pursued, while Chapter 4 proposes algorithms for a weighted estimation of these models.

1.2. Stage-wise analysis

Data from of MET often are high-dimensional, and the covariance matrix of the data may not be block-diagonal and contains numerous parameters that need to be estimated, thus one may face some computational challenges when estimating a model. To circumvent computational burdens, a stage-wise analysis was proposed to estimate genotype or genotype-environment means from MET data (Smith et al., 2001a; Piepho et al., 2012). Usually, the breeder aims to estimate the genotype means, treating the genotype effects either fixed or random. The stage-wise analysis pursues this goal by first estimating the genotype-environment means for each environment individually, taking the experimental design within environments into account. In a subsequent step, the estimated genotype-environment means are used to estimate the genotype means while accounting for environment and interaction effects. This stage-wise

estimation therefore makes use of the hierarchical structure of the MET data in the sense that the effects of the lower level of the hierarchy are estimated and used to estimate effects at the higher levels of the hierarchy. In the second stage, the assumptions on the genotype, environment and interaction effects may vary, depending on the objective of the breeder. The computational benefit in the hierarchical processing mostly comes from the stage-wise estimation of the covariance structure of the data. In particular, when genotypes, environment and interactions are treated as random effects, the covariance structure of the data contains covariances induced by those effects, and furthermore it contains independent blocks which refer to the environments. In the first stage of a stage-wise analysis, the independent blocks are analysed separately. The second stage then estimates the remaining covariances that were ignored in the first stage. In the second stage of a stage-wise analysis, the covariance structure estimated in the first stage can be taken into account by including a weighting matrix in the estimation of the second stage. As weighting matrix one may use either the inverse of the covariance matrix of the estimated means obtained in the first stage (Damesa et al., 2017). Alternatively, a diagonal matrix containing the diagonal entries of the inverse covariance matrix on the diagonal can be used as weighting matrix (Smith et al., 2001a). Taking the (co)variances in the second stage into account is required to approximate the single-stage analysis as good as possible. Models that use marker information can be estimated in two stages (Schulz-Streeck et al., 2013). With these models, estimation is done using statistical software for linear mixed models where the covariance matrix obtained in the first stage can be taken into account easily in the second stage. AMMI and GGE models are usually also estimated in two stages where the first stage consists of estimating genotype-environment means from the replicates while the second stage consists of estimating the main effects and the multiplicative interaction. In case of AMMI/GGE models, weighted estimation of the second stage was proposed using diagonal weighting matrices, i.e. when the genotype-environment means are independent with potentially heterogeneous variances (Rodriguez et al., 2014). Chapter 4 pursues the estimation of second stage of the AMMI models when the data is correlated.

1.3. Cross validation

In the application of statistical models, there usually is a set of candidate models that can be used to describe a trait. The most appropriate model needs to be determined by comparing the performance of different models in terms of a certain objective of the breeder. Frequent objectives in the application of the models are the prediction of genotypes that were not tested

in the field or to identify the most appropriate (prediction) model for an observed dataset. To evaluate these objectives, cross-validation (CV) can be used (Hastie and Tibshirani, 2001; James et al., 2013). In CV the data is divided into so-called training data and validation data. The model parameters are estimated from the training data and the validation data is predicted using the parameter estimates obtained by the training data. The goodness of the prediction is evaluated by a success criterion. The predictive ability of a model, defined as the Pearson correlation coefficient between the validation data and their predictions, or the test mean squared error (test-MSE) which is the mean of the squared differences between the validation data and their predictions (Hastie and Tibshirani, 2001; James et al., 2013) are often used as a success criterion.

CV can be used to evaluate different objectives the breeder may pursue with MET data by an appropriate choice of training and validation data. With models that use genetic marker information, either genotypes, environments or both may be sampled for validation in order to evaluate the predictive ability of the models in terms of predicting (i) unobserved genotypes, (ii) observed genotypes in an unobserved environment, and (iii) unobserved genotypes in an unobserved environment can be evaluated (Utz, 2000; Chapter 2).

In case of AMMI/GGE models, CV can be used to determine the number of multiplicative terms. With these models CV can be done by sampling one replicate of each genotype-environment combination randomly (Gauch and Zobel, 1988) such that the validation data come from different replications of the experiment. Another sampling strategy uses validation data from one complete replication of the experiment within each environment (Piepho, 1994). In this way, the validation data within an environment come from the same replication and furthermore this sampling strategy can be shown to add uncertainty only to the estimated environment effects while the sampling strategy by Gauch and Zobel (1988) adds uncertainty to the estimated interaction effects which may lead to an underestimation of the number of multiplicative terms (Piepho, 1994). The sampling procedure proposed by Piepho (1994) therefore mostly aims to improve the evaluation procedure by considering consequences associated with the model estimation; the model evaluation may further be improved by using a success criterion that is based on pairwise differences of the validation data (Piepho, 1998). Another approach to account for the experimental design in a CV is proposed in Chapter 3. In this approach, the data are adjusted for the design (replicate and block) effects before the application a CV scheme. This approach aims at improving the evaluation by an adjustment of the validation data such that the success criterion is largely unaffected by design effects

contained in the validation data. Generally, evaluation of a given objective by CV requires an adequate sampling strategy which may also need to take issues related to the model estimation and the validation procedure into account.

1.4. Objectives of this study

In this work, the determination of the most appropriate prediction model by CV was the major issue for essentially two different tasks that occur in plant breeding: Chapter 2 deals with models that the use of genetic marker information, while Chapter 3 and 4 focus on AMMI/GGE models. In Chapter 2, models that use genetic marker information were applied to describe two differently inherited traits of lettuce. In particular, a stage-wise analysis was applied and the predictive abilities of different models, comprising phenotypic selection, marker-assisted selection, and genome-wide selection were comparatively evaluated. For this purpose, different CV schemes, which apply different sampling strategies, were implemented to identify the most appropriate prediction model for a certain objective.

In AMMI/GGE models, the most appropriate number of multiplicative terms that are used to describe the interaction can be determined using CV. Thus, different CV schemes that can be used to determine the number of multiplicative terms in AMMI/GGE models were compared in Chapter 3. In the CV schemes, the number of multiplicative terms that lead to the lowest test-MSE can be used as an estimate for the most appropriate number of multiplicative terms. Using simulated MET data in which the interaction effects were generated by a known number of multiplicative terms, the performances of different CV schemes in terms of recovering the number of multiplicative terms underlying the simulated data were compared. As data from MET are typically laid out according to a certain experimental design, the design effects contained in the data need to be taken into account when the test-MSE is computed. In Chapter 3, the design effects were taken into account by pre-processing the data before applying a CV scheme.

The AMMI/GGE model is usually estimated in two stages where the first stage consists of estimating the genotype-environment means from the replicated MET data, and the second stage consists of estimating the main effects of genotypes and environment, and the multiplicative interaction. With replicated MET data, the estimated genotype-environment means obtained in the first stage may be correlated. Thus, the covariance matrix of the estimated genotype-environment means needs to be taken into account when the model of the

second stage is estimated. Therefore, three different algorithms for a weighted estimation of the AMMI/GGE model are proposed in Chapter 4. The estimates obtained by the weighted algorithms were compared to an unweighted estimation in a simulation study to investigate the effectiveness of the weighted estimation.

2. Comparing predictive abilities of phenotypic and marker-assisted selection methods in a biparental lettuce population

S. Hadasch[a], I. Simko[b], R.J. Hayes[b], J. O. Ogutu[a], H.P. Piepho[a]

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

[b] U.S. Department of Agriculture, Agricultural Research Service, Crop Improvement and Protection Unit, 1636 E. Alisal Street, Salinas, CA 93905, USA.

Abstract

Many agronomic traits of plants are polygenic such that breeding for such traits is a challenging task. Selection for such traits can be done by phenotypic selection, marker-assisted selection or genome-wide selection. In this study different selection models comprising phenotypic selection, a marker-assisted selection, and genome-wide selection were compared in terms of their predictive abilities under cross-validation. The data used in this analysis represent two traits (downy mildew resistance and shelf-life) of a biparental lettuce population genotyped with 95 SNP and 205 AFLP markers. The predictive abilities of the models were obtained under three different cross validation procedure that sample either genotypes, environments, or both. For the downy mildew resistance data, the predictive ability of the genome-wide selection model was often found to be significantly higher than the predictive ability of the marker-assisted selection model under the different cross validation schemes. For the shelf-life data, the predictive ability of the marker-assisted selection model was often significantly higher than the predictive ability of the genome-wide selection model. The results of the study furthermore show that the predictive ability also depends on the cross validation scheme and the heritability of the target trait.

3. Cross validation in AMMI and GGE models: A comparison of methods

S. Hadasch[a], J. Forkman[b], H.P. Piepho[a*]

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

[b] Swedish University of Agricultural Sciences, Department of Crop Production Ecology, PO-Box 7043, 750 07 Uppsala, Sweden

Abstract

Genotype-environment interaction often is of major interest for plant breeders because it allows to identify and develop genotypes that are well adapted to a set of target environments. The data used to investigate the interaction stem from mutli-environemnt trials (MET) which are typically laid out according to an experimental design to control for heterogeneity in the field conditions are conducted. The analysis of such data can be done by the additive main effect and multiplicative interaction (AMMI) or the genotype-and genotype-environment interaction (GGE) model. In the application of these models, the objectives are to determine the number of multiplicative terms and to estimate the genotype-environment means as precisely as possible. Here, the performances of different cross validation schemes, of which some sample replicates of the genotype-environment combinations and some sample a genotype-environment mean for validation were evaluated in terms of the two objectives using data simulated based on the parameters from real experiments laid out as randomized complete block designs (RCBD) and as resolvable incomplete block designs (rIBD). Furthermore, an F-test to determine the number of multiplicative terms was applied in case of the RCBD. The results of the study show that the cross validation schemes that sample replicates of the genotype-environment combinations for validation most often outperformed the cross validation schemes that sample genotype-environment means. The results also show that the F-Test outperformed the cross validation schemes in case of the RCBD.

4. Weighted estimation of AMMI and GGE models

S. Hadasch[a], J. Forkman[b], W.A. Malik[a], H.P. Piepho[a*]

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

[b] Swedish University of Agricultural Sciences, Department of Crop Production Ecology, PO-Box 7043, 750 07 Uppsala, Sweden

The additive main effects and multiplicative interaction (AMMI) and the genotype-and genotype-environment interaction (GGE) model can be used to describe a two-way table of genotype environment means and to investigate the genotype-environment interaction. When the data of such two-way table are independent and homoscedastic, ordinary least squares (OLS) provides optimal estimates of the model parameters. In plant breeding, the assumption of independence and homoscedasticity is frequently violated because the means are estimated from the replicates within environments which are typically arranged according to an experimental design. Thus, when the genotype-environment means are correlated and heteroscedastic, generalized least squares (GLS) estimation is appropriate. In this paper, three different GLS algorithms that take the correlation and heteroscedasticity of the genotype-environment means by using a weighting matrix into account are proposed. The GLS estimation was applied using three different weighting matrices including an identity matrix (OLS), the full inverse covariance matrix of the genotype-environment means and an approximation of latter matrix. Using data simulated based on the parameters of real experiments, the GLS estimation was compared to the OLS estimation in terms of the mean squared error of the underlying model parameters and genotype-environment means. The results of this study show that GLS estimation outperforms OLS estimation in terms of the mean squared error and the effectiveness of the weighted estimation increases when the heterogeneity in the variances of the genotype-environment means increased.

5. General Discussion

5.1. Stage-wise analysis

In this thesis both the models that use genetic information (Chapter 2) and the AMMI/GGE model (Chapters 3 and 4) were estimated using a stage-wise analysis. In case of a stage-wise analysis, the scope always is to reproduce the single-stage analysis by the stage-wise analysis as good as possible. In case of fixed genotype, environment and interaction effects, the fully efficient two-stage analysis (Damesa et al., 2017), yields exactly the same point estimates and standard errors of the estimated genotype, environment and interaction effects. When genotype, environment or interactions are treated as random effects, the stage-wise analysis is identical to the single-stage analysis provided that the variance components are known (Piepho et al., 2012). In practice, variances need to be estimated and therefore the single-stage and the stage-wise analysis may differ in terms of the estimated effects and variances. Comparisons between single-stage and weighted two-stage analyses indicate, however, that the variances and the estimated means obtained by a weighted stage-wise estimation using either a diagonal weighting matrix or the full covariance matrix are very similar to those obtained in a single-stage analysis (Piepho, 2012; Schulz-Streeck et al., 2013; Damesa et al., 2017). Computational advantages of a two-stage analysis compared to a single-stage analysis generally arise when the dataset is large, and when the covariance structure single stage model is complex.

5.2. Evaluation of the predictive performance in cross validation

When the training and validation data are independent, the expected test-MSE can be decomposed into the variance of the validation data, the variance of the predicted validation data, and the squared bias of the prediction (Hastie and Tibshirani, 2001; James et al., 2013). When the training and validation data are not independent, the test-MSE is also affected by the covariance between the validation data and their predictions. Thus, to minimize the test-MSE in practice, one may pursue to minimize the different quantities that influence the test-MSE. For this purpose, a success criterion that is based on pairwise differences (Piepho, 1996) can be used to reduce the variance of the validation data and their predictions. The use of pre-processed data in a CV scheme (Chapter 3) can also be seen as an option to reduce the variance of the validation data and their predictions.

The contribution of the individual variance components to the variance of the predicted validation data is hard to identify in general, whereas one may compare the variance of the

predicted values for two CV schemes. In Chapter 3, the variance of the predicted validation data in case the rr-CV scheme probably leads to a higher variance of the predicted values than the cr-CV scheme. This may be attributed to the fact that the blocking structure of the original experimental design is destroyed in the rr-CV scheme which increases the variance of both the estimated effects and the predictions of the validation data. Such an increase in variance may be large especially when the experimental design entails incomplete blocks.

The variation of the expected test-MSE caused by the covariance between the validation data and their predictions can be reduced by using an adequate sampling strategy, e.g. the cr-CV scheme applied in Chapter 3 (Piepho, 1994). In this CV scheme, the training and validation data are from different replicates (and blocks) and therefore independent. In contrast, when the validation data consist of different replicates (rr-CV in Chapter2; Gauch and Zobel 1988), the training and validation data are not independent as they share replicate and block effects. Another way to create independent training and validation data consists of a linear transformation of the data that is based on the spectral decomposition of the covariance matrix of the data (Schulz-Streeck et al., 2013). In this way, the transformed data are conditionally independent with unit variance. The transformed data are denoted by rotated data in the following.

Generally, when one aims to evaluate the predictive performance in a CV, the different aspects that influence the success criterion need to be taken into account in order to evaluate the predictive performance of a model adequately. The results of Chapter 3 indicate that in case of AMMI/GGE models, the performance of different CV schemes is quite similar when the different aspects that influence the success criterion are taken into account.

5.3. Marker-based prediction

5.3.1. Extensions of marker-based prediction

The results of Chapter 2 indicate that when QTL- linked markers with large effects are present, marker-assisted selection (MAS) using only QTL linked markers outperforms genomic prediction (GP) while GP outperformed MAS when the effect size of QTL is small. This is in line to other studies that compare MAS and GP (Bernardo and Yu, 2007; Heffner et al., 2009; Meuwissen et al., 2001). The model used in Chapter 2 may be extended to increase the predictive ability of the models. This may be done by modelling the interaction between alleles

of the same gene (dominance effects) and/or the interaction of alleles from different genes (epistatic effects, Falconer and Mackay, 1996). Including epistatic effects in the model may be worthwhile (Jiang and Reif, 2015; Zang et al., 2015; He et al., 2016; Wolfe et al., 2016; He et al., 2017), whereas other reports suggest not to include epistatis effects (Miedaner et al., 2013; Shikha et al., 2017). Furthermore, epistasis may also be taken into account in mapping QTL markers (Ogutu et al., 2011; He et al., 2017). If the QTL are mapped without accounting for epistasis, the interaction between the QTL may still be included in the model to account for these interactions (Lynch and Walsh, 1998; p.443). Moreover, the marker-by-environment interaction can be taken into account by modelling marker-by-environment interactions for all markers (Schulz-Streeck et al., 2013), and/or modelling QTL-by-environment interactions. One may also consider mapping QTL linked markers for each environment separately. In Chapter 2 all marker effects were assumed to have homogeneous variances. This assumption can be relaxed by using BayesA or BayesB estimation (Meuwissen et al., 2001). In some cases, allowing for heterogeneous marker variances improved the predictive accuracy (Shikha et al., 2017) whereas for other datasets, models with heterogeneous marker variances did not outperform models that assume homogeneous variances of the markers (Howard et al., 2014). Modelling marker effects with heterogeneous variances may be worthwhile in cases, where there are markers with large effects. In Chapter 2, the presence of QTL-linked markers indicate that markers with large effects exist, thus allowing for heterogeneous marker variances could increase the predictive ability with the data used in Chapter 2.

5.3.2. Training procedures in marker-based prediction

CV schemes are typically used to determine the most appropriate marker based prediction model using the predictive ability. There are different objectives a breeder might want to pursue with the GP model. The predictive performance of a model usually depends on the objective pursued and on the training and validation procedure used in the CV. For example, using genotypes from different populations in the training and validation sets was shown to lead to a lower predictive performance than using training and validation data across populations (Lehermeier et al., 2014). Similarly, including the parents in the training set was shown to improve the predictive ability in rye (Bernal-Vasquez et al., 2017). In Chapter 2, the parents were not used to estimate the marker effect, thus including them to estimate marker effects may increase the predictive ability. Furthermore, when data consist of several location-year combinations in which the genotypes differ in each year, the use of genotyped checks that are

planted in each location-year combination can increase the predictive ability as the interaction between location-year combinations and the checks can be accounted for (Bernal-Vasquez et al., 2017).

5.3.3. Evaluating the predictive performance in marker-based prediction

When, the objective of the breeder in using a GP model is to estimate the genotype means across all environments, the training and validation sets should be created by sampling genotype means that were estimated across all environments such that the validation data are adjusted for genotype-environment interaction (GS-CV in Chapter 2). With MET data, the adjusted genotype means typically are correlated with heterogeneous variances. To meet the assumption of independence and homogeneity of variances, rotated means may be used in a CV scheme (Schulz-Streeck et al., 2013). The use of rotated data was shown to increase the predictive performance compared to using unrotated adjusted means and furthermore, rotated data led to a predictive performance which was quite similar to the single-stage analysis (Schulz-Streeck et al., 2013). The advantage of using rotated data lies in the fact that the validation data is also rotated such that the success criterion is computed based on the rotated data. In contrast, the validation data in the GS-CV of Chapter 2 was not rotated which may lead to an underestimation of the predictive ability. Thus the use of rotated data might increase the predictive ability in the GS-CV of Chapter 2. When a breeder aims to evaluate the prediction of the genotypes in an unobserved environment, the training datasets can be taken from all but one environment while the left out environment represents the validation data (ES-CV in Chapter 2; Utz et al., 2000). In such a CV scheme, the interaction of the genotypes with the validation environment as well as the effect of the validation environment cannot be estimated, thus the predictive ability obtained in this way is influenced by the environment effects of the validation environment and by genotype (marker) by environment interaction. For example, the predictive ability was found to be lower when genotypes from different environments (year-location combination) were used in the training and validation datasets compared to using genotypes from the same environment (Jiang et al., 2017; Song et al., 2017). To compute a success criterion in an ES-CV that is mainly independent of effects that are not of interest but cause variation in the validation data, e.g. effects of the validation environment, one may use a success criterion that is based on pairwise differences of the validation data. In this way, the success criterion is unaffected by the environment effects of the validation environment. Therefore, the predictive ability of the ES-CV scheme in Chapter 2 may be improved by using

pairwise genotype differences as a success criterion. The interaction of the genotypes and the validation environment, however, cannot be accounted for in an ES-CV scheme. Thus the success criterion in an ES-CV scheme indicates if there is substantial genotype-environment interaction. When the environments represent year-location combinations, a success criterion in which the location and year effects largely drop out can be created to evaluate the predictive ability accurately.

5.4. AMMI/GGE models

5.4.1. Interpretation of multiplicative terms

The AMMI/GGE model describes the interaction (AMMI) or the sum of genotype and interaction effect (GGE) by a sum of multiplicative terms, e.g. for AMMI $w_{ij} = \sum_{m=1}^{N} \lambda_m u_{im} v_{jm}$, where the elements of the multiplicative terms are subject to the constraints of the SVD. The SVD represents a technique to describe data in a low dimension using latent variables and it is usually used only to describe the main pattern of the data without providing a causal relationship between the latent variables and the data. In the framework of quantitative genetics, the multiplicative interaction may still be interpreted: For this purpose, one may consider that the different genotypes share certain (unknown) sets of genes with genotype-specific allelic combinations. Generally, several of such sets of genes may be present in the genome, each of which is represented by a multiplicative term $\lambda_m u_{im} v_{jm}$, thus there may be N such set of genes in total. The latent variable that refers to the i-th genotype and the m-th multiplicative term u_{im} can be regarded as a value that represents the contribution of the m-th set of genes to the development of the trait of the i-th genotype. This latent variable of the genotype u_{im} can capture both additive genetic effects and dominance/epistasis effects. The latent variables of the environments v_{jm} can be regarded as value that determines the contribution of the m-th set of genes to the development of the trait. The singular values λ_m can be seen as a scaling factor that represents the importance of the set of genes compared to another set of genes. The addition of the multiplicative terms represents the assumption that the different sets of genes do not interact.

With this interpretation of the multiplicative terms, the latent variables u_{im} can be regarded as a regression coefficient referring to the i-th genotype while the product of the latent variable v_{jm} and the scaling factors λ_m represents a covariate referring to the j-th environment.

Therefore, AMMI and GGE models can be interpreted similarly to regression models with known covariates (Chapter 2), while the values of the latent covariates are unknown and need to be estimated from the data.

5.4.2. Note on simulation studies

Chapters 2 and 3 were based on simulation studies. Simulation studies are very helpful in the comparison of different estimation methods or model selection procedures as it allows evaluating the performance of the methods when the true, data generating process is known. In this way, different CV schemes/estimation methods can be compared in terms of recovering the underlying model parameters in order to judge the performance of the different methods accurately. Generally, results obtained in simulation studies hold in practice only when the model that was used to generate the data well approximates the biological process that generates the trait.

5.4.3. Evaluating the predictive performance in AMMI/GGE models

While the simulation study of Chapter 3 showed that using adjusted data in a CV outperforms using the raw data, the proposed procedures do not account for the fact that the adjusted data are not independent. Therefore, the predictive performance of the model in the CV schemes could be improved by a weighted estimation of the AMMI/GGE model (Chapter 4).

Furthermore, in case of incomplete block designs, the success criterion can be calculated based on the pairwise genotypes differences within incomplete blocks to increase the predictive performance of the model. In this way, not only the replicate effects but also the variation in the validation data caused by incomplete blocks does not influence the test-MSE which may increase the predictive performance of the model. In case of row-column designs, one may use pairwise differences within the factor containing a larger number of levels as it should explain more variation than the factor with a lower number of levels.

In the CV schemes applied in Chapter 3, the data were adjusted for the design effects estimated from all data before applying a CV. Generally, an adjustment of the data can also be done by estimating the design effects from the training data. In case of the rr-CV scheme, the adjustments of the validation data can be done using the estimates obtained from the training data. In case of the cr-CV scheme, the design effects of the validation data cannot be estimated,

thus a success criterion that is based on pairwise differences could be used to account for the design effects contained in the validation data.

5.4.4. Further options to determine the number of multiplicative terms

Instead of CV schemes, statistical tests can be used to determine the number of multiplicative terms. So far, the tests proposed in the literature were proposed for independent data with homogeneous variances. In such case the test is based on the fact that the ratio between the sum of squares caused by multiplicative terms and the residual sum of squares is approximately F-distributed (Gollob, 1968; Cornelius, 1992). Some tests were found to be robust against deviations from the assumption of normality (Piepho, 1995). In the case of independent data, tests based on bootstrap procedures can also be used to determine the number of multiplicative terms (Forkman and Piepho, 2014). Such bootstrap procedures could be implemented for correlated data, i.e. when the experimental design entails random design effects. With such data, the sampling strategy of the bootstrap samples may need to take the experimental design into account to determine the number of multiplicative terms adequately.

5.4.5. Single-stage estimation of AMMI/GGE models

When the genotype-environment combinations are replicated within each environment, the AMMI/GGE model is usually estimated in two stages where the first stage consists of estimating the genotype-environment means while the second stage estimates the main effects and the multiplicative interaction. Strictly speaking, this two-stage estimation is only equivalent to single-stage estimation when the data is independent with homogeneous variances, i.e. with balanced data in a completely randomized design or in a randomized complete block design with fixed blocks. When this assumption is violated, i.e. when random replicate and block effects are contained in the underlying model (equation (2) of Chapter 3), the two-stage estimation does not represent the underlying model. Therefore, the AMMI/GGE model displayed in equation (2) of Chapter 3 may be estimated in a single stage using Algorithm 1 proposed in Chapter 4. This algorithm can be implemented using a mixed model package to estimate equation (2) of Chapter 3 iteratively. When random design effects are present, the main difference between the single-stage and the two-stage estimation arises from the fact that in case of the two-stage estimation, the variances and random effects are estimated assuming the maximum number of multiplicative terms. In contrast, the variances and random

effects in the single-stage estimation are estimated using only the number of terms that is to be estimated. As the main effects and the multiplicative interaction depend on the variance components, the estimated genotype-environment means estimated in the single-stage and the two-stage estimation are probably different. It may therefore be worthwhile to investigate the performances of the single stage and the two-stage estimation.

5.4.6. Estimation of AMMI/GGE models in case of non-normal data

The AMMI/GGE model can also be estimated in case of non-normal data in the framework of generalized linear models (van Eeuwijk, 1995). For a weighted estimation with non-normal data, Algorithm 1 of Chapter 4 could be implemented using statistical packages for generalized linear mixed models. In packages which allow to specify the covariance matrix of the errors, e.g. in the GLIMMIX procedure of SAS, the weighting matrix can be specified in the REPEATED statement. If the covariance matrix of the errors cannot be specified in the package, the response vector and associated design matrices of the AMMI/GGE model can be rotated to account for the weights.

5.4.7. Factor-analytic models

AMMI/GGE models consider the main and interaction effects as fixed effects. The assumption of fixed effects may be relaxed by the use of factor-analytic covariance structures for the interaction (Piepho, 1997; Smith et al., 2001b). In this way, the interaction effects can be assumed to be correlated which may be an appropriate assumption in practice. When the multiplicative interaction is modelled by random effects, CV can be used to determine the most appropriate prediction model (Piepho, 1998; Studnicki et al., 2017). In the application of CV schemes with factor-analytic models, the correlation induced by random effects of the experimental design needs to be taken into account, for example by an adjustment of the data before applying a CV scheme.

6. References

Bernal-Vasquez, A.M., Gordillo, A., Schmidt, M., and Piepho, H.P. 2017. Genomic prediction in early selection stages using multi-year data in a hybrid rye breeding program. BMC Genetics. 18: 51. DOI 10.1186/s12863-017-0512-8

Bernardo, R., and Yu, J. 2007. Prospects for genomewide selection for quantitative traits in maize. Crop Science. 47: 1082-1090.

Cooper, M., Stucker, R.E., DeLacy, I.H., and Harch, B.D. 1997. Wheat breeding nurseries, target environments, and indirect selection for gain yield. Crop Science. 37: 1168-1176.

Cooper, M. 1999. Concepts and strategies for plant adaptation research in rainfed lowland rice. Field Crop Research. 64: 13-34.

Cornelius, P.L., M. Seydsadr, and J. Crossa. 1992. Using the shifted multiplicative model to search for "separability" in crop cultivar trials. Theoretical and Applied Genetics. 84: 161-172. DOI: 10.1007/BF00223996

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

Eeuwijk, van F.A. 1995. Multiplicative interaction in generalized linear models. Biometrics. 51: 1017-1032.

Falconer, D.S., and Mackay, T.F.C. 1996. Introduction to quantitative genetics. Oliver & Boyd, Edinburgh.

Forkman, J., and Piepho, H.P. 2014. Parametric bootstrap methods for testing multiplicative terms in GGE and AMMI models. Biometrics. 70: 639-647. DOI: 10.1111/biom.12162

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

Gauch, H.G. Jr., and Zobel, R.W. 1988. Predictive and postdictive success of statistical analyses of yield trials. Theoretical and Applied Genetics. 76: 1-10. DOI: 10.1007/BF00288824

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

Gauch, H.G. Jr. 2013. A simple protocol for AMMI analysis of yield trials: Crop Science. 53: 1860-1869. DOI: 10.2135/cropsci2013.04.0241

Gollob, H.F. 1968. A Statistical model which combines features of factor analytic and analysis of variance techniques. Psychometrika. 33 No.1: 73-115. DOI: 10.1007/BF02289676

Hastie TJ, Tibshirani R, and Friedman J. 2009. The elements of statistical learning, 2nd edn. Springer, New York.

He, S., Schulthess, A.W., Mirdita, V., Zhao, Y., Korzun, V., Bothe, R., Ebmeyer, E., Reif, J.C., and Jiang, Y. 2016. Genomic selection in a commercial winter wheat population. Theoretical and Applied Genetics. 129: 641-651. DOI 10.1007/s00122-015-2655-1

He, S., Reif, J.C., Korzun, V., Bothe, R., Ebmeyer, E., and Jiang, Y. 2017. Genome‑wide mapping and prediction suggests presence of local epistasis in a vast elite winter wheat populations adapted to Central Europe. Theoretical and Applied Genetics. 130: 635-647. DOI 10.1007/s00122-016-2840-x

Heffner, E.L., M.E. Sorrells, and Jannink, J.L. 2009. Genomic selection for crop improvement. Crop Science. 49: 1-12. DOI: 10.2135/cropsci2008.08.0512

Howard, R., Carriquiry, A.L., and Beavis, W.D. 2014. Parametric and nonparametric statistical methods for genomic selection of traits with additive and epistatic genetic architectures. Genes, Genomes, Genetics. 4: 1027-1046. DOI: 10.1534/g3.114.010298

James, G., Witten, D. Hastie, T., and Tibshirani, R. 2013. An introduction to statistical learning. Springer, New York.

Jiang, Y., and Reif, C.J. 2015. Modeling epistasis in genomic selection. Genetics. 201: 759-768. DOI: 10.1534/genetics.115.177907

Jiang, Y., Schulthess, A.W., Rodemann, B., Ling, J., Plieske, J., Kollers, S., Ebmeyer, E., Korzun, V., Argiller, o., Stiewe, G., Ganal, M.W., Röder, M.S., and Reif, J.C. 2017. Validation the prediction accuracies of marker-assisted and genomic selection of Fusarium head blight resistance in wheat using an independent sample. Theoretical and Applied Genetics. 130: 471-482. DOI:10.1007/s00122-016-2827-7

Lehermeier, C., Krämer, N., Bauer, E., Bauland, C., Camisan, C., campo, L., Flament, P., Melchinger, A.E., Menz, M., Meyer, N., Moreau, L., Moreno-Gonzales, J., Ouzunove, M., Pausch, H., Ranc, N., Schipprack, W., Schönleben, M., Walter, H., Charcosset, A., and Schön, C.C. 2014. Usefulness of multiparental populations of maize (Zea mays L.) for genome-based prediction. Genetics. 198: 3-16. DOI: 10.1534/genetics.114.16194

Lynch, M., and Walsh, B. 1998. Genetics and analysis of quantitative traits. Sinauer Associates, Massachusetts

Miedaner, T., Zhao, Y., Gowda, M., Longin, C.F.H., Korzun, V., Ebmeyer, E., Kazman, E., and Reif, J.C. 2013. Genetic architecture of resistance to septoria tritici blotch in European wheat. BMC genomics. 14: 858.

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

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

Piepho, H.P. 1994. Best linear unbiased prediction (BLUP) for regional yield trials: a comparison to additive main effects and multiplicative interaction (AMMI) analysis. Theoretical and Applied Genetics. 89: 647-654. DOI: 10.1007/BF00222462

Piepho, H.P. 1995. Robustness of statistical tests for multiplicative terms in the additive main effects and multiplicative interaction model for cultivar trials. Theoretical and Applied Genetics. 97: 195-201. DOI: 10.1007/BF00221987

Piepho, H.P. 1997. Analyzing genotype-environment data by mixed models with multiplicative terms. Biometrics. 53: 761-766.

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

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

Rodrigues, P.C., Malosetti, M., Gauch, H.G. Jr., and van Eeuwijk, F.A. 2014. A weighted AMMI algorithm to study genotype-by-environment interaction and QTL-by-Environment interaction. Crop Science. 54: 1555-1569. DOI: 10.2135/cropsci2013.07.0462

Schulz-Streeck, T., Ogutu, J., and Piepho, H.P. 2013. Comparisons of single-stage and two-stage approaches in genomic selection. Theoretical and Applied Genetics. 126: 69-82.

Shikha, M., Kanjka, A., Rao, A.R., Mallikarjuna, M.G., Gupta, H.S., and Nepolean, T. 2017. Genomic selection ofr drought tolerance using genome-wide SNPs in maize. Frontiers in plant science. 8: 550. DOI: 10.3389/fpls.2017.00550

Song, J., Carver, B.F., Powers, C., Yan, L., Klapste, J., El-Kassaby, Y.A., and Chen, C. 2017. Practical application of genomic selection in a doubled-haploid winter wheat breeding program. Molecular breeding. . DOI 10.1007/s11032-017-0715-8

Smith, A., Cullis, B., and Gilmour, A. 2001a. The analysis of cop variety evaluation data in Australia. Australian and New Zealand Journal of Statistics. 43: 129-145. DOI:10.1111/1467-842X.00163

Smith, A., Cullis, B.R., and Thompson, R. 2001b. Analysing variety by environment data using multiplicative mixed models and adjustment for spatial field trend. Biometrics. 57: 1138-1147. DOI: 10.1111/j.0006-341X.2001.01138.x

Studnicki, M., Paderewski, J., Piepho, H.P., and Wojcik-Gront, E. 2017. Prediction accuracy and consistency in cultivar ranking for factor-analytic linear mixed models for winter wheat multienvironmental trials. Crop Science. 57: 2506-2516.

Utz, H.F., Melchinger, A.E., and Schön, C.C. 2000. Bias and sampling error of the estimated proportion of genotypic variance explained by quantitative trait loci determined from experimental data in maize using cross validation and validation with independent samples. Genetics. 154: 1839-1849.

Wolfe, M.D., Kulakow, P., Rabbi, I.Y., and Jannik, J.L. 2016. Marker-based estimates reveal significant nonadditive effects in clonally propagated cassava (Manihot esculenta): Implications for the Prediction of Total Genetic Value and the Selection of Varieties. Genes, Genomes, Genetics. 6: 3497-3506. DOI: 10.1534/g3.116.033332

Yan, W., Hunt, L.A., Sheng, Q., and Szlavnics, Z. 2000. Cultivar evaluation and mega-environment investigation based on the GGE biplot. Crop Science. 40: 597-605. DOI:10.2135/cropsci2000.403597x

Yan, W., Kang, M.S., Ma, B., Woods, S., and Cornelius, P.L. 2007. GGE Biplot vs. AMMI analysis of genotype-by-environment data. Crop Science. 47: 643-655. DOI: 10.2135/cropsci2006.06.0374

Yates, F. and Cochran, W.G. 1938. The analysis of groups of experiments. The journal of agricultural science. 28: 556-580.

Zang, J, Singh, A., Mueller, D.S., and Singh, A.K. 2015. Genome-wide association and epistasis studies unravel the genetic architecture of sudden death syndrome resistance in soybean. The Plant Journal. 84: 1124-1136. DOI: 10.1111/tpj.13069

7. Summary

In plant breeding, estimation of the performance of genotypes across a set of tested environments (genotype means), and the estimation of the environment-specific performances of the genotypes (genotype-environment means) are important tasks. For this purpose, breeders conduct multi-environment trials (MET) in which a set of genotypes is tested in a set of environments. The data from such experiments are typically analysed by mixed models as such models for example allow modelling the genotypes using random effects which may be correlated according to their genetic information. The data from MET are often high-dimensional and the covariance matrix of the data may contain many parameters that need to be estimated. To circumvent computational burdens, the data can be analysed in a stage-wise fashion. In the stage-wise analysis, the covariance matrix of the data needs to be taken into account in the estimation of the individual stages. In the analysis of MET data, there is usually a set of candidate models from which the one that fits bets to the objective of the breeder needs to be determined. Such a model selection can be done by cross validation (CV). In the application of CV schemes, different objectives of the breeder can be evaluated using an appropriate sampling strategy. In the application of a CV, both the sampling strategy and the evaluation of the model need to take the correlation of the data into account to evaluate the model performance adequately.

In this work, two different types of models that are used for the analysis of MET were focused. In Chapter 2, models that use genetic marker information to estimate the genotype means were considered. In Chapters 3 and 4, the estimation of genotype-environment means using models that include multiplicative terms to describe the genotype-environment interaction, namely the additive main effects and multiplicative interaction (AMMI), and the genotype and genotype-environment interaction (GGE) model, were focused. In all the Chapters, the models were estimated in a stage-wise fashion. Furthermore, CV was used in Chapters 2 and 3 to determine the most appropriate model from a set of candidate models.

In Chapter 2, two traits of a biparental lettuce (*Lactuca sativa* L.) population were analysed by models for (i) phenotypic selection, (ii) marker-assisted selection using QTL-linked markers, (iii) genomic prediction using all available molecular markers, and (iv) a combination of genomic prediction and QTL-linked markers. Using different sampling strategies in a CV, the predictive performances of these models were compared in terms of different objectives of a breeder, namely predicting unobserved genotypes, predicting genotypes in an unobserved environment, and predicting unobserved genotypes in an unobserved environment. Generally,

the genomic prediction model outperformed marker assisted and phenotypic selection when there are only a few markers with large effects, while the marker assisted selection outperformed genomic prediction when the number of markers with large effects increases. Furthermore, the results obtained for the different objectives indicate that the predictive performance of the models in terms of predicting (unobserved) genotypes in an unobserved environment is reduced due to the presence of genotype-environment interaction.

In AMMI/GGE models, the number of multiplicative terms can be determined by CV. In Chapter 3, different CV schemes were compared in a simulation study in terms of recovering the true (simulated) number of multiplicative terms, and in terms of the mean squared error of the estimated genotype-environment means. The data were simulated using the estimated variance components of a randomized complete block design and a resolvable incomplete block design. The effects of the experimental design (replicates and blocks) need to be taken into account in the application of a CV in order to evaluate the predictive performance of the model adequately. In Chapter 3, the experimental design was accounted for by an adjustment of the data for the design effects estimated from all data before applying a CV scheme. The results of the simulation study show that an adjustment of the data is required to determine the number of multiplicative terms in AMMI/GGE models. Furthermore, the results indicate that different CV schemes can be used with similar efficiencies provided that the data were adjusted adequately.

AMMI/GGE models are typically estimated in a two-stage analysis in which the first stage consists of estimating the genotype-environment means while the second stage consists of estimating main effects of genotypes and environments and the multiplicative interaction. The genotype-environment means estimated in the first stage are not independent when effects of the experimental design are modelled as random effects. In such a case, estimation of the second stage should be done by a weighted (generalized least squares) estimation where a weighting matrix is used to take the covariance matrix of the estimated genotype-environment means into account. In Chapter 4, three different algorithms which can take the full covariance matrix of the genotype-environment means into account are introduced to estimate the AMMI/GGE model in a weighted fashion. To investigate the effectiveness of the weighted estimation, the algorithms were implemented using different weighting matrices, including (i) an identity matrix (unweighted estimation), (ii) a diagonal approximation of the inverse covariance matrix of the genotype-environment means, and (iii) the full inverse covariance matrix. The different weighting strategies were compared in a simulation study in terms of the

mean squared error of the estimated genotype-environment means, multiplicative interaction effects, and Biplot coordinates. The results of the simulation study show that weighted estimation of the AMMI/GGE model generally outperformed unweighted estimation. Furthermore, the effectiveness of a weighted estimation increased when the heterogeneity in the covariance matrix of the estimated genotype-environment means increased.

The analysis of MET in a stage-wise fashion is an efficient procedure to estimate a model for MET data, whereas the covariance structure of the data needs to be taken into account in each stage in order to estimate the model appropriately. When correlated data are used in a CV, the correlation can be taken into account by an appropriate choice of training and validation data, by an adjustment of the data before applying a CV scheme and by the success criterion used in a CV scheme.

8. Zusammenfassung

In der Pflanzenzüchtung ist die Schätzung der mittleren Leistung verschiedener Genotypen über mehrere Umwelten hinweg (Genotyp-Mittelwerte) sowie die Schätzung der umweltspezifischen Leistung von Genotypen (Genotyp-Umwelt-Mittelwerte) von zentraler Bedeutung. Um die Mittelwerte der Genotypen zu schätzen, werden Versuche durchgeführt, in denen mehrere Genotypen an mehreren Umwelten getestet werden. Daten solcher Multi-Umwelt Versuche (MUV) werden oft anhand gemischter Modelle analysiert. Diese Modelle sind in der Pflanzenzüchtung von besonderer Bedeutung, da Genotyp-Effekte durch korrelierte Zufallseffekte so modelliert werden können, dass sie entsprechend der genetischen Information der Genotypen korrelieren. Die Daten aus MUV sind oft hoch-dimensional und darüber hinaus kann die Kovarianzstruktur der Daten viele Parameter enthalten, die geschätzt werden müssen. Um Engpässe im Hinblick auf die Rechenkapazität zu vermeiden, können die Daten stufenweise analysiert werden. In einer solchen stufenweisen Analyse muss die Kovarianzstruktur der Daten bei der Schätzung der jeweiligen Stufen berücksichtigt werden. Bei der Analyse von Daten aus MUV gibt es in der Regel mehrere Modelle, die zur Analyse herangezogen werden können. Die Bestimmung des Modells, das am besten zu den Zielen des Züchters passt kann anhand von Kreuzvalidierung (KV) bestimmt werden. Mittels KV kann man durch eine geeignete Wahl von Trainingsdaten und Validierungsdaten (Stichproben-Ziehung) verschiedene Ziele des Züchters evaluieren. In einer KV muss sowohl die Stichproben-Ziehung als auch die Evaluation der Vorhersagefähigkeit des Modells die Korellation der Daten berücksichtigen, um die Vorhersagefähigkeit des Modells zu bestimmen.

In dieser Arbeit werden zwei Modelle behandelt, die zur Analyse von MUV herangezogen werden können. Kaptiel 2 handelt von Modellen, die genetische Marker nutzen um die Genotyp Mittelwerte zu schätzen. Kaptiel 3 und 4 beinhalten Modelle zur Schätzung der Genotyp-Umwelt Mittelwerte. In diesen Modellen werden die Genotyp-Umwelt Interaktionen anhand von multiplikativen Termen modelliert. Eines der betrachteten Modelle beinhaltet additive Haupteffekte für Genotypen und Umwelten und multiplikative Interaktionseffekte (englisch: AMMI), wohingegen das andere Modell aus Haupteffekten für Umwelten und einer mutliplikativen Genotyp und Genotyp-Umwelt Interaktion (englisch: GGE) besteht. Die Modelle der jeweiligen Kapitel wurden alle stufenweise geschätzt. Darüber hinaus wurde in Kaptiel 2 und 3 KV angewendet, um das Modell, das am besten zum Ziel des Züchters passt, zu bestimmen.

In Kaptiel 2 werden zwei verschiedene Merkmale einer biparentalen Blattsalat (*Lactuca sativa* L.) Population anhand von Modellen für die (i) phenotypische Selektion, (ii) markergestützte Selektion mit QTL-assoziierten Markern, (iii) genomische Selektion anhand aller verfügbaren genetischen Marker, und (iv) einer Kombination von genomischer Selektion und markergestützter Selektion analysiert. In diesem Kapitel wurden verschiedene Ziele, die ein Züchter verfolgen kann anhand verschiedener Methoden der Stichproben-Ziehung analysiert. Die Ziele bestanden aus der Vorhersage von unbeobachteten Genotypen, der Vorhersage von beobachteten Genotypen in unbeobachteten Umwelten, und der Vorhersage von unbeobachteten Genotypen in unbeobachteten Umwelten. Die Ergebnisse der Analysen zeigen, dass die genomische Vorhersage die markergestützte und die phänotypische Selektion übertrifft, wenn es nur wenige Marker mit großen Effekten gibt. Dahingegen übertrifft die markergestützte Selektion die anderen Modelle, wenn es mehrere Marker mit großen Effekten gibt. Im Weiteren deuten die Ergebnisse darauf hin, dass die Vorhersage von (un)beobachteten Genotypen in unbeobachteten Umwelten durch Genotyp-Umwelt Interaktionen verringert wird.

Im AMMI/GGE Modell kann die optimale Anzahl multiplikativer Termen mittels KV bestimmt werden. In Kapitel 3 werden verschiedene KV-Methoden, die zur Bestimmung der Anzahl multiplikativer Terme herangezogen werden können, in einer Simulationsstudie verglichen. Die verschiedenen KV-Methoden wurden in Bezug auf die Bestimmung der wahren (simulierten) Anzahl multiplikativer Terme, und in Bezug auf die Schätzung der wahren Genotyp-Umwelt-Mittelwerte verglichen. Die Simulation der Daten erfolgte entsprechend einer randomisierten vollständigen Blockanlage und einer unvollständigen Blockanlage. Bei der Anwendung einer KV muss das Versuchsdesign (Effekte von Wiederholungen und unvollständigen Blöcken) berücksichtigt werden, um die Anzahl der multiplikativen Terme zu schätzen. In Kapitel 3 wurde das Versuchsdesign berücksichtigt, indem die Daten vor der Anwendung einer KV um die Designeffekte korrigiert wurden. Die Ergebnisse der Simulationsstudie zeigen, dass die Daten vor Anwendung einer KV um die Designeffekte korrigiert werden sollten, um die Anzahl der multiplikativen Terme zu bestimmen. Ausserdem zeigen die Ergebnisse, dass verschiedene Kombinationen von Datenkorrektur und Stichproben-Ziehung zu sehr ähnlichen Ergebnissen führen.

Das AMMI/GGE Modell wird üblicherweise in zwei Stufen geschätzt. Dabei besteht die erste Stufe aus der Schätzung der Genotyp-Umwelt-Mittelwerte, wohingegen die zweite Stufe die Haupteffekte von Umwelten (und Genotypen) und die multiplikativen Interaktionen schätzt.

Die geschätzten Genotyp-Umwelt-Mittelwerte aus der ersten Stufe sind nicht unabhängig wenn die Designeffekte durch Zufallseffekte modelliert werden. Daher sollte die zweite Stufe die Kovarianzmatrix der geschätzten Genotyp-Umwelt-Mittelwerte anhand einer gewichteten Schätzung (generalisierte Kleinstquadrat-Schätzung) berücksichtigen, um die Effekte des AMMI/GGE Modells zu schätzen. In Kapitel 4 werden drei verschiedene Algorithmen anhand derer eine gewichtete Schätzung möglich ist vorgestellt. Um die Effektivität der gewichteten Schätzung zu untersuchen, wurden die Algorithmen mit verschiedenen Gewichtungsmatrizen implementiert. Als Gewichtungsmatrizen dienten (i) die Einheitsmatrix (ungewichtete Schätzung), (ii) eine diagonale Approximation der Inversen der Kovarianzmatrix, und (iii) die Inverse der Kovarianzmatrix. Die verschiedenen Gewichtungsmethoden wurden in einer Simulationsstudie in Bezug auf die mittlere quadratische Abweichung der Genotyp-Umwelt-Mittelwerte, der multplikativen Interaktionseffekte, und der Biplot Koordinaten verglichen. Die Ergebnisse zeigen, dass eine gewichtete Schätzung bezüglich der untersuchten Kriterien deutlich besser als eine ungewichtete Schätzung abschneidet. Die Ergebnisse zeigen ausserdem, dass die Effektivität einer gewichteten Schätzung zunimmt, wenn die Heterogenität der Kovarianzmatrix steigt.

Die stufenweise Analyse von Daten aus MUV ist eine effiziente Strategie, um ein Modell zu schätzen. In den jeweiligen Stufen muss die Kovarianzmatrix der Daten berücksichtigt werden, um das Modell zu schätzen. Wenn eine KV mit korrelierten Daten durchgeführt wird, kann die Korrelation durch eine geeignete Stichproben-Ziehung, eine Korrektur der Daten vor Anwendung einer KV, und durch das Evaluationskriterium berücksichtigt werden.

9. Acknowledgements

First of all, thanks to my supervisor Prof. Dr. Hans-Peter Piepho for his excellent support, very helpful advice, and his patience during the development of the individual papers and this thesis. Thanks also for providing the opportunity to work in projects that were not related to this thesis.

Thanks also to Dr. Johannes Forkman and Dr. W. Ahmed Malik for very helpful discussions, feedback, and proofreading during the development of Chapters 3 and 4.

Many thanks to my office mate Dr. Joseph Ogutu for many nice discussions and conversations about work and the world.

Special thanks to my office neighbours Dr. Angela-Maria Bernal-Vasquez, Paul Schmidt, and Dr. Nha Vo-Tan for perfect management of the coffee and lunch breaks, for sharing everyday issues, and for numerous nice conversations during walks around the castle.

Special thanks also to Sofie Pødenphant Jensen for deciding to visit our group and for an unforgettable summer.

Thanks also to all members of the biostatistics group that were not mentioned.

Lastly, great thanks to my family and friends that do not deal with biostatistics for the support throughout the years.

10. List of Publications

Hadasch, S., Simko, I., Hayes, R.J., Ogutu, J., and Piepho, H.P. 2016. Comparing predictive abilities of phenotypic and marker-assisted selection methods in a biparental lettuce population. The Plant Genome. 9(1): 1-11. DOI: 10.3835/plantgenome2015.03.0014

Hadasch, S., Forkman, J., and Piepho, H.P. 2017. Cross validation in AMMI and GGE models: A comparison of methods. Crop Science. 57: 264-274. DOI: 10.2135/cropsci2016.07.0613

Hadasch, S., Forkman, J., Malik, W.A., and Piepho, H.P. 2017. Weighted estimation of AMMI and GGE models. Journal of Agricultural, Biological, and Environmental Statistics. 23: 255-275. https://doi.org/10.1007/s13253-018-0323-z