Aus dem

Institut für Kulturpflanzenwissenschaften

Universität Hohenheim

Fachgebiet: Bioinformatik

Prof. Dr. H.-P. Piepho

Mixed modelling for phenotypic data from plant breeding

Dissertation

zur Erlangung des Grades eines Doktors

der Agrarwissenschaften

vorgelegt

der Fakultät Agrarwissenschaften

von

Diplom-Agrarbiologe

Jens Möhring

aus Sprockhövel

2010

Die vorliegende Arbeit wurde am 09.02.2011 von der Fakultät Agrarwissenschaften der Universität Hohenheim als "Dissertation zur Erlangung des Grades eines Doktors der Agrarwissenschaften" angenommen.

Tag der mündlichen Prüfung: 17.03.2011

1. Prodekan:	Prof. Dr. A. Fangmeier
Berichterstatter, 1. Prüfer:	Prof. Dr. H.-P. Piepho
Mitberichterstatter, 2. Prüfer:	Prof. Dr. A. E. Melchinger
3. Prüfer:	Prof. Dr. C. P. W. Zebitz

Content

---

[1] Fischer*, S., J. Möhring*, C.C. Schön, H.-P. Piepho, D. Klein, W. Schipprack, H.F. Utz, A.E. Melchinger and J.C. Reif. 2008. Plant Breeding 127:446-451.

[2] Möhring, J., and H.-P. Piepho. 2009. Crop Sci. 49:1977-1988.

[3] Möhring, J., A.E. Melchinger, and H.-P. Piepho. Crop Sci.51:470-478.

[4] Fischer*, S., J. Möhring*, H.P. Maurer, H.-P. Piepho, E.-M. Thiemt, C.C. Schön, A.E. Melchinger and J.C. Reif. 2009. Crop Sci. 49:2119-2122.

* Both authors contributed equally.

Abbreviations

AMMI	additive main effect and multiplicative interaction
ANOVA	analysis of variance
BLUE	best linear unbiased estimator
BLUP	best linear unbiased prediction
DNA	deoxy-ribo-nucleic acid
GCA	general combining ability
GGE	genotype genotype-environment
GREG	genotype regression
LD	linkage disequilibrium
MAR	missing at random
MAS	marker assited selection
MCAR	missing completely at random
MET	multi-environment trial
MINQUE	minimum norm quadratic unbiased estimation
MNAR	missing not at random
MSEP	mean square error
QTL	quantitative trait loci
p-rep	partially replicated
RCBD	randomised complete block design
REML	restricted maximum likelihood
SCA	specific combining ability
SHMM	shifted multiplicative model
SNP	single nucleotid polymorphism
SREG	site regression

1. General Introduction

Plant breeding is a process of creating new variability, e.g. by crossing genotypes and evaluating these genotypes to select more preferable genotypes under limited financial resources. During this process, large amounts of data are measured and analysed. In the beginning of plant breeding these data were almost exclusively phenotypic data like ratings or yield measurements. Since around 30 years, genetic information like DNA-marker data have been used in addition. They are either linked to a trait of interest (e.g. restorer, resistance) or are used to describe the genetic relationships between genotypes. Genetic data are needed in quantitative trait loci (QTL) studies, association mapping or genome-wide selection, but despite the differences between these methods, finally all of them are based on analysis of phenotypic data. Thus an efficient analysis of phenotypic data is an important prerequisite for successful plant breeding programs.

In classical QTL studies (Schön et al., 1993) offspring of a biparental cross are analysed using DNA markers and a phenotypic trait of interest to detect marker-trait correlations. Linked markers can then be used to indirectly select for this trait in a target population. Depending on the costs, the heritability, availability, and genetic as well as environmental variances and covariances, direct phenotypic or indirect marker assisted selection (MAS) is preferred (Ribaut and Hoisington, 1998; Dubcovsky, 2004). Instead of using offspring of only two parents with limited genetic variation, multiple crosses can be used to explain more genetic variation. In both cases the resolution of QTL detection is limited because of a limited number of meioses and therefore a limited number of recombinations in the offspring. Further, if the genetic variablility in the crosses used for detecting QTLs varies from the one in the target population, transferability of results is limited. While the resolution and transferability are limited, the analysis of

phenotypic data from biparental crosses or multi-crosses is mostly simple, due to the fact that these experiments are designed for this special purpose.

In contrast, association studies use phenotypic data from regular plant breeding processes, thus the same data as normally used for phenotypic selection in plant breeding. With these data a higher resolution is possible because of the large number of meioses accumulated in breeding history. Additionally, the genetic variation in test and target population are more similar, which makes transferability more likely. Association studies use the linkage disequilibrium (LD) between marker and a trait of interest. The main important point in association studies is to separate LD caused by population structure or relatedness between offspring from LD caused by marker-trait associations. Several methods, e.g. using a kinship matrix or using estimates of the population structure, were proposed (Yu et al., 2006; Stich et al., 2008). The analysis of phenotypic data for association mapping requires more advanced statistical methods compared to the analysis of designed experiments for QTL detection. The methods must account for field and mating design and cultivar specifics in the analysis. This thesis will show how to model breeding data for a range of cultivars in a mixed model framework. The models can be used either for phenotypic selection or for genetic studies.

New technologies for obtaining cheap single nucleotid polymorphism (SNP)-marker data allow the evaluation of thousands of markers, so often more markers are analysed than genotypes are available (Meuwissen et al., 2001). Thus, a standard multiple regression analysis for all marker data is impossible. In this case, either markers are pre-selected or machine-learning approaches like boosting (Bühlmann and Hothorn, 2007) and support vector machines or ridge regression are used (Piepho, 2009). These approaches, which essentially use all markers or a large fraction of those available, are refered to as genome-wide or genomic selection (Jannink et al., 2010). For quantitative traits genomic selection

outperforms MAS (Bernardo, 2007). As in association studies, genome-wide selection requires the efficient analysis of phenotypic data from plant breeding processes. So the analysis of phenotypic data is the basis of phenotypic selection as well as for association studies or genomic selection.

As costs for marker data decrease and available marker information increases, the relative importance of cost and time-efficient analysis of phenotypic data is getting more important. For this reason the main focus of this thesis is the development of methods for an adequate analysis of phenotypic data, which requires appropriately considering the experimental field and mating design and the genetic structure of the breeding data.

Experimental design of plant breeding trials

Plant breeders conduct and analyse large series of multi-environment plant breeding trials, either for selection in the breeding process or for linkage/ association studies. A large number of genotypes is tested in several locations and years. Depending on the selection stage and thus depending on the available amount of seed, unreplicated or replicated field trials are conducted (Kempton, 1984). If the amount of seed is limited, simple augmented designs with replicated checks (Federer, 1961), augmented lattice square designs (Federer, 2002; Williams and John, 2003), partially replicated designs (p-rep, Cullis et al., 2006; Smith et al., 2006), or augmented p-rep designs (Williams et al., 2010) are used. Checks in augmented designs are placed at random within blocks or replicated on a fixed grid, e.g. each tenth plot. Within the experimental field designs, errors of observations can be modelled as uncorrelated or spatially correlated. In the latter case the correlation can be handled by a large number of spatial models like autoregressive (Gilmour et al., 1997) and linear variance (Piepho and Williams, 2010). It depends on the point of view whether a spatial model is the baseline

model (Gilmour et al., 1997) and additional effects for blocks, rows, columns, tractor reeling etc. are added afterwards as needed, or the randomization-based block model is the baseline model and spatial correlation is added in case this improves the model fit (Piepho and Williams, 2010; Müller et al., 2010).

In replicated experiments, for practical reasons plant breeders in Germany often subdivide genotypes into groups of 25-100 genotypes and test them in separate trials, connecting the trials by common checks. The trials are often designed as randomised complete block design (RCBD) or as lattices and mostly genotypes are filled in subgroups cross by cross. If a subdivision is not required for technical or logistical reasons, using a single α-design is preferable (Piepho et al., 2006), mainly because of reduced space required for check plots. Additionally, the randomization includes all genotypes, thus less heterogeneous standard errors of pairwise comparisions are obtained.

Multi-environment trials (METs) are often analysed using mixed models (Smith et al., 2001; Smith et al., 2005; Piepho et al., 2008). To represent the data structure in METs, effects for year and location as well as effects for the randomization structure (trial, replicate and block effects) can be taken as random. To reduce demand on computing resources, effects not involving genotypes can be treated as fixed. For example, when year and location main effects are taken as fixed, we are ignoring inter-environment information, which is usually low (Piepho and Möhring, 2006).

MET from plant breeding programs usually include a large number of genotypes. The data are unbalanced because of selection or limitations of financial resources or seed availablility. Using a mixed model framework with large numbers of genotypes and an unbalanced data structure requires complex models and large computational ressources. To reduce the demand on computing time or for other

practical reasons the analysis is usually subdivided in two stages (Frensham et al., 1997). In the first stage adjusted means of genotypes for each trial or location are estimated. Often, several trials are conducted at the same location and each trial is analysed separately. This implies the assumption of heterogeneous residual variances between trials in the same location, which may not generally be plausible or efficient. Therefore in this thesis, data are analysed across trials for each location. In the second stage, the adjusted means are then submitted to a mixed model. The same procedure of stage-wise analysis is used in cultivar testing in official registration trials with one trial per location (Hühn, 1997; Piepho and Michel, 2000; Smith et al., 2001; Laidig et al., 2008). In the case of official field trials, stage-wise analysis resulted in mean estimates comparable to one-stage analysis (Hühn, 1997). In plant breeding the structure of unbalanceness and selection is more complex and it is not known, whether a stage-wise analyses has an influence on the results in the case of unbalanced plant breeding MET under German conditions. This thesis will analyse four large series of plant breeding METs to answer the question, whether the widely used stage-wise analysis can be recommended for plant breeding data or a single-stage analysis is required. In the two-stage analysis each trial is analysed separatly and genotypes are often not randomized within trials. Potential problems of this incomplete randomization will be discussed in this thesis.

Mating designs in plant breeding

Depending on the crop species, different breeding strategies are used. Schnell (1982) distinguished four types of varieties: pure-line varieties, synthetic varieties, hybrids and clone varieties. In self-pollinating crops like wheat and barley, crosses are performed within one genepool. Classically, progenies of crosses become homozygous by using doubled haploids or repeated selfing in single seed descent or pedigree selection during the breeding process. Depending on the selection

stage and the breeding scheme the genetic variances vary. For cross-fertilizated crops like maize, hybrid breeding is used. Depending on the number of genepools used for crossing, the mating design is called factorial or diallel. In a factorial, parents belong to two genepools, e.g. Flint and Dent in European maize (Schrag et al., 2009) or Lochow and Petkus in rye (Fischer et al., 2010). In wheat hybrid breeding, parents belong to one genepool. If both parents are from the same genepool, the mating design is called a diallel. Thus a diallel can be understood as a factorial with the additional restriction, that the male and female genepool are identical. Diallels can be divided in four methods, depending on whether parents or reciprocal crosses or both are included (Griffing, 1956). In hybrid breeding, the interest lies in estimating general combining ability (GCA) effects and specific combining ability (SCA) effects.

For diallel analysis Hayman (1954), Griffing (1956) and Gardner and Eberhart (1966) proposed models with analysis based on ANOVA tables. All of these analyses are based on a quantitative-genetic model and some allow the separation of additive and non-additive effects. Mixed model-based methods of analysis for diallel methods without parents have been proposed by Zhu and Weir (1994a, b, 1996a, b) and Xiang and Li (2001), but no extension for diallel methods with parents is available. In the case of the diallel method without parents, the ANOVA-based methods and REML-based methods result in identical estimates for all models when data are completely balanced and all ANOVA estimates are non-negative. If parents (i.e. selfed crosses for fully inbreds) are used within the diallel, the models vary because of different restrictions within the model, as shown in this thesis. A general framework to handle all kinds of diallels within a mixed model package using REML is lacking. This thesis will therefore propose a general framework for analysing diallels. It will be shown that the other mentioned models can be seen as special cases within this general framework and that they can be imlemented by adding restrictions to this general model.

Objectives

This thesis will demonstrate the use of BLUP for analysing plant breeding data, i.e. the use of correlated information to enhance estimation of breeding values. Plant breeding data often exhibit special experimental designs and structures resulting in some problems using mixed models in a time- and cost-efficient analysis. Some of these problems will be addressed here.

In Chapter 2 the handling of unbalanced maize (*Zea mays*) data in multi-environmental trials (METs) with missing data and a factorial mating design in each environment are described. Trends of variance component ratios and means are calculated over 30 years.

In Chapter 3 the impact of using two stages for analysing METs, thus estimating adjusted means per location and then averaging these estimates over locations in a second stage, instead of estimating genotype effects in one stage is investigated for four series of breeding trials. The structures of these series are described and recognized in the analysis. Arising missing data problems are discussed.

In Chapter 4 several models for diallel analysis are compared and a general REML-based model for analysing data arising from all diallel methods is proposed.

An extension of the diallel model developed in Chapter 4 as well as the model for factorials in Chapter 2 are used in Chapter 5 for triticale data to compare analysis assuming one or two genepools within a group of 21 parental lines in triticale.

References:

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

Bühlmann, P., and T. Hothorn. 2007. Boosting algorithms: Regularization, prediction and model fitting. Statist. Sci. 22:477-505.

Cullis, B.R., A.B. Smith, and N.E. Coombes. 2006. On the design of early generation variety trials with correlated data. J. Agric. Biol. Environ. Stat. 11:381-393.

Dubcovsky, J. 2004. Marker-assisted selection in public breeding programs: The wheat experience. Crop Sci. 44:1895-1898.

Federer, W.T. 1961. Augmented designs with one-way elimination of heterogeneity. Biometrics 17:447-473.

Federer, W.T. 2002. Construction and analysis of an augmented lattice square design. Biom. J. 44:251-257.

Fischer, S., A.E. Melchinger, V. Korzun, P. Wilde, B. Schmiedchen, J. Möhring, H.-P. Piepho, B.S. Dhillon, T. Würschum, and J.C. Reif. 2010. Molecular marker assisted broadening of the Central European heterotic groups in rye with Eastern European germplasm. Theor. Appl. Genet. 120:291-299.

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

Gardner, C.O., and A.S. Eberhart. 1966. Analysis and interpretation of the variety cross diallel and related populations. Biometrics 22:439-452.

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.

Griffing, B. 1956. Concept of general and specific combining ability in relation to diallel crossing systems. Aust. J. Biol. Sci. 9:463-493.

Hayman, B.I. 1954. The analysis of variance of diallel tables. Biometrics 10:235-244.

Hühn, M. 1997. Weighted means are unnecessary in cultivar performance trials. Crop Sci. 37:1745–1750.

Jannink, J.-L., A.J. Lorenz, and H. Iwata. 2010. Genomic selection in plant breeding: from theory to practice. Briefings in functional genomics & proteomics 9:166-77.

Kempton, R.A. 1984. The design and analysis of unreplicated field trials. Vortr. Pflanzenzucht. 7:219-242.

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.

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

Piepho, H.-P., A. Büchse, and B. Truberg. 2006. On the use of multiple lattice designs and α-designs in plant breeding trials. Plant Breeding 125:523-528.

Piepho, H.-P., and V. Michel. 2000. Überlegungen zur regionalen Auswertung von Landessortenversuchen. Informatik. Biomet. Epidemiol. Med. Biol. 31:123–136.

Piepho, H.-P., and J. Möhring. 2006. Selection in cultivar trials - Is it ignorable? Crop Sci. 46:192-201.

Piepho, H.-P., and E.R. Williams. 2006. A comparison of experimental designs for selection in breeding trials with nested treatment structure. Theor. Appl. Genet. 113:1505-1513.

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

Ribaut, J.-M., and D.A. Hoisington. 1998. Marker assisted selection: new tools and strategies. Trends Plant Sci.3:236–239.

Schön, C.C., M. Lee, A.E. Melchinger, W.D. Guthrie, and W. Woodman. 1993. Mapping and characterization of quantitative trait loci affecting resistance against second-generation European corn borer in maize with the aid of RFLPs. Heredity 70:648–659.

Schrag, T.A., J. Möhring, A.E. Melchinger, B. Kusterer, B.S. Dhillon, H.-P. Piepho, and M. Frisch. 2010. Prediction of hybrid performance in maize using molecular markers and joint analyses of hybrids and parental inbreds. Theor. Appl. Genet. 120:451-461.

Schnell, F.W. 1982. A synoptic study of the methods and categories of plant breeding. Z. Pflanzenzüchtung 89:1-18.

Smith, A.B., B.R. Cullis, and A.R. Gilmour. 2001. The analysis of crop variety evaluation data in Australia. Aust. N. Z. J. Stat. 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. 143:449-462.

Smith, A.B., P. Lim, and B.R. Cullis. 2006. The design and analysis of multi-phase plant breeding experiments. J. Agric. Sci. 144:393-409.

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.

Williams, E.R., and J.A. John. 2003. A note on the design of unreplicated trials. Biom. J. 45:751-757.

Williams, E.R., H.-P. Piepho, and D. Whitaker. 2010. Augmented p-rep designs. Biometrical Journal (accepted).

Xiang, B., and B.L. Li. 2001. A new mixed analytical method for genetic analysis of diallel data. Can. J. Forest Res. 31:2252-2259.

Yu, J., G. Pressoir, W.H. Briggs, I.V. Bi, M. Yamasaki, J.F. Doebley, M.D. McMullen, B.S. Gaut, D.M. Nielsen, J.B. Holland, S. Kresovich, and E.S. Buckler.

2006. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat. Genet. 38:203-208.

Zhu, J., and B.S. Weir. 1994a. Analysis of cytoplasmic and maternal effects. I. A genetic model for diploid plant seeds and animals. Theor. Appl. Genet. 89:153-159.

Zhu, J., and B.S. Weir. 1994b. Analysis of cytoplasmic and maternal effects. II. Genetic model for triploid endosperms. Theor. Appl. Genet. 89:160-166.

Zhu, J., and B.S. Weir. 1996a. Diallel analysis for sex-linked maternal effects. Theor. Appl. Genet. 92:1-9.

Zhu, J., and B.S. Weir. 1996b. Mixed model approaches for diallel analysis based on a bio-model. Genet. Res. 68:233-240.

Trends in genetic variance components during 30 years of hybrid maize breeding at the University of Hohenheim

Fischer, S., J. Möhring, C.C. Schön, H.-P. Piepho, D. Klein, W. Schipprack, H.F. Utz, A.E. Melchinger, and J.C. Reif.

Plant Breeding 127:446-451

The original publication is available at
http://onlinelibrary.wiley.com/doi/10.1111/j.1439-0523.2007.01475.x/full

Abstract

The efficiency of hybrid performance prediction based on GCA effects depends on the ratio of variances due to specific (SCA) vs. general (GCA) combining ability ($\sigma^2_{SCA}:\sigma^2_{GCA}$). Therefore, we analyse the changes in estimates of σ^2_{GCA}, σ^2_{SCA} and their ratio during 30 years of hybrid maize breeding. The observed trends in genetic variances will be compared to theoretical results expected under a simple genetic model. The analysed multi-location trials were conducted at the University of Hohenheim from 1975 to 2004 and were designed as North Carolina Design II. Grain yield (GY) and dry matter content (DMC) were measured. GY showed a significant ($P < 0.05$) annual increase of 0.17 Mg/ha, but no linear trend was found for DMC, the variances of GCA or SCA and their ratio. The predominance of the sum of estimates of σ^2_{GCA} of the flint and dent heterotic groups compared to their σ^2_{SCA} retained with ongoing inter-population improvement. Consequently, hybrid performance can be predicted on the basis of their parental GCA effects.

Comparison of Weighting in Two-Stage Analysis of Plant Breeding Trials

J. Möhring and H.-P. Piepho

Crop Science 49:1977-1988.

The original publication is available at https://www.crop.scijournals.org

Abstract

Series of plant breeding trials are often unbalanced with a complex genetic structure. It is common practice to employ a two-stage approach, where adjusted means per location are estimated and then a mixed model analysis of these adjusted means is performed. This reduces computing costs. An important question is how means from the first step should be weighted in the second step. We therefore compare different weighting methods in the analysis of four typical series of plant breeding trials using mixed models with fixed or random genetic effects. We used four published weighting methods and proposed three new methods. For comparing methods we used the one-stage analysis as benchmark and computed four evaluation criteria. With the assumption of fixed genetic effects we found that the two-stage analysis gave acceptable results. The same is found in three of four datasets when genetic effects were taken as random in stage two. In both cases differences between weighting methods were small and the best weighting method depended on the dataset but not on the evaluation criteria. A two-stage analysis without weighting also produced acceptable results, but weighting mostly performed better. In the fourth dataset the missing data pattern was informative, resulting in violation of the missing-at-random (MAR) assumption in one- and two-stage analysis. In this case both analyses were not strictly valid.

REML-based diallel analysis

J. Möhring, A. E. Melchinger, and H.-P. Piepho

Crop Science 51:470-478.

The original publication is available at http://crop.scijournals.org

Abstract

Depending on the model and the experimental design diallel analyses can give different results. A unified framework to fit and compare different models by mixed-model packages is lacking. We, therefore, present a general diallel model and the required additional restrictions for the genetic variance–covariance structure to become equivalent to other commonly used diallel models. We discuss the definitions and requirements of the commonly used models compared with the general model. To exemplify your comparison we provide analysis of three real datasets. We observed biased variance-component estimates for general and specific combining ability if the assumptions regarding the genetic variance–covariance structure are not fulfilled and random genetic effects are assumed. We give detailed program codes in statistical software packages SAS and ASReml for the implementation of the general diallel model. If no a priori information about the genetic model for genotypes is available, the general model can be used to analyze all diallel designs easily with standard statistical software.

Impact of Genetic Divergence on the Ratio of Variance Due to Specific vs. General Combining Ability in Winter Triticale

S. Fischer, J. Möhring, H.P. Maurer, H.-P. Piepho, E.-M. Thiemt, C.C. Schön, A.E. Melchinger, and J.C. Reif

Crop Science 49:2119-2122.

The original publication is available at http://crop.scijournals.org

Abstract

This paper examine the influence of genetic divergence on the ratio of the components of variance for specific (σ^2_{SCA}) and general (σ^2_{GCA}) combining ability using experimental data in triticale (×*Triticosecale* Wittm.). In total, 21 lines and their 210 crosses were evaluated for grain yield in field trials. The re-analysis of published molecular data indicate an optimum of two genetically distinct subgroups. Estimates of σ^2_{SCA} and σ^2_{GCA} were determined either for the total diallel or between and within the two subgroups. The ratio of σ^2_{SCA} vs. σ^2_{GCA} tended to be lower for crosses between than within the subgroups. This can be interpreted as an indicator of a more favourable ratio of σ^2_{SCA} vs. σ^2_{GCA} in situations with two genetically distinct subgroups than in situations without genetically distinct subgroups.

6. General Discussion

A time- and cost-efficient breeding strategy often requires the analysis of large series of field experiments that generate phenotypic data. Knowledge of the experimental design, the mating design, the aim of the analysis and requirements of statistical analysis are important for finding an adequate model for analysis. This thesis demonstrates the use of a mixed model framework for solving several problems in analysing phenotypic data collected during plant breeding programs.

Fixed or random genetic effects

If genotypes are taken as fixed, best linear unbiased estimators (BLUEs) based on generalized least squares are used to estimate genotype means. Best linear unbiased predictors (BLUPs; Henderson, 1984; Searle et al., 1992) are used if genetic effects are taken as random. Whether effects for genotypes are taken as random or as fixed depends on the aim of the analysis (Smith et al., 2001a, 2005) and on the way the genotypes have been generated. If interest is in estimating genotype means, genotypes are taken as fixed. If the focus is on predicting the potential breeding value of genotypes in future experiments and genotypes can be regarded as randomly drawn from a base population, genotype effects are taken as random (Henderson, 1984). In plant breeding the prediction of breeding values is of interest, but due to selection, a base population in which idealized conditions hold, such as random mating, linkage equilibrium and lack of inbreeding, does not exist (Piepho et al., 2008). Up to now, plant breeders have often treated genotypes as a fixed factor, ignoring all covariances between genotypes coming from ancestry or evaluation process. Taking genotypic effects as random has the potential advantage that additional correlated information from relatives can be exploited, e.g. by using the numerator relationship matrix (Henderson, 1984; Bernardo, 1994) or by modelling nested and crossed genetic effects representing

the pedigree (Gallais, 1980; Piepho and Williams, 2006; Piepho et al., 2008). Alternatively, the relationship of genotypes can be estimated through marker information (Bernardo, 1993). Furthermore, a separation of genetic effects into additive and non-additive effects is possible (Bernardo, 2002; Piepho and Möhring, 2010). The disadvantage of taking the genetic effect as random is the requirement of estimating a variance component. If there is little information for estimation the variance component, both the variance component estimate and the BLUPs are uncertain. Thus, Searle et al. (1992) proposed to consider effects as random, if the number of genotypes is large. Van Eeuwijk (1995) suggested to have at least ten degrees of freedom for estimating variance components. In this thesis the genotype effect is taken as random. In addition, in Chapter 3 the genotype effect is taken as fixed for comparison. Taking genotypes as random is in accordance with Robinson (1991) and Piepho and Williams (2006), as well as empirical results showing that BLUP is preferable to BLUE (Hill and Rosenberger, 1985; Kleinknecht et al., 2010).

Missing data

During selection, newly created genotypes are added while culled genotypes are discarded, thus plant breeding data are almost always selected and unbalanced. This results in missing data, which complicates analysis, for example in the estimation of heritability (Piepho and Möhring, 2007).

Little and Rubin (2002) distinguish three kinds of missing data patterns: informative missing or missing not at random (MNAR), missing at random (MAR) and missing completely at random (MCAR). Their definition is based on considering all observations from one subject, where a subject refers to a group of correlated data points in repeated measures designs. In medical applications with these types of design, a subject corresponds to a patient/person. Data points of

different subjects are uncorrelated by definition. Another frequent application of repeated measures designs is in animal sciences, where subjects are animals with repeated measures (Littell et al., 1998). In these cases all observations from one patient or animal are correlated, so they belong to the same subject. In plant breeding all genotypes and thus all data belong to the same subject because of correlations between genotypes (coancestry) and correlation because of testing them in the same year, location or block.

The data from one subject can be subdivided into observed and missing data. If a missing data pattern depends on observed data, but not on missing data, the missing data pattern is MAR. If it depends on both observed and missing data, it is informative. If it is independent of both observed and unobserved data, it is MCAR. MCAR and, with the additional assumption of seperability, the MAR pattern is ignorable if REML is used (Verbeke and Molenberghs, 2000). In plant breeding, the missing data pattern is often informative, due to missing information for selection decisions or missing pedigree information. Breeders often use pedigree information during designing their experiments. It is common that genotypes from the same cross are tested within the same trial, often side by side. If trials for a set of genotypes were not performed in every location, pedigree information influences the missing data pattern. Piepho and Möhring (2006) showed that missing data due to selection can be ignored, if all data used for selection are available and are included in the analysis.

The analyses in this thesis demonstrated several ways for including available additional information, e.g. the *per se* performance of hybrid parents for selecting rape seed cultivars (Chapter 3). It was further shown, that if pedigree information for analysis is missing, but was used for the decision of testing or non-testing genotypes in environments, the commonly used model is invalid (Chapter 3).

Alternative genotype-environment variance-covariance structures

The multi-environmental trial (MET) data from plant breeding programmes in this thesis were analysed by a linear mixed model with additive main effects for genotype and environment and additive genotype-environment interaction effects. If genotypes are assumed as random, as it is the case in this thesis, the genotype-environment interaction is random, too. Assuming that genotypes *i* are sorted within environments *j*, the variance-covariance matrix of the genotype-environment mean y_{ij} is

$$\mathrm{var}(\mathbf{y}) = \mathbf{I}_e \otimes \mathbf{I}_g \sigma^2_{ge} + \mathbf{J}_e \otimes \mathbf{I}_g \sigma^2_g = \left(\mathbf{I}_e \sigma^2_{ge} + \mathbf{J}_e \sigma^2_g\right) \otimes \mathbf{I}_g ,$$

where $\mathbf{I}_n$ is a identity matrix of size n, $\mathbf{J}_n$ is an $n \times n$ matrix with all elements equal to unity, $\mathbf{y}$ is the vector of all y_{ij}, n, g and e are the numbers of observations, environments and genotypes, respectively, and σ^2_g and σ^2_{ge} are the genotype and genotype-environment interaction variances, respectively. This model corresponds to a compound symmetry structure for the environment-within-genotype effects and is the simplest structure.

Finlay and Wilkinson (1963) suggested using a product of a sensity parameter for each genotype and the environment main effect as interaction effect in addition to additive main effects. The sensity parameters are slopes for a regression on environmental means. The general form of this linear-bilinear or multiplicative model with two general factors was proposed by Mandel (1961). Thus the factors can be interchanged resulting in a regression on genotypic means (Piepho, 1999). An extension of the environmental effect (Cornelius, 1978) in the Finlay-Wilkinson model resulted in the very popular additive main effect and multiplicative interaction (AMMI) model (Gollob, 1968; Gabriel, 1978; Zobel et al., 1988; Gauch,

1988) with one principal component. Models based on principal components were first introduced by Mandel (1971). Variants of this model can be formulated by dropping the additive environmental main effect (GGE or GREG) (Cornelius et al., 1996), dropping the genotype main effect (site regression, SREG) or dropping both plus the intercept and adding a so-called shift parameter (shifted multiplicative model, SHMM; Cornelius et al., 1992; Piepho 1998). Gogel (1995) proposed a mixed-model analogue for the Finlay-Wilkinson model, while Piepho (1997) proposed the same for AMMI using a factor-analytic variance-covariance structure for random environments and random genotype-environment interaction effects. In Piepho (1998) and Smith et al. (2001b), mixed model analogues to SREG, GREG and AMMI with random genotype effect are given. For the latter case the variance-covariance for $\mathbf{y}$ can be described by

$$\mathrm{var}(\mathbf{y}) = \left(\mathbf{\Lambda}_e \mathbf{\Lambda}_e' + \mathbf{\Psi}_e\right) \otimes \mathbf{I}_g \sigma^2_{ge},$$

where $\mathbf{\Lambda}_e$ is, in the simplest case of one factor, a vector of e environmental sensitivities, $\mathbf{\Psi}_e$ is a diagonal matrix of size e and $\mathbf{I}_g$ is a identity matrix of size g. For factor-analytic models with more than one factor, $\mathbf{\Lambda}_e$ is a matrix with one column for each factor. Factor-analytic structures give more flexibility for estimating heterogeneity in the variance-covariance matrix, especially if the number of factors is increased. In case the number of genotypes exceeds that of environments, the maximum number of factors is equal the number of environments, in this case the number of required parameters is $\frac{e^2 + e}{2}$, where e is the number of environments, which is identical to an unstructured variance-covariance matrix and just a different parametrization, otherwise the number of factors is limited by the number of genotypes. The advantage of more flexible variance-covariance structures for genotype-environment interaction in analysing METs was demonstrated with independent genotypes (Piepho, 1997, 1998; Smith

et al., 2001b) and related genotypes (Crossa et al., 2006; Kelly et al., 2007). Piepho (1998) and So and Edwards (2009) proposed to use a model selection approach for evaluating which variance-covariance structure is best. The disadvantage of more complex genotype-environment structures is the requirement of more variance component estimates. For a larger number of environments the increasing number of required variance component estimates can result in convergence problems (Welham et al., 2010), loss of efficiency and increased computational demands. In the analysis of the four large series in Chapter 3, it was not possible to fit factor-analytic structures or other more flexible structures for the genotype-environment interaction effects, even with a reduced genetic model. So essentially, for each genetic effect a random main effect and a random genotype-environment interaction effect were fitted, which corresponds to a compound symmetry structure. Further simplifications of this commonly used compound symmetry structure are required, if the model otherwise leads to convergence problems. This can happen, e.g. if a numerator relationship matrix for a large number of genotypes is used within the genotype-environment structure. To avoid convergence problems, the genetic effect is then simplified by ignoring the covariances within the numerator relationship matrix. Crossa et al. (2006) showed that using a factor-analytic genotype-environment structure with nine or two factors just slightly influenced the BLUPs of genotype main effects, even if the model fit improves significantly and the standard errors decrease. This suggests that even where factor-analytic models can be fitted, the change in BLUPs of main effects compared to the CS model would also be minor.

An alternative to changing the variance-covariance structure of genotype-environment interaction is the attempt to explain the genotype-environment interaction effects by a regression on covariables for each location (Denis, 1988; Baril et al., 1995; van Eeuwijk, 1996; Crossa, 1999). Using covariables has the advantage of easy biological interpretation of genotype-environment effects. In

analysing variety trials it was shown, however, that the environmental covariables often explain only small parts of genotype-environment interaction variance in large series (Piepho et al., 1998). For MET data, Vargas et al. (1999) show that the bilinear part of the genotype-environment interaction can be related to environmental covariables. Van Eeuwijk et al. (2005) proposed to combine statistical models for the genotype-environment interaction and models with covariables. For the datasets analysed in this thesis, no environmental covariables were available.

Alternative error structures

For all analysis in this thesis a heterogeneous residual variance with an environment-specific residual variance was found by the likelihood-ratio test (Wolfinger, 1993) to be superior to a homogeneous residual variance. We therefore used the approximation of Stram and Lee (1994) due to its simplicity, even if the requirement of a large number of independent subjects is often not given in plant breeding data and more complex methods are preferable (Crainiceanu and Ruppert, 2004). No correlation between residual errors within an environment is assumed. This result is in accordance with So and Edwards (2009). Piepho and Michel (2000) proposed to use the reciprocal values of the variance of adjusted means per environment as weights in calculating genotype means in a MET analysis. These weights imply different residual variances for each environment, which agrees with the models used in this thesis.

An alternative error structure assumes a spatial correlation between residuals within an environment. Observations on plots with short distance between them are more similar than observations with large distance between them. So the residual variance of the difference of two observations is a function of the distance between them. If the coordinates for each observation are available, a range of

spatial models can be fitted, e.g. autoregressive (Gilmour, 1997) or linear (Piepho et al., 2008) variance structures. In general, a spatial error structure can replace the independent error structure after (Piepho et al., 2008) or before (Gilmour et al., 1997) fitting randomisation-based effects like row, column or block effects. Additionally, both structures can be used simultaneously. In analysing plant breeding data, Müller et al. (2010a) found that on the one hand often some spatial models slightly outperformed a model with independent errors (baseline model) in terms of model fit, but on the other hand the baseline model with no spatial terms was most often the best model. Hu and Spilke (2009) concluded that the spatial model has to be defined separately for each environment. In their analysis only one trial per environment was present, so they defined spatial error structures separately for each trial. The optimal use of spatial residual variance structures in routine MET analysis is not straightforward and a model selection approach to find the best spatial model is required. For the analysed data in this thesis most often spatial information was lacking. Additionally, the number of trials within the analysis was generally large (up to several hundred trials), requiring a large number of variance component estimates and a large number of analyses to find the optimal model for each trial.

Besides these problems, the commonly practiced incomplete randomization of genotypes causes some additional problems. Plant breeders often randomize genotypes in two stages. First they randomize the crosses (if at all) and later they randomize genotypes within crosses. The reason for this procedure is the interest of breeders to select crosses as well as genotypes within crosses for highly heritable traits, e.g. resistances directly in the field. Additionally this arrangement helps breeders to get visual information about the variability of a trait within a cross. Therefore testing genotypes from the same cross in spatial neighbourhood simplifies field selection, but at the same time results in restricted randomization. As a result, correlation due to genetic similarity tends to decrease with spatial

distance, because the larger the spatial distance the more likely it is that the concerned genotypes belong to different crosses. The main problem is that spatial models for the residual also assume that correlation decays with spatial distance. Thus, if spatial error models are used for data that are randomized with the restrictions discussed above, the spatial error potentially explains both, the genetic correlation and the spatial correlation. In other words, the spatial and genetic correlation structures are confounded. A separation is not strictly possible. If genetic variance is captured in the residual error term, as a result of confounding, then genotype effect estimates (BLUPs) may be over-shrunken towards the general mean. The amount of confounding and the effect of analysing incompletely randomized data with spatial models are up to now not known. The effect can either be evaluated by overlaying simulated genotype effects onto uniformity trial data using incomplete and complete randomization and analysing the whole data structure or by full Monte Carlo simulation. The former type of analysis is currently underway (Bettina Müller, personal communication).

Relative merits of two-stage analysis depend on aim of analysis

Chapter 3 and Welham et al. (2010) both compare two-stage analysis with different weighting methods with the single-stage analysis. In Chapter 3 just small differences in estimated genotype main effects are found even by an analysis with incorrect weighting, e.g. by using no weights explicitly. The reason for this observation was the generally large number of environments per genotype resulting in an averaging out of the effect of incorrect weights. In contrast, Welham et al. (2010) found a preference of weighting and larger differences between single-stage and two-stage analysis. Besides the assumed spatial correlation of errors, which resulted in more heterogeneous weights within a location, the main reason for these, at first sight, contrasting results is the focus on estimating genotype effects in specific environments in Welham et al. (2010) as

compared to the focus on main genotype main effects in Chapter 3. Depending on the aim of the analysis the influence of different weighting varies. Crossa et al. (2006) observed the same for the variance-covariance structure of the genotype-environment interaction. Comparable to Chapter 3 they calculated the correlation between BLUPs for genotype main effects with different models. Through they observed differences in the standard error and differences in the model fit, they observed high correlations, indicating close similarities between these models in terms of the BLUPs. Again, if there are a large number of environments, the influence of genotype-environment interaction effects on the genotype main effect is averaged out. Therefore the choice of the best genotype-environment interaction variance-covariance structure and the correct residual variance model is less important for analysis aiming at estimation of genotype main effects than for analysis aiming at estimating genotype effects in a specific environment, or for calculating the stability of genotypes (Eberhart and Russell, 1966; Shukla, 1972; Piepho, 1992). In this thesis we always concentrated on genotype main effects, thus the influence of the genotype-environment variance-covariance structure or the residual variance-covariance structure was less critical. Also, it turned out to be difficult to fit more complex models. We therefore chose the simplest genotype-environment variance-covariance structure.

Alternative estimation methods

Historically, ANOVA-based methods for analysing MET (Fisher, 1921) or data from diallel mating designs were used (Griffing, 1956a; Eberhart and Gardner, 1966). REML (Patterson and Thompson, 1971) and ANOVA methods yield identical results for balanced, if simple variance component models are used and all ANOVA estimates are non-negative (Searle et al., 1992). REML is the state-of-the-art method in analysing animal breeding data (Hudson and van Vleck, 1982; Dong and van Vleck, 1989; Meyer and Smith, 1996) and plant breeding data

(Piepho, 2008). Another estimation method is MINQUE (Rao, 1970, 1971), which is equivalent to just the first iteration of REML (Searle et al., 1992). MINQUE has been proposed for analysing complex diallels (Zhu and Weir, 1994a, b, 1996a, b). All these methods are based on a frequentistic view of statistical probability. In this thesis REML was used for all analyses. For diallel analysis including parents, thus for models with more complex variance-covariance structure, REML and ANOVA methods result in different variance component estimates as shown in Chapter 4.

In contrast to the frequentistic view taken in this thesis, the Bayesian approach uses a prior distribution for estimable parameters. This prior distribution is updated with the distribution of actual data to obtain the posterior distribution (Gelman et al., 2000). If good prior information is available, the Bayesian approach can give better variance component estimates and therefore reduced the uncertainty of empirical BLUP. Edwards and Jannink (2006) proposed to use the Bayesian approach for modelling heterogeneous error and genotype-environment interactions. The Bayesian approach was used for modelling animal data (Gianola and Fernando, 1986) and MET data from variety testing (Theobald et al., 2002). The Bayesian approach has also been used for plant breeding data in order to exploit marker information (Bauer et al., 2009).

The Bayesian approach will be preferable, if the amount of information for a variance component, which can be extracted directly from the data, is limited. In plant breeding one possible application is the estimation of cross-specific genetic variances, when the number of genotypes per cross is limited and smaller than required for assuming a random effect when frequentist methods are used for analysis. The Bayesian view also relaxes the division between random or fixed effects, so what is regarded as fixed effect in a frequentist setting can be seen from a Bayesian perspective as a random variable on which prior knowledge is diffuse

or vague (Gianola and Fernando, 1986). The disadvantage of the Bayesian approach is the influence of the prior distribution on the results of the analysis (Koehler, 1993), when the amount of experimental data is limited, which makes a priori expected results more likely and unexpected results less likely. The expected benefit of the Bayesian approach for large series of trials with a large amount of information for estimating variance component estimates can be assumed to be limited.

Use of correlated information

If available, the analyses in this thesis included pedigree information, and thus used the information of correlated genotypes to achieve better estimates for genotype effects. The pedigree information can be used via the numerator relationship matrix (Bernardo, 1993) or by modelling nested and crossed genetic effects representing the pedigree (Gallais, 1980; Piepho and Williams, 2006; Piepho et al., 2008). For simple structures of coancestry, a model with nested and crossed genetic effects has the advantage that the covariance between ancestors is estimated directly within the model. By contrast, if the numerator relationship matrix is used, e.g. the covariance between fullsibs is fixed to half of the genetic variance. Thus, using nested and crossed genetic effects allows a more flexible variance-covariance structure to be fitted. For complex structures, however, the number of required variance component estimates increases, which makes the analysis complex and potentially inefficient. In both cases genotype effects are assumed to be random.

For constructing the numerator relationship matrix we used the coancestry determined by the pedigree. Alternatively marker information can be used to estimate the "realized" coancestry between genotypes (Bernardo, 1994). Marker information can potentially better reflect sampling effects and effects from

selection and drift (Piepho et al., 2008). Bernardo (1994) shows that the marker-based numerator relationship matrix can perform better in plant breeding data under selection. For the datasets analysis in this thesis, the use of the pedigree-based or marker based numerator relationship matrix or both resulted in better model fits than for models without using the numerator relationship matrix. This is in agreement with results in Bernardo (1994), Xu and Virmani (2000), Davik and Honne (2005), Bauer (2006), and Bauer et al. (2006) for additive genotype-environment interaction and Crossa et al. (2006) and Kelly et al. (2009) for multiplicative models. In addition to the numerator relationship matrix, Bernardo (1993) proposed to fit a dominance relationship matrix (D-matrix), which is also used in Flachenecker et al. (2006).

By either the use of correlated information via the numerator relationship matrix or by nested and crossed genetic effects, it is possible to exploit the correlation between genotypes and their parents or ancestors for enhancing the estimation of genotype effects. This information is often used implicitly in practical breeding work, e.g., by looking at the whole cross to evaluate a single genotype. Plant breeders often design field trials such that related genotypes are tested close to each other, so that a selecting between crosses and within crosses directly on the field is possible. But this information is rarely included directly in a mixed model framework even though normally the parents are carefully selected and information about the *per se* performance is available prior to the analysis. Also, information about correlated relatives in previous generations is often available but rarely used. Piepho and Möhring (2010) used different selfing generations to separately estimate additive and dominance variance. For diallel crosses, Zhu and Weir (1994a, b, 1996a, b) proposed models for using additional backcross generations or the groups of male and female offspring. The reason for not using this information lies in technical difficulties to include the often selected and unbalanced data from previous years or in genetical difficulties to include a

different base population or genotypes with different degrees of inbreeding. For the special case of a diallel analysis, Griffing (1956b), Wricke and Weber (1986) and Zhang and Kang (1997) stated that estimating variance components including parents is not useful. Wright (1985) proposed the use of parents, if the reference population is descendant, while if the reference population is ancestral, parents should not be used. In contrast, Curnow (1980) stated that using two types of information, i.e., the performance as parent and the performance as parent in a cross has the potential of increasing selection gain. He also discussed the possibility of weighting both sources of information in a selection index. Using a selection index is essentially identical to using BLUP (Piepho et al., 2008), when assuming a joint variance-covariance structure for genetic effects of parents and genotypes. Chapter 4 proposed a method for using correlated information like *per se* performance within a diallel and showed the potential advantages of this method. But this method is not restricted to parents in diallels. Schrag et al. (2009) used a comparable model with genotypes and the *per se* performance of their parents in a series of partial factorials in maize. Chapter 5 combines the analysis of diallels and factorials using the joint variance-covariance structure. The analysis of the diallel mating design with 21 parents in triticale was either performed by a diallel assuming one population or by two diallels (within each population) and a factorial (between populations) assuming two populations. Further examples using correlated information were given in Piepho and Möhring (2006, 2010) and Falke et al. (2010). The variance-covariance structure can be used to estimate the genetic correlation between successive breeding stages, e.g. to evaluate an early selection.

For the special case of analysing diallels this thesis presents a unified framework for analysing data with a joint variance-covariance structure for effects of parents and crosses. It is shown that estimating the correlation between genotypes and their parents can be superior to assuming a fixed ratio of variance components and

hence a fixed correlation, which can be derived from theoretical genetic models. For using correlated information the important point is a careful separation of different genetic effects and a careful modelling of correlated and non-correlated effects. The same methods for describing the pedigree or for using correlated information that were used in this thesis for the mere purpose of estimating genetic effects, can also be used in mixed modelling for QTL detection by linkage mapping (Schrag et al., 2010) or association mapping (Yu et al., 2006; Crossa et al., 2007; Stich et al., 2008; Müller et al., 2010b).

Alternative field trial designs

In plant breeding trials of many German breeders, the genotypes are typically separated in subgroups of 25 up to 100 genotypes. This is done, e.g., for ease of handling large numbers of genotypes or for faster analyses per subgroup. The genotypes of each subgroup are tested together in one design, mostly a block design or a lattice design. Normally, in these trials genotypes are tested cross by cross, so related genotypes from the same cross are more likely to be in the same subgroup and more likely to be neighbours than genotypes from different crosses. This type of subgrouping of genotypes into trials results in problems with models assuming spatially correlated errors and problems with the missing data pattern, if the information for grouping the genotypes is missing. Furthermore, breeders are interested in selecting the best genotypes overall, not just the best genotypes within a subgroup. Thus, a comparison of all pairs of genotypes is needed. To compare genotypes coming from different subgroups, and hence coming from different trials, additional checks are required. Piepho et al. (2006) showed that using a single α-design that includes all or several crosses outperforms a subgrouping of crosses into several designs, mainly through a reduced number of required replicated check plots and more homogeneous standard errors of pairwise differences. Alternatively, replicated checks can be replaced by

genotypes of interest, thus a part of the genotypes is replicated more often than the rest. This idea is called partially replicated (p-rep) design (Cullis et al., 2006; Smith et al., 2006; Williams et al., 2010). The p-rep design is preferable especially in trials with no particular interest in check results themselves.

Genetic model, additive and non-additive effects

The diallel analysis proposed by Griffing (1956a), Hayman (1954) or Gardner and Eberhart (1966) are based on a Mendelian model introduced in Fisher (1918), which assumes the inheritance of several independent genes, so-called Mendelian factors. The sum of these factors is defined as additive effect, while interactions between genes are modelled by non-additive effects like dominance and epistasis. Thus, a distinction of additive and non-additive effects is possible, which is very attractive for interpretation. But these genetic models are based on strong assumptions (Hayman, 1954) like diploid segregation, which results in the restriction of a fixed ratio of variance component estimates. Often, the strong restrictions are not realistic (Gilbert, 1958). The same holds for using the numerator relationship matrix. Again, fixed ratio between variance component estimates and covariance estimates within the pedigree are assumed and a distinction between additive and non-additive effects is possible. Alternatively, the variance components and covariances are directly estimated, either by using nested and crossed genetic effects or by multivariate analyses. For diallels it was shown in Chapter 4, that if no information about the true model is available, a general model with a more flexible variance-covariance structure, where variances and covariances are directly estimated from the data, can be preferable. This flexible model can be converted to the genetic models by adding appropriate restrictions to the general model and thus still allows the interpretation of genetic effects as additive and non-additive effects.

Conclusion

The mixed model framework is a powerful tool for analysing phenotypic plant breeding data, either for selection or as essential part in genetic studies like QTL detection, linkage studies, association studies or genomic selection. This thesis demonstrated the use of mixed models for analysing large series of plant breeding trials with factorial or diallel mating designs. The often-used stage-wise analysis of large series of plant breeding trials is found to give acceptable results. To avoid informative missing data and to get more precise genotype main effect estimates, correlated information is included via the numerator relationship matrix, nested and crossed genetic effects or a joint variance-covariance structure. The estimated variance component estimates can be used to evaluate changes in the population of cultivars through recurrent selection or to evaluate the optimal grouping of genotypes into heterotic pools for hybrid breeding. For diallels it is shown that the proposed mixed model including crosses and their parents generalized, and can outperform, other diallel models proposed in the literature. Thus, the thesis demonstrates that a valid and efficient analysis of all available phenotypic data is an essential part of many plant breeding processes.

References

Baril, C.P., J.B. Denis, R. Wustman, and F.A. van Eeuwijk. 1995. Analysing genotype by environment interaction in Dutch potato variety trials using factorial regression. Euphytica 82:149-155.

Bauer, A.M., T.C. Reetz, and J. Léon. 2006. Estimation of breeding values of inbred lines using best linear unbiased prediction (BLUP) and genetic similarities. Crop Sci. 46:2685-2691.

Bauer, A.M. 2006. BLUP-Zuchtwertschätzung bei selbstbefruchtenden Getreidearten unter Berücksichtigung aller Verwandtschaftsinformationen und der Inzuchtverhältnisse. Diss. Uni Bonn, Shaker Verlag.

Bauer, A.M., T.C. Reetz, F. Hoti, W.-D. Schuh, J. Léon, and M.J. Sillanpää. 2009. Bayesian prediction of breeding values by accounting for genotype-by-environment interaction in self-pollinating crops. Genet. Res. 91:193-207.

Bernardo, R. 1993. Estimation of breeding values of inbred lines using best linear unbiased prediction (BLUP) and genetic similarities. Crop Sci. 46:2685-2691.

Bernardo, R. 1994. Prediction of maize single-cross performance using RFLPs and information from related hybrids. Crop Sci. 34:20-25.

Bernardo, R. 2002. Breeding for quantitative traits in plants. Stemma, Woodbury, Minn.

Cornelius, P.L. 1978. Empirical models for the analysis of unreplicated lattice-split-plot cultivar trials. Crop Sci. 18:627-633.

Cornelius, P.L., M. Seyedsadr, and J. Crossa. 1992. Using the shifted multiplicative model to search for “separability” in crop cultivar trials. Theor. Appl. Genet. 84:161-172.

Cornelius, P.L., J. Crossa, J. Seyedsadr, M.S. Kang, and H.G. Gauch. 1996. Statistical tests and estimators of multiplicative models for genotype-by-environment cultivar trials. Crop Sci. 46:1722-1733.

Crainiceanu, C.M., and D. Ruppert. 2004. Likelihood ratio tests in linear mixed models with one variance component. J. R. Statist. Soc. B 66:165-185.

Crossa, J., J. Burgueno, P.L. Cornelius, G. McLaren, R. Trethowan, and A. Krishnmachri. 2006. Modeling genotype × environment interaction using additive genetic covariances of relatives for predicting breeding values of wheat genotypes. Crop Sci. 46:1722-1733.

Crossa, J., M. Vargas, F.A. van Eeuwijk, C. Jiang, G.O. Edmeades, and D. Hoisington. 1999. Interpreting genotype × environment interaction in tropical maize using linked molecular markers and environmental covariables. Theor. Appl. Genet. 99:611-625.

Crossa, J., J. Burgueno, S. Dreissigacker, M. Vargas, S.A. Herrera-Foessel, M. Lillemo, R.P. Singh, R. Trethowan, M. Warburton, J. Franco, M. Reynolds, J.H. Crouch, and R. Ortiz. 2007. Association analysis of historical bread wheat germplasm using additive genetic covariance of relatives and population structure. Genetics 177:1889-1913.

Cullis, B.R., A.B. Smith, and N.E. Coombes. 2006. On the design of early generation variety trials with correlated data. J. Agric. Biol. Environ. Stat. 11:381-393.

Curnow, R.N. 1980. Selecting crosses using information from a diallel cross. Biometrics 36:1-8.

Davik, J., and B.I. Honne. 2005. Genetic variance and breeding values for resistance to a wind-borne disease [Sphaerotheca macularis (Wallr. ex Fr.)] in

strawberry (Fragaria × ananassa Duch.) estimated by exploring mixed and spatial models and pedigree information. Theor. Appl. Genet. 111:256-264.

Denis, J.B. 1988. Two-way analysis using covariates. Statistics 19:123-132.

Dong, M.C., and L.D. van Vleck. 1989. Effect of relationship on estimation of variance components with an animal model and restricted maximum likelihood. J. Dairy Sci. 71:3047–3052.

Eberhart, S.A., and W.A. Russell. 1966. Stability parameters for comparing varieties. Crop Sci. 6:36-40.

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

Falke, K.C., P. Wilde, H. Wortmann, B.U. Müller, J. Möhring, H.-P. Piepho, and T. Miedaner. 2010. Correlation between per se and testcross performance in rye (Secale cereale L.) introgression lines estimated with a bivariate mixed linear model. Crop Sci.50:1863-1873.

Finlay, K.W., and G.M. Wilkinson. 1963. The analysis of adaptation in plant breeding programme. Aust. J. Agric. Res. 14:742-757.

Fisher, R.A. 1918. The correlation between relatives on the supposition of Mendelian inheritance. Trans. Roy. Soc. Edinb. 52:399-433.

Fisher, R.A. 1921. Studies in crop variantion. I. An eximination of the yield of dressed grain from Broadbalk. J. Agric. Sci. 11:107-135.

Flachenecker, C., K.C. Falke, M. Frisch, and A.E. Melchinger. 2006. Trends in population parameters and best linear unbiased prediction of progeny performance in a European F_2 maize population under modified recurrent full-sib selection. Theor. Appl. Genet. 112:481-493.

Gabriel, K.R. 1978. Least squares approximation of matrices by additive and multiplicative model. J. R. Statist. S. Ser. B 40:186-196.

Gallais, A. 1980. Is Fisher´s model necessary for the theory of population improvement? Theor. Appl. Genet. 58:177-180.

Gardner, C.O., and A.S. Eberhart. 1966. Analysis and interpretation of the variety cross diallel and related populations. Biometrics 22:439-452.

Gauch, H.G. 1988. Model selection and validation for trials with interaction. Biometrics 44:705-715.

Gelman, A., J.B. Carlin, H.S. Stern, and D.B. Rubin. 1995. Bayesian data analysis. Chapman & Hall, New York.

Gianola, D., and R.L. Fernando. 1986. Bayesian methods in animal breeding theory. J. Anim. Sci. 63:217-244.

Gilbert, N.E.G. 1958. Diallel cross in plant breeding. Heredity 12:477-492.

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.

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

Gollob, H.F. 1968. A statistical model that combines features of factor analysis and analysis of variance techniques. Psychometrika 33:73-115.

Griffing, B. 1956a. Concept of general and specific combining ability in relation to diallel crossing systems. Aust. J. Biol. Sci. 9:463-493.

Griffing, B. 1956b. A generalised treatment of the use of diallel crosses in quantitative inheritance. Heredity 1:31-50.

Hayman, B.I. 1954. The analysis of variance of diallel tables. Biometrics 10:235-244.

Henderson, C.R. 1984. Estimation of variances and covariances under multiple trait models. J. Dairy Sci. 67:1581–1589.

Hill Jr.,R.R., and J.L. Rosenberger. 1985. Methods for combining data from germplasm evaluation trials. Crop Sci. 25:467-470

Hu, X., and J. Spilke. 2009. Comparison of various spatial models for the analysis of cultivar trials. N. Z. J. Agric. Res. 52:277-287.

Hudson, G.F.S., and L.D. van Vleck. 1982. Estimation of components of variance by method 3 and Henderson's new method. J. Dairy Sci. 65:435-441.

Kelly, A.M., A.B. Smith, J.A. Eccleston, and B.R. Cullis. 2007. The accuracy of varietal selection using factor analytic models for multi-environment plant breeding trials. Crop Sci. 47:1063-1070.

Kelly, A.M., B.R. Cullis, A.R. Gilmour, J.A. Ecclestone, and R. Thompson. 2009. Estimation in a multiplicative mixed model involving a genetic relationship matrix. Gen. Sel. Evol. 41:33.

Kleinknecht, K., F. Laidig, H.-P. Piepho, and J. Möhring. 2010. Best Linear Unbiased Prediction (BLUP): Is it beneficial in official variety performance trials? (accepted).

Koehler, J.J. 1993. The influence of prior beliefs on scientific judgments of evidence quality. Organizational Behavior and Human Decision Processes 56:28-55.

Mandel, J. 1961. Non-additivity in two-way analysis of variance. J. Am. Statist. Ass. 56:878-888.

Mandel, J. 1971. A new analysis of variance model for non-additive data. Technometrics 13:1-18.

Meyer, K., and S.P. Smith. 1996. Restricted Maximum Likelihood estimation for an animal model using derivates of the likelihood. Gen. Sel. Evol. 28:23-49.

Littell, R.C., P.R. Henry, and C.B. Ammerman. 1998. Statistical analysis of repeated measures data using SAS procedures. Anim. Sci. 76:1216-1231.

Little, R.J.A., and Rubin D.B. 2002. Statistical analysis with incomplete data. 2nd ed. John Wiley & Sons, New York.

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

Müller, B.U., B. Stich, and H.-P. Piepho. 2010b. A general method for controlling the genome-wide Type I error rate in linkage and association mapping experiments in plants. Heredity 106:825-831.

Patterson, H.D., and R. Thompson. 1971. Recovery of inter-block information when block sizes are unequal. Biometrika 58:545-554.

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

Piepho, H.-P. 1998. An algorithm for fitting the shifted multiplicative model. J. Statist. Comput. Simul. 62:29-43.

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. 1999. Fitting a regression model for genotype-by-environment data on heading dates in grasses by methods for nonlinear mixed models. Biometrics 55:1120-1128.

Piepho, H.-P., and V. Michel. 2000. Überlegungen zur regionalen Auswertung von Landessortenversuchen. Informatik, Biomet. Epidemiol. Med. Biol. 31:123-136.

Piepho, H.-P. 1992. Vergleichende Untersuchungen der statistischen Eigenschaften verschiedener Stabilitätsmaße mit Anwendungen auf Hafer, Winterraps, Ackerbohnen sowie Futter- und Zuckerrüben. Diss., Universität Kiel.

Piepho, H.-P., C. Richter, and E.R. Williams. 2008. Nearest neighbour adjustment and linear variance models in plant breeding trials. Biom. J. 50:164-189.

Piepho, H.-P., and J. Möhring. 2006. Selection in cultivar trials - Is it ignorable? Crop Sci. 46:192-201.

Piepho, H.-P., and J. Möhring. 2007. Computing heritability and selection response from unbalanced plant breeding trials. Genetics 177:1881-1888.

Piepho, H.-P., and E.R. Williams. 2006. A comparison of experimental designs for selection in breeding trials with nested treatment structure. Theor. Appl. Genet. 113:1505-1513.

Piepho, H.-P., J. Möhring, A.E. Melchinger, and A. Büchse. 2008. BLUP for phenotypic selection in plant breeding and variety testing. Euphytica 161:209-228.

Piepho, H.-P., and J. Möhring. 2010.Generation means analysis using mixed models.Crop Sci. 50:1674-1680.

Rao, C.R. 1970. Estimation of heteroscedastic variances in linear models. J. Amer. Stat. Ass. 65:161-172.

Rao, C.R. 1971. Estimation of variance and covariance components-MINQUE theory. Journal of Multivariate Analysis 1:257-275.

Robinson, G.K. 1991. That BLUP is a good thing: the estimation of random effects. Stat.Sci. 6:15-51.

Schrag, T.A., J. Möhring, H.P. Maurer, B.S. Dhillon, A.E. Melchinger, H.-P. Piepho, A.P. Sørensen, and M. Frisch. 2009. Molecular marker-based prediction of hybrid performance in maize using unbalanced data from multiple experiments with factorial crosses. Theor. Appl. Genet. 118:741-751.

Schrag, T.A., J. Möhring , A.E. Melchinger, B. Kusterer, B.S. Dhillon, H.-P. Piepho, and M. Frisch. 2010. Prediction of hybrid performance in maize using molecular markers and joint analyses of hybrids and parental inbreds. Theor. Appl. Genet. 120:451-461.

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

Shukla G.K. 1972. Some statistical aspects of partitioning genotype-environmental components of variability. Heredity 29:237-245.

Smith, A.B., B.R. Cullis, and A.R. Gilmour. 2001a. The analysis of crop variety evaluation data in Australia. Aust. N. Z. J. Stat. 43:129-145.

Smith, A.B., B.R. Cullis, and A.R. Gilmour. 2001b. Analyzing variety by environment trials using multiplicative mixed models and adjustments for spatial field trend. Biometrics 57:1138-1147.

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. 143:449–462.

Smith, A.B., P. Lim, and B.R. Cullis. 2006. The design and analysis of multi-phase plant breeding experiments. J. Agric. Sci. 144:393-409.

So, Y.-S., and J. Edwards. 2009. A comparision of mixed-model analyses of the Iowa crop performance test for corn. Crop Sci. 49:1593-1601.

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.

Stram, D.O., and J.W. Lee. 1994. Variance components testing in the longitudinal mixed effects model. Biometrics 50:1171-1177.

Theobald, C.M., M. Talbot, and F. Nabugoomu. 2002. Bayesian approach to prediction from crop trials. J. Agric. Biol. Environ. Statist. 7:403-419.

Vargas, M., J. Crossa, F.A. van Eeuwijk, M.E. Ramirez, and K. Sayre. 1999. Using AMMI, factorial regression, and partial least squares regression models for interpretating genotype × environment interaction. Crop Sci. 39:955-967.

van Eeuwijk, F.A. 1995. Linear and bilinear models for the analysis of multi-environment trials: I. An inventory of models. Euphytica 84:1-7.

van Eeuwijk, F.A., M. Malosetti, X. Yin, P.C. Struik, and P. Stam. 2005. Statistical models for genotype by environment data: from conventional ANOVA models to eco-physiological QTL-models. Aust. J. Agric. Res. 56:883-894.

van Eeuwijk, F.A., J.B. Denis, and M.S. Kang. 1996. Incorporating additional information on genotypes and environments in models for two-way genotype by environment tables. In Genotype-by-environment interaction. (Eds M.S. Kang, H.G. Gauch), CRC Press, Boca Raton, p. 15-50.

Verbeke, G., and G. Molenberghs. 2000. Linear mixed models for longitudinal data. Springer Verlag, Berlin.

Welham, S.J., B.J. Gogel, A.B. Smith, R. Thompson., and B.R. Cullis. 2010. A comparision of analysis methods for late-stage variety evaluation trials. Aust. N. Z. J. Stat. 52:125-149.

Williams, E.R., H.-P. Piepho, and D. Whitaker. 2010. Augmented p-rep designs. Biom. J. 53:19-27.

Wolfinger, R. 1993. Laplace's approximation for nonlinear mixed models. Biometrika 80:791-795.

Wricke, G., and W.E. Weber. 1986. Quantitative genetics and selection in plant breeding. De Gruyter, Berlin.

Wright, A.J. 1985. Diallel designs, analyses, and reference populations. Heredity 54:307-311.

Xu, W., and S.S.Virmani. 2000. Prediction of hybrid performance in rice: Comparisons among best linear unbiased prediction (BLUP) procedure, midparent value, and molecular marker distance. Int. Rice Res. Notes 25:12-13.

Yu, J., G. Pressoir, W.H. Briggs, I.V. Bi, M. Yamasaki, J.F. Doebley, M.D. McMullen, B.S. Gaut, D.M. Nielsen, J.B. Holland, S. Kresovich, and E.S. Buckler. 2006. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat. Genet. 38:203-208.

Zhang, Y., and M.S. Kang. 1997. DIALLEL-SAS: A SAS program for Griffing's diallel analyses. Agron. J. 89:176-182.

Zhu, J., and B.S. Weir. 1994a. Analysis of cytoplasmic and maternal effects. I. A genetic model for diploid plant seeds and animals. Theor. Appl. Genet. 89:153-159.

Zhu, J., and B.S. Weir. 1994b. Analysis of cytoplasmic and maternal effects. II. Genetic model for triploid endosperms. Theor. Appl. Genet. 89:160-166.

Zhu, J., and B.S. Weir. 1996a. Diallel analysis for sex-linked maternal effects. Theor. Appl. Genet. 92:1-9.

Zhu, J., and B.S. Weir. 1996b. Mixed model approaches for diallel analysis based on a bio-model. Genet. Res. 68:233-240.

Zobel, R.W., M.J. Wright, and H.G. Gauch. 1988. Statistical analysis of a yield trial. Agron. J. 80:388-393.

7. Summary

Phenotypic selection and genetic studies for QTL detection, linkage and association mapping or genomic selection require an efficient and valid analysis of phenotypic plant breeding data. Therefore, the analysis must take the mating design, the field design and the genetic structure of tested genotypes into account. The analysis is often performed in stage-wise fashion by analysing each trial or location separately and estimating adjusted genotype means per trial or location. These means are then submitted to a mixed model to calculate genotype main effects across trials or locations.

In Chapter 2 an analysis of unbalanced multi-environment trials (METs) in maize using a factorial design are performed. The dataset from 30 years is subdivided in periods of up to three years. Variance component estimates for general and specific combining ability are calculated for each period. While mean grain yield increased with ongoing inter-pool selection, no changes for the mean of dry matter yield or for variance component estimate ratios were found. The continous preponderance of general combining ability variance allows a hybrid selection based on general combining effects.

Chapter 3 studies the influence of stage-wise analysis on genotype main effect estimates for models which take account of the typical genetic structure of genotype effects within plant breeding data. For comparison, the genetic effects were assumed both fixed and random. The performance of several weighting methods for the stage-wise analysis are analysed by correlating the two-stage estimates with results of one-stage analysis and by calculating the mean square error (MSE) between both types of estimate. In case of random genetic effects, the genetic structure is modelled in one of three ways, either by using the numerator relationship matrix, a marker-based kinship matrix or by using crossed and nested

genetic effects. It was found that stage-wise analysis results in comparable genotype main effect estimates for all weighting methods and for the assumption of random or fixed genetic effect if the model for analysis is valid. In case of choosing invalid models, e.g., if the missing data pattern is informative, both analyses are invalid and the results can differ. Informative missing data pattern can result from ignoring information either used for selecting the analysed genotypes or for selecting the test environments of genotypes, if not all genotypes are tested in all environments.

While correlated information from relatives is rarely directly used for analysis of plant breeding data, it is often used implicitly by the breeder for selection decisions, e.g. by looking at the performance of a genotye and the avarage performance of the underlying cross. Chapter 4 proposed a model with a joint variance-covariance structure for related genotypes in analysis of diallels. This model is compared to other diallel models based on assumptions regarding the inheritance of several independent genes, i.e. on genetic models with more restrictive assumptions on the relationship between relatives. The proposed diallel model using a joint variance-covariance structure for parents and parental effects in crosses is shown to be a general model subsuming other more specialized diallel models, as these latter models can be obtained from the general model by adding restrictions on the variance-covariance structure. If no a priori information about the genetic model is available, and thus no need for using a genetic model is given, the proposed general model can outperform the more restrictive models. Using restrictive models can result in biased variance component estimates, if restrictions are not fulfilled by the data analysed.

Chapter 5 evaluates, whether a subdivision of 21 triticale genotypes into heterotic pools is preferable. Subdividing genotypes into heterotic pools implies a factorial mating design between heterotic pools and a diallel mating design within each

heterotic pool. Without subdivision the mating design is a diallel, therefore the proposed model in Chapter 4 is used for analysis. For two (or more) heterotic pools the model is extended by assuming a joint variance-covariance structure for parental effects and general combing ability effects within the diallel and within the factorial. It is shown that a model with two heterotic pools shows the best model fit. The variance component estimates for the general combing ability decrease within the heterotic pools and increase between heterotic pools.

The results in Chapter 2 to 5 show, that an efficient and valid analysis of phenotypic plant breeding data is an essential part of the plant breeding process. The analysis can be performed in one or two stages. The used mixed models recognizing the field and mating design and the genetic structure can be used for answering questions about the genetic variance in cultivar populations under selection and of the number of heterotic pools. The proposed general diallel model using a joint variance-covariance structure between related effects can further be modified for factorials and other mating designs with related genotypes.

8. Zusammenfassung

Eine effiziente und valide Auswertung von pflanzenzüchterischen Daten wird für phänotypische Selektion einerseits und genetischen Studien wie QTL-Kartierung, Kopplungs- und Assoziationsstudien sowie genomweiter Selektion anderseits benötigt. Hierfür muss in der Auswertung das Versuchsdesign, das Kreuzungsdesign und die genetische Struktur der zu testenden Genotypen berücksichtigt werden. Die Auswertung erfolgt oft in zwei Stufen. Zunächst werden Mittelwerte pro Versuch oder pro Ort geschätzt. Diese Mittelwerte werden anschließend in einer Serienauswertung verwendet, um genotypische Schätzwerte über die Versuchsserie hinweg zu erhalten.

In Kapitel 2 wird die Auswertung eines 30-jährigen, mehrortigen und unbalancierten Maisdatensatzes mit faktoriellem Kreuzungsdesign durchgeführt. Der Datensatz wird zunächst in bis zu dreijährige Versuchsserien unterteilt. Für diese werden dann Gesamtmittelwerte sowie Varianzkomponenten für generelle und spezifische Kombinationseignung ermittelt und zwischen den Versuchsserien verglichen. Während der Kornertrag mit der Zeit zunimmt, kann für die Trockensubstanzmenge und das Verhältnis der Varianzkomponenten keine Veränderung nachgewiesen werden. Der stets hohe Anteil der allgemeinen Kombinationseignungsvarianz an der gesamten genetischen Varianz erlaubt eine Hybridselektion aufgrund der allgemeinen Kombinationseignung.

Kapitel 3 untersucht den Einfluss einer zweistufigen Auswertung auf genotypische Gesamtmittelwerte für den Fall, dass die für Pflanzenzüchtungsdaten typischen Verwandtschaftsverhältnisse zwischen Genotypen berücksichtigt werden. Hierbei werden Zweischrittauswertungen mit unterschiedlichen Gewichtungsmethoden im zweiten Schritt mit einer Einschrittauswertung verglichen. Die genetischen Effekte werden als zufällig

angenommen, wobei zur Integration der Verwandtschaftsinformation der Genotypen drei Verfahren verwendet werden: Eine abstammungsbasierte Ähnlichkeitsmatrix, eine markerbasierte Ähnlichkeitsmatrix oder ein Modell mit geschachtelten und gekreuzten genetischen Effekten. Zum Vergleich werden die selben Datensätze auch mit festen genetischen Effekten ausgewertet. Als Gütekriterium werden die Korrelation der Gesamtmittelwertschätzungen zu den Schätzwerten der Einschrittauswertung sowie der mittlere quadratische Fehler (MSE) zwischen den Schätzwerten aus Ein- und Zweischrittauswertung bestimmt. Dabei ergeben sich sowohl für die Annahme fester, als auch für die Annahme zufälliger genetischer Effekte vergleichbare Mittelwertschätzwerte für alle Gewichtungsmethoden. Im Fall von nicht zulässigen Modellen, zum Beispiel wenn das Fehlmuster der Daten nicht zufällig ist, ergeben sich Unterschiede zwischen Ein- und Zweischrittauswertung. In dem Fall sind beide Auswertungen nicht zulässig. Informative Fehlmuster können durch fehlende Verwandtschaftsinformationen verursacht werden, wenn diese Information zur Selektion der geprüften Genotypen oder der geprüften Genotyp-Umwelt-Kombinationen genutzt wird.

Während korrelierte Information von Verwandten im Modell für die Auswertung pflanzenzüchterischer Daten selten direkt verwendet wird, nutzen Züchter diese Information oft implizit. So wird zur Bewertung der Leistung eines Genotypen oft auch die Eignung der gesamten Kreuzung betrachtet. In Kapitel 4 wird für die Auswertung eines Diallels ein Modell vorgeschlagen, das eine gemeinsame Varianz-Kovarianzmatrix für alle korrelierten genetischen Effekte verwendet. Im Diallel wird also eine Korrelation zwischen dem Elterneffekt eines Elters und dem generellen Kombinationseignungseffekts des selben Elters modelliert. Dieses Model wird verglichen mit anderen Diallelmodellen, die auf der Vererbung vieler unabhängiger Gene und somit auf restriktiveren Annahmen bezüglich der Varianz-Kovarianzmatrix basieren. Es kann gezeigt werden, dass das

vorgeschlagene Modell eine Verallgemeinerung für die anderen verwendeten Diallelmodelle darstellt und dass sich diese spezielleren Diallelmodelle durch Hinzufügen von Restriktionen in der Varianz-Kovarianzmatrix aus dem vorgeschlagenen Modell ableiten lassen. Fehlen Vorabinformationen, dass das wahre genetische Vererbungsmodell durch die anderen Diallelmodelle besser abbilden wird, so kann das vorgeschlagene Modell Dialleldaten potentiell besser beschreiben. Außerdem können Abweichungen von restriktiveren Varianz-Kovarianzstrukturen zu verzerrten Varianzkomponentenschätzungen führen.

Kapitel 5 untersucht, ob eine Unterteilung von 21 Triticalegenotypen in heterotische Gruppen sinnvoll ist. Eine Unterteilung der Genotypen in heterotische Gruppen impliziert faktorielle Kreuzungsdesigns zwischen den heterotischen Gruppen und diallele Kreuzungsdesigns innerhalb der heterotischen Gruppe. Ohne Unterteilung ist das Kreuzungsdesign ein Diallel und das in Kapitel 4 vorgeschlagene Modell kann genutzt werden. Für zwei oder mehr heterotische Gruppen wird das Modell erweitert, in dem eine gemeinsame Varianz-Kovarianzmatrix für den Eltereffekt und die generelle Kombinationseignung des Elters im Diallel und im faktoriellen Design angenommen wird. Ein Modell mit zwei heterotischen Gruppen zeigt die beste Modellanpassung. Die Varianz der generellen Kombinationseignung schrumpft innerhalb der heterotischen Gruppe und erhöht sich zwischen den heterotischen Gruppen.

Die Ergebnisse in den Kapiteln 2 bis 5 zeigen, dass eine effiziente und valide Auswertung phänotypischer Pflanzenzüchtungsdaten ein essentieller Teil der Pflanzenzüchtung ist. Die Auswertung selbst kann ein oder zweistufig erfolgen. Die gemischten Modelle berücksichtigen das Versuchs- und Kreuzungsdesign und können verwendet werden, um Fragen über die Entwicklung genetischer Varianzen in Züchtungspopulationen oder zur optimalen Anzahl heterotischer

Gruppen zu beantworten. Das vorgeschlagene Diallelmodell mit einer gemeinsamen Varianz-Kovarianzstruktur für alle korrelierten genetischen Effekte lässt sich für faktorielle Designs und andere Kreuzungsdesigns mit korrelierten Genotypen erweitern.

Curriculum vitae

Personal data

Name	Jens Möhring
Adresse	Winterlinger Weg 17, 70567 Stuttgart, Germany
Born	Hattingen, 30.06.78
Phone	0711/7824319
e-mail	moehring@uni-hohenheim.de

Education

1984-1989	Elementary school (Grundschule Niedersprockhövel)
1989-1998	High school (Gymnasium im Schulzentrum Holthausen (Hattingen)) Final Examination (Grade: 1.7)
10/1998-09/1999	Studies of Agricultural Biology at the University of
10/2000-07/2003	Hohenheim, Stuttgart, Germany; focus on biometry and population genetics. University degree: Diplom-Agrarbiologe (Grade: 1.2)
10/2004-09/2005	Studies of Informatics at the University of Stuttgart, Germany
07/2005-03/2011	PhD student, Institute of Crop Science, Bioinformatics Unit (Prof. Dr. H.-P. Piepho), University of Hohenheim, Stuttgart, Germany.

Professional Experience

08/1999-06/2000	Civil service and Internship (Johannesanstalten Mosbach, Branch Schwarzacher Hof, Agricultural Unit)
09/2000	Harvest hand, Max-Planck-Institute for Plant Breeding Research, Cologne, Germany
08/2001	Internship at the Max-Planck-Institute for Plant Breeding Research, Research Group of PD Dr. C. Gebhardt, Cologne,

	Germany
08/2002	Internship at the Institute of Plant Breeding, Seed Science and Population Genetics, University of Hohenheim in Stuttgart, Germany: Calculating the linkage disequilibrium with R
04/2002-03/2003	Teaching assistant, Institute of Applied Mathematics and Statistics, University of Hohenheim, Stuttgart, Germany
11/2002-02/2003	Teaching assistant, Institute of Crop Science, Bioinformatics Unit, University of Hohenheim, Stuttgart, Germany
09/2003-12/2003	Consultancy study: Optimierung des Sortenprüfsystems bei Winterweizen. Commissioned by: Bund Deutscher Pflanzenzüchter, Bundessortenamt and Verband der Landwirtschaftskammern
01/2004-03/2004	Consultancy study: Optimierung des Prüfsystems bei Winterweizen. Commissioned by: Bund Deutscher Pflanzenzüchter, Bundessortenamt and Verband der Landwirtschaftskammern
04/2004-08/2004	Consultancy study: Bereitstellung von Werkzeugen für die „Auswertung von überlappenden Anbaugebieten“ am Beispiel Winterweizen. Commissioned by: Bund Deutscher Pflanzenzüchter and Verband der Landwirtschaftskammern

List of publications (with peer review)

1. Piepho, H.-P., and J. Möhring. 2005. Best linear unbiased prediction in subdivided target regions. Crop Sci. 45:1151-1159.
2. Piepho, H.-P., and J. Möhring. 2006. Selection in cultivar trials- Is it ignorable? Crop Sci. 46:192-201.
3. Piepho, H.-P., and J. Möhring. 2007. Computing heritability and selection response from unbalanced plant breeding trials. Genetics 177:1881-1888.

4. Fischer, S., J. Möhring, C.C. Schön, H.-P. Piepho, D. Klein, W. Schipprack, H.F. Utz, A.E. Melchinger, and J.C. Reif. Trends in genetic variance components during 30 years of hybrid maize breeding at the University of Hohenheim. 2008. Plant Breeding 127:446-451.
5. Falke, K.C., Z. Susic, P. Wilde, H. Wortmann, J. Möhring, H.-P. Piepho, H.H. Geiger, and T. Miedaner. 2008. Testcross performance of rye introgression lines developed by marker-assisted backcrossing using an Iranian accession as donor. Theor. Appl. Genet. 118:1225-1238.
6. Piepho, H.-P., J. Möhring, A.E. Melchinger, and A. Büchse. 2008. BLUP for phenotypic selection in plant breeding and variety testing. Euphytica 161:209-228.
7. Gutjahr, C. M., Weis, M. Sökefeld, C. Ritter, J. Möhring, A. Büchse, and H.-P. Piepho. 2008. Creating decision rules for site-specific weed managment [Erarbeitung von Entscheidungsalgorithmen für die teilflächenspezifische Unkrautbekämpfung]. Journal of Plant Diseases and Protection 21:143-148.
8. 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.
9. Stich, B., A.E. Melchinger, M. Heckenberger, J. Möhring, A. Schechert, and H.-P. Piepho. 2008. Association mapping in multiple segregating populations of sugar beet (Beta vulgaris L.). Theor. Appl. Genet. 117:1167-1179.
10. Möhring, J., and H.-P. Piepho. 2009. Comparison of weighting in two-stage analyses of series of experiments. Crop Sci. 49:1977-1988.
11. Fischer, S., J. Möhring, H.P. Maurer, H.-P. Piepho, E.-M. Thiemt, C.C. Schön, A.E. Melchinger, and J.C. Reif. 2009. Impact of genetic divergence on the ratio of variance due to specific vs general combining ability in winter-triticale. Crop Sci. 49:2119-2122.

12. Schrag, T.A., J. Möhring, H.P. Maurer, B.S. Dhillon, A.E. Melchinger, H.-P. Piepho, A.P. Sørensen, and M. Frisch. 2009. Molecular marker-based prediction of hybrid performance in maize using unbalanced data from multiple experiments with factorial crosses. Theor. Appl. Genet. 118:741-751.
13. Möhring, J., A.E. Melchinger, and H.-P. Piepho. 2010. REML based diallel analysis. Crop Sci. 51:470-478.
14. Falke, K.C., P. Wilde, H. Wortmann, B.U. Müller, J. Möhring, H.-P. Piepho, and T. Miedaner. 2010. Combined analysis of per se and testcross performance in rye (Secale cereale L.) introgression lines. Crop Sci. 50:1863-1873.
15. Fischer, S., H.P. Maurer, T. Würschum, J. Möhring, H.-P. Piepho, C.C. Schön, E.-M. Thiemt, B.S. Dhillion, E.A. Weissmann, A.E. Melchinger, and J.C. Reif. 2010. Development of heterotic groups in triticale. Crop Sci. 50:584-590.
16. Fischer, S., A.E. Melchinger, V. Korzun, P. Wilde, B. Schmiedchen, J. Möhring, H.-P. Piepho, B.S. Dhillon, T. Würschum, and J.C. Reif. 2010. Molecular marker assisted broadening of the Central European heterotic groups in rye with Eastern European germplasm. Theor. Appl. Genet. 120:291-299.
17. Gruber, S., A. Bühler, J. Möhring, and W. Claupein. 2010. Sleepers in the soil – Vertical distribution by tillage and long-term survival of oilseed rape seeds compared with plastic pellets. European Journal of Agronomy 33:81-88.
18. Schrag T.A., J. Möhring , A.E. Melchinger, B. Kusterer, B.S. Dhillon, H.-P. Piepho, and M. Frisch. 2010. Prediction of hybrid performance in maize using molecular markers and joint analyses of hybrids and parental inbreds. Theor. Appl. Genet. 120:451-461.

19. Reif, J.C., W. Liu, M. Gowda, H.P. Maurer, J. Möhring, S. Fischer, A. Schechert, and T. Würschum. 2010. Genetic basis of agronomically important traits in sugar beet (Beta vulgaris L.) investigated with joint linkage association mapping. Theor Appl Genet. 121:1489-1499.
20. Will, S., T. Eichert, V. Fernández, J. Möhring, T. Müller, and V. Römheld. 2011. Absorption and mobility of foliar-applied boron in soybean as affected by plant boron status and application as a polyol complex. Plant Soil (DOI: 10.1007/s11104-011-0746-6).
21. Piepho, H.-P., and J. Möhring. 2011. On estimation of genotypic correlations and their standard errors by multivariate REML using the MIXED procedure of the SAS System. Crop Science (accepted).

Reports, presentations and publications without peer review:

1. Möhring, J., A. Büchse, and H.-P. Piepho. 2003. Optimierung des Sortenprüfsystems bei Winterweizen. Bericht für Bund Deutscher Pflanzenzüchter, Bundessortenamt und Verband der Landwirtschaftskammern.
2. Möhring, J., A. Büchse, and H.-P. Piepho. 2004. Optimierung des Prüfsystems bei Winterweizen. Bericht für Bund Deutscher Pflanzenzüchter, Bundessortenamt und Verband der Landwirtschaftskammern.
3. Möhring, J., A. Büchse, and H.-P. Piepho. 2004. Bereitstellung von Werkzeugen für die „Auswertung von überlappenden Anbaugebieten“ am Beispiel Winterweizen. Bericht für Bund Deutscher Pflanzenzüchter und Verband der Landwirtschaftskammern.
4. Möhring, J., A. Büchse, H.-P. Piepho, V. Michel, J. Rath, and F. Laidig. 2004. Gesundsparen ohne Nachteile! DLG-Mitteilungen 6:22-23.
5. Möhring, J., A. Büchse, and H.-P. Piepho. 2005 Auswertung von landwirtschaftlichen Sortenversuchen mit PROC MIXED – Spagat zwischen

Theorie und Praxis. p. 279-188. *In* E. Rödel and R.H. Bödeker (ed.) SAS: Verbindung von Theorie und Praxis. Shaker-Verlag, Aachen.

6. Zenk, A., J. Möhring, and V. Michel. 2005. Einbindung neuer Methoden zur Routineauswertung von landwirtschaftlichen Versuchen mit Hilfe von SAS Makros. p. 407-418. *In* E. Rödel and R.H. Bödeker (ed.) SAS: Verbindung von Theorie und Praxis. Shaker-Verlag, Aachen.
7. Michel, V., A. Zenk, J. Möhring, A. Büchse, and H.-P. Piepho. 2007. Die Hohenheim-Gülzower-Serienauswertung als bundesweites Basisverfahren im regionalisierten Sortenwesen. Beiträge zum Sorten- und Versuchswesen und zur Biostatistik. Mitteilungen aus der Landesforschungsanstalt für Landwirtschaft und Fischerei Mecklenburg-Vorpommern:72-82.
8. Michel, V., A. Zenk, R. Graf, J. Möhring, A. Büchse, and H.-P. Piepho. 2007. The Hohenheim-Gülzow method for analysis of series of trials as basic procedure for PIAF and PIAFStat and SAS in a regionalized trial system. pp. 136-141. *In* H.-P. Piepho and H. Bleiholder Proceedings of the International Symposium Agricultural Field Trials – Today and Tomorrow. Verlag Grauer, Beuren.
9. Piepho, H.-P., and J. Möhring. 2007. On weighting in two-stage analysis of series of experiments. Biuletyn Oceny Odmian 32:109-121.
10. Möhring, J., and H.-P. Piepho. 2008. Comparision of one-stage and two-stage analysis in series of experiments. p. 107-108. *In* L.A. Hothorn (ed.) Abstracts of Talks and Posters. First Conference of the Central European Network Lifestat 2008, 10.-13. March 2008, Munich, Germany. (ISBN: 978-3-86541-266-9)
11. Möhring, J., A.E. Melchinger, and H.-P. Piepho. 2010. REML-based diallel analysis in plant breeding. 2nd Joint Statistical Meeting DAGStat Statistics under one umbrella. Dortmund, 23.-26. März 2010.

Acknowledgments

First and foremost, I am very grateful to my academic supervisor Prof. Dr. H.-P. Piepho for his ongoing availability, the uncomplicated way of communication, his helpful hints in all statistical issues, and especially for his endless patience during creation of this thesis.

Sincere thanks to Dr. A. Büchse and Dr. K. Hartung for helpful hints and hundrets of statistical and non-statistical discussions. I also want to thank all other actual and alumni members of the Institute of Crop Science, especially in the bioinformatik unit, for helpful hints and the pleasant working climate.

Special thanks goes to Prof. A.E. Melchinger for the efficient and professional management of GABI BRAIN and GABI GAIN, his scientific suggestions and last but not least the assumption to examinate this thesis.

I would like to thank Dr. J.C. Reif, Dr. T.A. Schrag, Dr. S. Fischer, and several other not mentioned members of the Institute of Plant Breeding, Seed Science and Population Genetics and the State Plant Breeding Institute for productive cooperations and linking my statistical knowledge with practical plant breeding problems.

Best thanks to Prof. C.P.W. Zebitz for the uncomplicated assumption to examinate this thesis.

My thanks are extended to I. Gräschus, S. Meyer and all other non-scientific members of the Institute of Crop Science and the Institute of Plant Breeding, Seed Science and Population Genetics for organization and taking over many bureaucratic work, so that it was possible to concentrate of scientific problems.

I would like to thank the cooperating breeding companies in GABI BRAIN and GABI GAIN for providing data and information about their breeding programs. Discussions during meetings helped to allocate actual breeding problems.

Finally, I am deeply indebted to my girlfriend, my friends, and my family at home who supported me in manifold ways.

This research was supported by the German Federal Ministry of Education and Research within the GABI BRAIN project (grant no. 0313126 B) and the GABI GAIN project (grant no. 0315072 B).

Erklärung

Hiermit erkläre ich an Eides statt, dass die vorliegende Arbeit mit dem Titel

„Mixed modelling for phenotypic data from plant breeding“

von mir selbst verfasst und lediglich unter Zuhilfenahme der angegebenen Quellen und Hilfsmittel angefertigt wurde. Wörtlich oder inhaltlich übernommene Stellen wurden als solche gekennzeichnet.

Die vorliegende Arbeit wurde in gleicher oder ähnlicher Form noch keiner anderen Institution oder Prüfungsbehörde vorgelegt.

Insbesondere erkläre ich, dass ich nicht früher oder gleichzeitig einen Antrag auf Eröffnung eines Promotionsverfahrens unter Vorlage der hier eingereichten Dissertation gestellt habe und ich nicht die Hilfe einer kommerziellen Promotionsvermittelung oder -beratung in Anspruch genommen habe.

Stuttgart, den 03.11.2010

Jens Möhring