Genomic prediction in rye

A THESIS SUBMITTED TO
THE FACULTY OF AGRICULTURAL SCIENCES
OF UNIVERSITY OF HOHENHEIM
IN FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF
DR.SC.AGR. / PH.D. IN AGRICULTURAL SCIENCES

Angela Maria Bernal Vasquez

Faculty of Agricultural Sciences
University of Hohenheim

March 2017

FACULTY OF AGRICULTURAL SCIENCE

Institute of Crop Science

University of Hohenheim

Biostatistics Unit

Prof. Dr. Hans-Peter Piepho

Genomic prediction in rye

Dissertation

submitted in fulfillment of the regulations to acquire the degree

"Doktor der Agrarwissenschaften"

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

to the

Faculty of Agricultural Sciences

presented by:

Angela Maria Bernal Vasquez

born in:

Bogota, Colombia

submitted in: March 2017

This thesis was accepted as a doctoral thesis (Dissertation) in fulfilment of the regulations to acquire the doctoral degree “Doktor der Agrarwissenschaften” by the Faculty of Agricultural Sciences at University of Hohenheim on August 4th, 2017

Date of the oral examination: October 24th, 2017

Examination Committee

Chairperson of the oral examination	Prof. Dr. Bennewitz
Supervisor and Reviewer	Prof. Dr. Hans-Peter Piepho
Co-Reviewer	Prof. Dr. Chris-Carolin Schön (Technische Universität München)
Additional examiner	Prof. Dr. Albrecht E. Melchinger

To my closest ancestors, my full-sib and my tester

Acknowledgments

During this long journey of making my PhD I have met people who have walked with me and have helped me either with words, with facts or with energy. The first one I am indebted to, and to whom I will never have enough words to thank is my "Doktorvater" Prof. Dr. Hans-Peter Piepho. He has guided my research with accurate advice and scientific curiosity. Thank you for the endless support and patience, for having always time for answers and for your constant inspiration. It has been a great honor to work with you.

Many thanks also to Prof. Dr. Chris-Carolin Schön for her always fine-tuned insights as project partner, co-author and now as part of the examination committee. Thanks to Prof. Dr. A. E. Melchinger for his plant breeding lectures building a baseline for this work and also for agreeing to serve on my dissertation committee.

Massive dues to Dr. Andres Gordillo, who made the bridge for me towards Germany and had always accurate personal and professional advices. He has contributed with very practical views of the sometimes encrypted results and has always allocated time to follow-up and discuss our work.

It was a pleasure for me to be part of the elite group of Rye-Select. I learnt so many things during the project meetings and profited from the discussions and ideas exchanged. Thanks to all the group led by Dr. Peer Wilde and Dr. Viktor Korzun, for allowing the use of the data and the crew support. Big big thanks to my co-authors Malthe Schmidt and Manfred Schönleben (wherever you are) for such good job cooperation and empathy. To Dr. H. F. Utz, who shared all his PlabStat outlier secrets and watched out this publication. Double thanks to my colleague Jens Möring for useful hints at the beginning of the way, for co-authoring and his unconditional support over these years.

I am very grateful to my colleagues in the Biostatistics group, the old and the new ones, for

the cheerful and positive atmosphere, useful discussions and good laughs. To Zeynep Akyildiz for taking care of my logistics and her sweet willingness to help. To Joseph O. Ogutu for his genomic selection input and Waqas Malik for his R guidance, to Vanda Lourenço for her cheerful advice, to Torben Schulz-Streeck my SAS sensei, and to Steffen Hadash for walk&talk company while philosophical and statistical musing. Extra thanks to my Hohenheim friends, the ones that are still there and the ones that have already left, because celebrating success felt double with your company and sharing sadness felt half: Marcelo (my soul buddy), Oscar, Joaquín, Hanna & Juan & Isabel, Manuel, Yudith, Valheria & José, Lina, Lucia & Marco. The thanks also cross the ocean and go to my friends in Colombia (Pipe, Migue, Natali, AngelaP) and to my former professors José Miguel Cotes and Luis Ernesto Rodríguez. I could always count on them to solve contextualization doubts and to keep connected to my roots and spirit.

I want to thank my German family, who took me in their heart with care and warmth. To my Colombian family, my mother, my father and my brother to whom I dedicate this work. You have been my infinite sun at every moment filling me with love, passion and "empuje": Gracias por su infinito apoyo, ánimo y energía!. Thank you to my dearest Gerald, for biking, walking and swimming with me, no matter what the conditions were, hand in hand all throughout this Gran Fondo. Thank you soooo much for ever!.

I finally thank you unknown reader for searching answers in my work.

Table of Contents

Nomenclature

Abbreviations

AIC	Akaike information criterion
BLUE	best linear unbiased estimation
BLUP	best linear unbiased prediction
cAIC	conditional AIC
CMS	cytoplasmic male sterility
CV	cross validation
EG-BLUP	extended genomic BLUP
FV	forward validation
GBLUP	genomic BLUP
GCA	general combining ability
GE	genotype-by-environment
GEBV	genomic estimated breeding value
GLMM	generalized linear mixed models
GP	genomic prediction
GS	genomic selection
GWAS	genome-wide association analysis
GY	genotype-by-year
LD	linkage disequilibrium
LMM	linear mixed models
MAD	median absolute deviation

MAS	marker assisted selection
MET	multi-environment trial
ML	maximum likelihood
QTL	quantitative trait loci
REML	restricted maximum likelihood
RKHS	reproducing kernel Hilbert space
RR-BLUP	ridge regression BLUP
SNP	single nucleotide polymorphism
SRR	Studentized residual razor
TBV	true breeding values
TPE	target population of environment
TS	training set
VS	validation set
WGS	whole-genome sequence

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

Rye (*Secale cereale L.*) production has tremendously improved in the past 150 years since the start of rye breeding. In the past few decades the enormous boost has been mainly due to the discovery of the P-cytoplasm and application of cytoplasmic male sterility (CMS) technology and, more recently, due to the onset of molecular techniques in genetics. Since rye is an allogamous species, rye hybrid breeding exploits heterosis resulting from sequential selfings that generate inbreeding depression, and benefits from the general and specific combination ability in the final variety [Schlegel, 2016].

Genomic selection (GS) is a tool in the breeding process that uses genomic estimated breeding values (GEBVs) based on molecular marker information to improve efficiency and cost-effectiveness [Meuwissen et al., 2001]. Genomic prediction (GP) is the term coined for the statistical procedures involved in obtaining those predicted GEBVs. The fact that rye is known to have low linkage disequilibrium (LD) [Li et al., 2011b] makes GP challenging, because there are many genes with small effects and few genes with strong effects, but also appealing, because by using many markers, quantitative trait loci (QTL) are more likely to be captured.

The present study was carried out as part of the RYE-SELECT project, funded by the German Federal Ministry of Education and Research (BMBF). The project involves several academic partners plus the breeding company KWS-Lochow and aims to study molecular tools and genetic methods to develop appropriate breeding strategies and evaluate the potential of GS. To integrate GS into the hybrid rye breeding program of KWS-Lochow, a thorough study

of GP involving phenotypic and genotypic analyses is imperative. This includes data pre-processing, evaluation of efficient phenotypic models, identification of effective and suitable GP approaches, methods to integrate crop growth models within GP, definition of optimal breeding schemes and studies on appropriate statistical methods for inference on markers effects. To address the aims of this project, in this thesis outlier detection methods, phenotypic spatial and non-spatial models and GP approaches were evaluated. The methods were implemented using empirical data produced in the hybrid rye breeding program of KWS-Lochow during the years 2009 to 2014.

1.2 The hybrid rye breeding program

Within small cereals, rye is of utmost importance due to its tolerance to cold, drought or poor soils [Schlegel, 2016]. Hybrid rye commercial breeding can be described in two stages: (i) population improvement, where parent line development is carried out, and (ii) product development, where the hybrid candidates are submitted to agronomic ranking and evaluation before commercialization [Wilde et al., 2015].

Parent line development involves: (i) recombination of pre-tested lines, (ii) subsequent selfing of their progenies, (iii) line *per se* selection, (iv) testcross and seed production, and (v) evaluation of general combining ability (GCA) to the opposite pool [Wilde et al., 2015]. A reliable CMS system is required for the seed parent pool candidates, whereas lines of the pollen parent pool must carry efficient fertility restoration genes [Geiger and Miedaner, 2009].

Testcross progenies (coming from S_2-lines) are evaluated using multi-environment trials (METs) in so-called GCA1 experiments to assess the GCA of the testcrosses. In a subsequent experiment (GCA2), the best candidates from the GCA1 experiments are subjected to further evaluations across more locations and usually with the same testers as in the GCA1 experiments. Following GCA2, the best lines are tested in a factorial experiment (FACT or GCA3) involving different and more testers and also more locations than in GCA1 and GCA2 experiments. GCA2 and FACT/GCA3 are considered follow-up experiments to confirm the selection decisions on the initial GCA1 experiment. After FACT/GCA3 experiments, a new breeding cycle starts. The new cycle uses the selected genotypes from the previous cycle as parents, so that between GCA1 in the first cycle and GCA1 of the next cycle five years elapse. Genetic gain accumulates

from cycle to cycle as a result of genetic progress in variety development [Wilde et al., 2015].

Two breeding programs (one in Germany and the other in Poland) run constantly, that is, in a given year one breeding cycle starts the parent-crossing stage, and in the same year, another cycle already undergoes the stage of early-generation testing. In this work, the data analyzed correspond to GCA1, GCA2 and FACT/GCA3 experiments of both countries. Within each experiment (which runs in a single year), the series or trials are inter-connected through checks (commercial well known lines) and subsets of genotypes are evaluated in more than one location. Hence, connectivity within a year is well established. The connection within breeding cycles (GCA1 + GCA2 + GCA3) is achieved by the selected genotypes carried forward from one year to the next, which depending on the stage (and the selection intensity) can amount to 1% to 10% of the genotypes in the preceding year. Different breeding cycles, in the opposite, remain disconnected, i.e. there are no common genotypes between cycles.

1.3 Residual analysis

Before starting any analysis, breeders (and many other specialists) deal with the uncertainty as to the reliability of the data. Often the question is whether the assumptions of the linear (mixed) models are fulfilled, i.e., normally distributed (and independent) errors with zero mean and homogeneous variance. Breeders often use a standard pipeline with which the data is quickly analyzed, residuals are estimated and a set of diagnostic plots is produced. This is the starting point to identify suspicious observations that may further result in poorly selected models, poor inference and inappropriate decisions [Gumedze et al., 2010]. Some breeders have developed their own routines to identify those suspicious observations. Scientists and academics have come up with several solutions, some of them already in-built in statistical software [Pinheiro and Bates, 2000; Schützenmeister and Piepho, 2012; Utz, 2003] and some others using more advanced methodologies [Gumedze et al., 2010; Nobre and Singer, 2007]. This thesis scrutinizes from a statistical and a practical point of view whether methods automatized in statistical packages and breeder rules of thumb to detect outliers are valid in terms of control of the Type I error rate (i.e. the probability to falsely declare a null effect to be real or non-zero).

One of the most popular statistical software packages for plant breeding trial analysis in Germany is PlabStat [Utz, 2003]. Its outlier detection method is one of the most beloved and

trusted versions among plant breeders. The software is ANOVA-based, freely distributed and many junior breeders are familiar with PlabStat since they learn how to use the software at universities. Statistical packages are now mainly using variance estimation methods based on likelihood theory, e.g., maximum likelihood (ML) estimation and restricted maximum likelihood (REML). The properties of ANOVA and REML have been extensively studied [Littell, 2002; Searle et al., 1992], most of the times giving preference to REML estimation, because with large sample size, it provides consistent (i.e., asymptotically unbiased and variance tending to zero), fully efficient (i.e., minimum variance) and normally distributed parameter estimators [Burnham and Anderson, 1998]. The transition from ANOVA to REML poses a challenge for users of ANOVA packages such as PlabStat accustomed to in-built facilities, e.g. outlier identification tools. Other breeders not familiar with PlabStat may have developed their own rules based on their experiences or stick to the popular ones, such as one that in the present study is called Studentized residual razor (SRR), also widely used in practical analysis of plant breeding trials. In this rule, if the absolute value of the Studentized residual of an observation is beyond the $(1 - \alpha/2)$-quantile of the normal distribution (with α the expected proportion of falsely flagged residuals), that observation should be considered an outlier and handled accordingly.

The problem of judging an observation as an outlier can be seen as a hypothesis test, requiring a test of significance. When there are several observations to be judged, a multiple testing problem arises. This problem refers to the fact that with lots of tests, the probability of finding at least one significant but acceptable outlying observation by chance alone may be inappropriately large [Snedecor and Cochran, 1980]. Many different methods have been described to deal with the multiple testing problem, i.e., methods controlling the experiment-wise Type I error rate. In this work, the Bonferroni-Holm test [Holm, 1979] is considered using different types of standardized residuals.

The area of diagnostic statistics is very broad and actively studied because many questions arise in judging reliability and quality of the data. In plant breeding, where data undergo several analyses (e.g., phenotypic analysis and then genotypic analysis), one of the questions arising is whether dropping outliers in a preliminary stage has a positive effect on a further stage, e.g., in GP. And in such a case, what attributes should the method have to provide a secure outlier detection for safe and reliable further analyses. This topic is covered in Chapter 2 using a case study and published field trial datasets.

1.4 Phenotypic analysis towards genomic prediction

Phenotypic analyses presented in this thesis follow a stage-wise approach. In the first stage or stages, adjusted genotype means are obtained from fitting a statistical model to the raw plot data. Next, the adjusted means can be used to obtain GEBVs in a GP stage [Piepho et al., 2012a]. In this section, the general basis for the phenotypic analysis approach is described followed by the principles of the GP methodology.

1.4.1 A stage-wise approach for phenotypic and genomic prediction analyses

Linear mixed models (LMM) have been used to analyze advanced trial layouts allowing to account for random sources of field variation as well as for genetic and environmental covariances commonly found in plant breeding METs [Cooper et al., 2014; Piepho et al., 2008b]. LMM can be implemented as a single-stage model or by a stage-wise approach [Piepho et al., 2012a; Smith et al., 2001]. A potential drawback of the former approach is the computational load, specially if there are large sets of genotypes, many locations and complex variance-covariance structures of the genotype-by-location effects. Thus, in practice, the use of stage-wise analysis alleviates the computational burden, although accounting for heteroscedasticity and heterogeneity of covariances among the adjusted means can be challenging [Möhring and Piepho, 2009]. In addition, when single locations need specific adjustment to characterize the within-location variation, a stage-wise approach may be more convenient [Piepho et al., 2012a], and more intuitive because of its simplicity and understandability [van Eeuwijk et al., 2016].

Stage-wise approaches have been found to reproduce well a single-stage analysis [Möhring and Piepho, 2009; Piepho et al., 2012a]. In general, adjusted genotype means per location are computed in the first stage and then, to minimize loss of information, a variance-covariance matrix of those means approximated by some diagonal matrix is used for weighting in the next stage [Piepho et al., 2012a; Schulz-Streeck et al., 2013a; Smith et al., 2001]. Attention should be paid to the hierarchy of factors used in the experimental design, as well as to fixed or random status of an effect in the model. The factors used for the analysis are for example: genotype, testers, locations, trials within locations, replicates within trials, incomplete blocks within replicates, and the effects are the model terms defined for these factors [Piepho et al., 2012a]. Further, a distinction between treatment factors and design factors should be made:

design factors are innate to the experimental design units, whereas the treatment factors are those whose levels are randomly allocated to experimental units [Brien, 1983; Piepho et al., 2012a]. This is important because for the first stage, the level where data will be summarized to compute adjusted means should be decided. Additionally, caution should be taken when setting effects as random or fixed across the stages in order to avoid "double shrinkage" [Smith et al., 2001] and to ensure a proper estimation of the effects across stages [Piepho et al., 2012a]. Paying due attention to these issues, a stage-wise approach fits perfectly into the GP framework, where adjusted genotype means are the substrate for the prediction of GEBVs.

1.4.2 Genomic prediction and related factors

Genomic selection is the term coined for the process of using not only large QTL but all markers available along the genome in order to be able to predict GEBVs of genotypes that have not been tested in the field, i.e., they have no phenotypic observations but genotypic information is available. GEBVs refer to the purely additive genetic effect calculated as the sum of marker effects for a given genotype. Selection decisions are informed by the GEBVs [Meuwissen et al., 2001]. The methodology was developed as an option to overcome the shortcomings of marker assisted selection (MAS), which was based only on the use of large contrasting QTL effects [Bernardo, 2008; Jannink et al., 2010]. Thus, GS requires that markers cover all the genome and are sufficiently separated from one another to use the LD and in this way, the effect of the marker in a given position can be calculated. The advantage of GS is the increase of the cost-effectiveness of the breeding cycle, i.e., the gain per unit time [Heffner et al. 2009, 2010], because it is possible to: (i) pre-select good crosses and (ii) avoid phenotyping all genotypes in the field, thus increasing the number of evaluated individuals per cross.

In general, GS consist of the following steps:

- Genotype and phenotype a set of genotype entries. This population is called the training set (TS).
- Only genotype another set of genotype entries, whose GEBVs are to be predicted. This population is called the validation set (VS).
- Train a model using the TS (phenotyped and genotyped entries) and the molecular marker information of the VS to predict GEBVs of unphenotyped entries.

- Select candidate genotypes from the VS according to the predicted GEBVs.

Multiple interrelated factors affect GS accuracy: training and validation sets size [Auinger et al., 2016; Crossa et al., 2014; Schulz-Streeck et al., 2012], number and type of markers [Heslot et al., 2013b], level of LD between markers and QTL [Habier et al., 2007; Meuwissen et al., 2001; Riedelsheimer et al., 2012], heritability of the trait under scrutiny [Daetwyler et al., 2013; Goddard, 2009; Guo et al., 2014b; Habier et al., 2007], non-additive genetic variation [Toro and Varona, 2010; Wang et al., 2014; Wellmann and Bennewitz, 2011], population structure [Guo et al., 2014b; Isidro et al., 2015; Windhausen et al., 2012], degree of genetic relatedness between TS and VS [Crossa et al., 2014; Wientjes et al., 2013] and genotype-by-environment (GE) interaction [Burgueño et al., 2012; Heslot et al., 2014].

1.4.3 Genomic prediction models

In animal breeding, pedigree-based on the best linear unbiased prediction (BLUP) has been used long before the advent of molecular markers since, often, the breeding value of an animal cannot be estimated through direct observations but only through the evaluation of the progeny. For example, a bull's breeding value for milk yield can only be estimated via its daughters' and grand daughters' milk yield. To exploit this pedigree information, genetic correlation needs to be modeled. BLUP allows modeling of those correlations. Conversely, for plant breeders it is possible to test the same genotype in multiple environments, where a simple mean across environments provides a rather accurate estimate of the genotypic value, so in the past there has been little pressure to exploit information from relatives in order to improve precision. Plant breeders have only fairly recently started to embrace the adoption of BLUP [Piepho et al., 2008a]. Animal breeders have always used BLUP, so the step towards GP BLUP-based models was straightforward for them, whereas for plant breeders, who had never used BLUP before and always relied on simple arithmetic means or best linear unbiased estimators (BLUE), GP took off later.

High-density marker systems, characterized by a large number of markers (p), represent a problem in GP since number of genotypes (n) are far less than p, so that standard multiple regressions are impossible when $p >> n$. A plethora of GP methods have been proposed to overcome this limitation and to allow the estimation of marker effects. Most common approaches involve mixed models [Meuwissen et al., 2001; Piepho, 2009b] and Bayesian methods

[Gianola, 2013; Meuwissen et al., 2001], but choices are not restricted to these [Heslot et al., 2012; Ogutu et al., 2012].

Throughout this work, GP will be done exclusively using LMM. The GP model uses the markers as random factors to be able to compute BLUPs of the marker effects. The model is based on the assumption that the marker effects are sampled from the same normal distribution. Thus, all markers have the same variance and are assumed to be very small. The estimated marker effects are shrunk by a penalty parameter (λ^2), thus avoiding over-fitting and stabilizing the estimation. The GEBV of a genotype is thus the sum of its estimated marker effects. The ridge regression BLUP (RR-BLUP) method allows to use REML to estimate the penalty parameter [Piepho, 2009b]. The method RR-BLUP is equivalent to genomic BLUP (GBLUP) [Goddard, 2009; Habier et al., 2007] and leads to the same results as long as kinship matrices are equally scaled [Habier et al., 2007] but computation load is different depending on the size of n and p. If $p >> n$, GBLUP may be more favorable, whereas for $p < n$, RR-BLUP is more convenient.

The disadvantage of assuming marker effects with homogeneous variances (as RR-BLUP and GBLUP do) is that large QTL effects may be underestimated [Meuwissen et al., 2001]. Several Bayesian methods have been developed to overcome this restriction. In general, Bayesian methods differ in the marker effects' prior distribution, which may allow for each marker effect to be shrunk differently and for some marker effects to be zero [Gianola, 2013; Habier et al., 2013]. Simulations and real data studies that compare GP methods (BLUP- and Bayesian-based) have shown that similar results are achieved [Piepho, 2009b; Technow et al., 2014; Wimmer et al., 2013; Zhao et al., 2013], but when there are large QTL effects Bayesian methods perform better [Habier et al., 2013; Kemper et al., 2015; Pryce et al., 2011; Van den Berg et al., 2015; Wellmann and Bennewitz, 2012].

Accuracy of predicted GEBVs can be measured using techniques such as k-fold cross validation (CV), leave-one-out CV or forward validation (FV). Details of these approaches are introduced in the sub-section 1.5.2. The correlation between predicted GEBVs and the observed phenotypic values (usually adjusted genotype means) is known as the predictive ability of the model. When the heritability is taken into account (i.e. predictive ability divided by square root of heritability), one refers to predictive accuracy [Dekkers, 2007], which assesses the correlation between true breeding values (TBV) and GEBVs. Throughout the chapters

of this thesis, the predictive abilities are used to assess the prediction performance of the GP approaches considered.

1.5 The model choice: linking phenotypes to genotypes

The phenotypic analysis preceding the GP stage may have an influence on the prediction of GEBVs. The challenges that arise regarding model choice in the phenotypic analysis and genomic prediction are introduced in this section.

1.5.1 Spatial models as an *add-on* of the phenotypic model

Spatial models have been used to account for heterogeneity of the growing conditions at the trial (or location) level [Cullis et al., 2006; Gilmour et al., 1997; Piepho et al., 2008b; Williams et al., 2006]. The fact that adjacent plots are more similar than non-adjacent plots favors the use of geostatistical variance-covariance structures. The key idea of modeling a spatial trend is to fit a "mixture of spatial covariances and/or deterministic functions of spatial coordinates" [Gilmour et al., 1997].

One- and two-dimensional methods to model spatial trend in the context of field trial have been extensively studied [Besag and Kempton, 1986; Cullis and Gleeson, 1991], as well as more sophisticated methods such as using random effects to model large-scale trends [Gilmour et al., 1997] or using smoothing splines [Verbyla et al., 1999]. In practice, approaches based on mixed models with a spatial component for local trend have gained popularity because of their flexibility and step-wise approach to model selection [Piepho et al., 2008b; Piepho and Williams, 2010; Williams et al., 2006].

The trials of the rye hybrid program from KWS-Lochow are laid out as α-designs with two replicates. Fields are rectangular and incomplete blocks are laid out longitudinally across the fields. The experimental plots are identified by row and column numbers, thus besides the blocks accounting for field heterogeneity, row and column factors can also be included in the model as post-blocking effects and to account for spatial trend, as an *add-on* to the non-spatial models. In Chapter 3, a comparison between spatial and non-spatial models is explored towards implementation of GP in the hybrid rye program.

1.5.2 Genomic prediction - cross validation (GP-CV) as a tool for model selection

Model selection in biological sciences in the context of LMM has been addressed by the use of the Akaike Information Criterion (AIC). The relation of the discrepancy between two models and the ML allowed practical and theoretical development of model selection theory. The model with the smallest AIC value represents the best approximation for the information in the data to the truth, relative to the other models considered in a set of candidate models. It is possible that none of the models in the set are good, but AIC will select the best approximating model of those in the candidate set [Burnham and Anderson, 1998].

Cross validation (CV) has been suggested and well studied as basis for model selection [Arlot and Celisse, 2010]. In general, it consists in partitioning the data into two parts, one is used for training the model (training set, TS) and the other for validating the model (validation set, VS). Then, another partition is considered and the process is repeated many times [Burnham and Anderson, 1998]. The several approaches for CV differ in the number of partitions (folds) and the number of partitions that are used for the training and the validation set [Arlot and Celisse, 2010], e.g., leave-one-out, where each data point is successively "left out" from the sample and used for validation, or k-fold CV, where the data is split into k subsamples of approximately equal size and each subsample is successively used for validation. Recently, the prediction of untested scenarios, say future years or unobserved environments, has become important due to its practical relevance. This validation approach is called forward validation (GP-FV) [Battenfield et al., 2016; Plieschke et al., 2015].

In the context of GP, k-fold CV is one of the most popular approaches to evaluate the predictive ability or predictive accuracy (if heritability is considered) of a certain model. In Chapter 3, spatial and non-spatial models are compared using the AIC and, since marker data are available, the predictive accuracy of a GP procedure can also be taken as a selection criterion to decide which phenotypic model fits best the data. This latter approach is called genomic prediction - cross validation (GP-CV).

To compare models via AIC when REML is used, it is required that the models have the same fixed effects. If this is not the case, ML can be used [Vaida and Blanchard, 2005; Wolfinger, 1996]. It would be preferable, however, to use REML given its smaller bias property [Searle et al., 1992]. The advantage of using GP-CV is that it can also be used with REML to compare models with different fixed effects structures.

1.5.3 Accounting for genotype-by-environment (*GE*) interaction effects

In Section 1.2 the general structure of the rye hybrid program was presented. It is important to emphasize that breeding cycles, in particular GCA1 experiments from different years, are disconnected, meaning that they do not share common genotypes. This situation becomes a problem once it is desired to pull the data together in order to increase the TS-size for GP. In this disconnected scenario, a genotype-by-year (GY) interaction effect will become confounded with genotype main effects, so that an estimation of the mean value of a genotype becomes inaccurate. For these reasons, plant breeders have preferred to analyze their datasets by year, where replicates and locations allow them to estimate a reasonably accurate genotype mean. A paradigm change has come into play with the popularity of GP, where large TS sizes are important to achieve high accuracy of predictions. Merging multiple years of data from different cycles is the obvious way to increase TS-size, which is necessary to improve predictive accuracy. Further, the use of genetic relationships among genotypes across years constitutes a potential tool to improve the estimates of genotypes across disconnected scenarios. In Chapters 3 and 4 some insights are provided, first, to deal with the disconnectivity problem, and then, to separate GY from genotype main effects.

Burgueño et al. [2012] report important gains in predictive accuracy MET compared to single environments. Heslot et al. [2014], Jarquín et al. [2014] and Malosetti et al. [2016] incorporate GE effects in the GP model by explicitly modeling environmental covariables. That is, the variance-covariance matrix of the GE effects corresponds to the Kronecker product of a matrix reflecting similarities among genotypes in terms of their markers' profile (e.g., kinship matrix), and a matrix reflecting similarities among environments induced by the growing conditions. The authors showed that including environmental covariables allowed predicting new environments. In the case environmental information is not available, the kinship can still be used to model GE, where the covariance matrix is the product between an identity matrix, multiplied by a variance component, and the kinship matrix.

In Chapter 4, the fact is exploited that a rye breeding cycle runs across years, so that splitting the environmental effect into year, location and interaction effects would allow to separate the GY effects from the genetic effects. The challenge is that GCA1 trials across years are disconnected. Hence, one option is to use complete cycles (GCA1 + GCA2 + GCA3) in the TS, disregarding the fact that GCA2 and GCA3 data are selection-biased, which is appropriate

as long as all the data used to make selection decisions are included in the analysis [Piepho and Möhring, 2006]. Another option is to use the kinship to model the genetic correlations between genotypes across years (i.e. use kinship to model GY). This is possible, despite lack of common genotypes between years, because there is plenty of replication of alleles between genotypes.

Another problem associated with environmental effects is that, e.g., in a year where there was a disease pressure or environment stress factors, yield-QTL can be confounded with non-yield QTL. If there is no sufficient phenotypic information to explicitly separate these effects, e.g., due to disconnected years, a pre-selection in the TS may potentially help to reduce the confounding effect. This approach is studied in Chapter 4.

1.6 Objectives and hypotheses

In the RYE-SELECT project, the main objective was to develop genome-based breeding strategies to improve selection efficiency for yield and other agronomically important traits in the rye hybrid breeding program. Since all of the steps considered in the process are important and can contribute to the accuracy of predictions in GS, efforts to improve any of these steps are worth undertaking. The specific objectives in this thesis were to:

1. Scrutinize outlier detection tools currently used by many breeding programs and propose reliable and easy-to-use outlier detection methods

2. Evaluate the merits of spatial models for phenotypic data analysis

3. Evaluate and compare approaches that allow dissecting genotype from genotype-by-year effects in disconnected datasets

4. Use a forward validation approach to evaluate the predictive ability of genomic prediction models under scenarios of common practice for plant breeders, such as prediction of VS of different relatedness degree with the TS, and prediction of a selection of top performance genotypes.

In line with the objectives, the hypotheses tested in this thesis are:

1. Outliers have an effect on the GP predictive ability and choice of detection method is important

2. Spatial models for phenotypic analysis produce more accurate genotype estimates leading to a better predictive ability than afforded by purely randomization-based models

3. The genotype effect can be separated from the genotype-by-year effect in disconnected year-data using the kinship matrix

4. The higher the relatedness between TS and VS, the higher the predictive abilities

1.7 Outline of the thesis

This doctoral work is conceived as a cumulative thesis, where each chapter is a journal article. All the articles are framed as case studies using empirical data. In Chapter 2, ANOVA- and REML-based analysis are reviewed towards understanding the PlabStat procedure for outlier detection, and additional methods are proposed. Chapter 3 covers the analysis of single-year data using spatial and non-spatial models and a comparison of GP approaches when datasets are weakly connected. In Chapter 4, multi-year data is used to present different ways to fit genotype-by-year effects. The models are compared and evaluated using a forward validation approach, under scenarios of different relatedness between TS and VS and top-yield selection in the TS. A general discussion is presented in Chapter 5.

Chapter 2

Outlier detection methods for generalized lattices: A case study on the transition from ANOVA to REML[1]

Angela-Maria Bernal-Vasquez[a], H. Friedrich Utz[b], Hans-Peter Piepho[a]

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

[b] Plant Breeding Institute, University of Hohenheim, Fruwirthstrasse 21, 70599 Stuttgart, Germany

2.1 Abstract

Key message: **We review and propose several methods for identifying possible outliers and evaluate their properties. The methods are applied to a genomic prediction program in hybrid rye.**

Many plant breeders use ANOVA-based software for routine analysis of field trials. These programs may offer specific in-built options for residual analysis that are lacking in current REML software. With the advance of molecular technologies, there is a need to switch to REML-based approaches but without losing the good features of outlier detection methods that have proven useful in the past. Our aims were to compare the variance component estimates between ANOVA and REML approaches, to scrutinize the outlier detection method of the ANOVA-based package PlabStat and to propose and evaluate alternative procedures for outlier detection.

[1] A version of this chapter is published as:
Bernal-Vasquez, A.-M., Utz, H.F., & Piepho, H.-P. (2016). Outlier detection methods for generalized lattices: a case study on the transition from ANOVA to REML. *Theoretical and Applied Genetics*, 129:787-804.

We compared the outputs produced using ANOVA and REML approaches of four published datasets of generalized lattice designs. Five outlier detection methods are explained step by step. Their performance was evaluated by measuring the true positive rate (TPR) and the false positive rate (FPR) in a dataset with artificial outliers simulated in several scenarios. An implementation of genomic prediction using an empirical rye multi-environment trial was used to assess the outlier detection methods with respect to the predictive abilities of a mixed model for each method.

We provide a detailed explanation how the PlabStat outlier detection methodology can be translated to REML-based software together with the evaluation of alternative methods to identify outliers. The method combining the Bonferroni-Holm test to judge each residual and the residual standardization strategy of PlabStat exhibited good ability to detect outliers in small and large datasets and under a genomic prediction application. We recommend the use of outlier detection methods as a decision support in the routine data analyses of plant breeding experiments.

Keywords Generalized lattices (GL), residual, outlier, generalized least squares estimation (GLSE), restricted maximum likelihood (REML), ANOVA, genomic prediction (GP), receiver operating characteristic (ROC) curve.

2.2 Introduction

Plant breeding companies in Germany have been examining their phenotypic datasets routinely using statistical programs especially designed for analysis of breeding trials. A popular example is PlabStat [Utz, 2003], which implements Cochran and Cox [1957] theory based on analysis of variance (ANOVA) methods. As hardware improved, more complex statistical methods for mixed models [e.g., maximum likelihood (ML), restricted maximum likelihood (REML), Bayesian methods] became feasible. Although the transition to these methods may produce some sting when the outputs do not exactly match or do not offer the same information, the motivation to use them is permanent since new powerful breeding tools, such as genomic prediction (GP), are usually applied under these statistical frames.

Genomic prediction computes genomic breeding values (GEBV) by using marker information in best linear unbiased prediction (BLUP) based on phenotypic data [Meuwissen et al., 2001]. Ridge-regression BLUP (RR-BLUP) has been identified as a simple and accurate method to obtain GEBV [Piepho, 2009b]. GP is rapidly becoming part of the routine data analysis in breeding companies; thus, handling the phenotype and genotype analysis under the same format (and software) helps avoiding the extra file editing and exchange across programs.

A popular part of the outputs offered by the PlabStat software is, for example, a structured residual

analysis for the identification of possible outlying observations based on analysis of variance (ANOVA) approach and a special treatment of missing observations, which does not fully coincide with the standard output of REML-based packages. In this paper we therefore compare the theoretical underpinnings and results obtained by both approaches.

Breeding trials are typically laid out as multi-environmental trials (METs), that is, trials are performed in several locations and comprise a large number of genotypes. For a large number of genotypes, generalized lattice (GL) designs are a popular class of designs. GL designs are defined as "block designs for $v = ks$ varieties in $b = rs$ blocks of k units such that the blocks can be arranged into r complete replications, i.e., the designs are resolvable" [Williams, 1977]. Square lattice designs ($k = s$) and rectangular lattice designs ($k = s - 1$) are included in this definition, as well as α-designs, which involve a cyclic method of construction based on α-arrays. Analysis of such designs requires mixed models, for which ANOVA and REML-based analyses are not identical.

Statistical software such as PlabStat makes use of Cochran and Cox [1957] theory. The analysis of METs in PlabStat uses a stage-wise approach. In the first stage, individual trials are analyzed. The genotype means obtained in the first stage are then summarized across environments in the second stage. Normally, in many advanced-generation testing programs, there are several trials in one environment laid out as GL designs [Piepho et al., 2006], each described by the model

$$y_{ijh} = \mu + \gamma_i + \tau_h + b_{ij} + e_{ijh}, \tag{2.1}$$

where y_{ijh} is the observation of the hth genotype in the jth block within the ith complete replicate, μ is the general mean, γ_i is the effect of the ith complete replicate, τ_h is the effect of the hth genotype, b_{ij} is the effect of the jth incomplete block nested within the ith complete replicate and e_{ijh} is the residual plot error associated with y_{ijh}. For the analysis in PlabStat, it is assumed that the error is normally and independently distributed with mean zero and variance σ_e^2 $[e_{ijh} \sim N(0, \sigma_e^2)]$ and the block effects are normally and independently distributed with mean zero and variance σ_b^2 $[b_{ij} \sim N(0, \sigma_b^2)]$. This latter assumption implies that ordinary least squares estimation (OLSE) cannot be applied to estimate the fixed effects because the observations that are in the same block are positively correlated [Cochran and Cox, 1957, p.382]. Instead, generalized least squares estimation (GLSE) can be used. Particularly, PlabStat follows the iterative procedure suggested by Williams [1977].

The predictions, $\hat{y}_{ijh}$, are defined as

$$\hat{y}_{ijh} = \hat{\mu} + \hat{\gamma}_i + \hat{\tau}_h + \hat{b}_{ij}, \tag{2.2}$$

where $\hat{b}_{ij}$ is the adjusted block effect calculated by weighting factors (or shrinkage) using ANOVA estimates of the variance components σ_b^2 and σ_e^2 [Cochran and Cox, 1957]. The block effect estimator corresponds to the BLUP of b_{ij}. The estimators $\hat{\mu}$, $\hat{\gamma}$ and $\hat{\tau}$ are the best linear unbiased estimators (BLUEs) of the fixed effects for intercept, replicates and genotypes, respectively.

In PlabStat, missing values are imputed iteratively minimizing the residual mean square following the method of Yates (1933) (cited by Utz, 2003). The imputation procedure is equivalent to inserting values for the missing observations by means of the "intra-block" formula, that is, fitting incomplete blocks as fixed effect and estimating the model parameters using OLSE [Cochran and Cox, 1957]. These estimates are used as if they were observed data, so that the degrees of freedom for the residual error sum of squares is the only change made in the analysis of the combined dataset (observed and missing). The degrees of freedom are equal to those for complete data reduced by the number of estimated missing observations [Searle, 1987, p.364].

PlabStat uses an ANOVA method to estimate variances, while most common mixed model packages use REML. Some software packages have the option to switch to the ANOVA method of variance estimation, so that the PlabStat output can be exactly resembled. For example, the MIXED procedure of SAS has the option *method = type1*, which allows ANOVA estimators to be computed from sequential sum of squares.

Outputs of residuals, predictions and block effects are equal between PlabStat and the REML approach for balanced datasets, provided the ANOVA estimates of the variance components are positive [Searle et al., 1992]. However, GL designs exhibit planned unbalancedness (which is sometimes mistaken as balanced data), because the blocks are incomplete [Littell, 2002].

In the case of missing values, a REML-package fits the model to the incomplete data whereas PlabStat, by default, imputes the missing values; hence, the data analyzed are in the end different. Therefore, even in the case of switching to ANOVA-estimation in the REML-software, variance component estimates remain different.

Variance components estimation is a permanent area of research in data analysis. Developments in the last decades relate to the transition from ANOVA to likelihood- and generalized least squares-based inference [Littell, 2002]. Which variance estimation approach is best, has been widely discussed. The estimators using the maximum likelihood principle are consistent and asymptotically normal, and the asymptotic sampling dispersion matrix is known so that confidence intervals and hypothesis test of parameters are available. ANOVA estimators, on the other hand, are unbiased but often have larger variance [Searle et al., 1992, p.255].

It is also worth mentioning the recent developments in outlier detection methods for linear mixed models (LMM) and generalized linear mixed models (GLMM), that include the variance shift outlier model (VSOM), determining if each observation has an inflated variance by assessing its individual likelihood ratio and using score test statistics [Gumedze et al., 2010], a decomposition of a generalized leverage matrix of the LMM that helps detecting leverage points for marginal and conditional fitted values [Nobre and Singer, 2011], using residual plots to compare empirical residual distributions to appropriate null distributions constructed using parametric bootstrap [Schützenmeister and Piepho, 2012], and an extension of the Cook's distance to factors that allow identifying the influence on the fixed effect estimation or on the random effects prediction [Pinho et al., 2015].

This work aims to (1) compare the variance component estimates of ANOVA and REML-based approaches in analysis of plant breeding designs, (2) elucidate the outlier detection method implemented by PlabStat, and (3) evaluate the ability of alternative procedures to identify outlying observations.

2.3 Materials and Methods

2.3.1 Statistical model

For convenience, we introduce the matrix form of model (2.1):

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \mathbf{e}, \tag{2.3}$$

where $\mathbf{y}$ is the vector of observations, $\mathbf{X}$ and $\mathbf{Z}$ are the design matrices of fixed and random effects, respectively, β is a vector of fixed effects, $\mathbf{u}$ is the vector of random effects and $\mathbf{e}$ is the vector of errors. The fixed effects in vector β comprise the intercept μ, the replicate effect γ_i, and the treatment effect τ_h, whereas the vector $\mathbf{u}$ comprises the block effect nested within replicates, b_{ij}.

An outlier is defined as an observation that has a large residual in comparison with most of the other observations, so that it may need to be treated specially, e.g., as a missing value [Anscombe and Tukey, 1963]. A raw residual, r_{ijh}, of the ijhth observation, y_{ijh}, is defined as

$$r_{ijh} = y_{ijh} - \hat{y}_{ijh}, \tag{2.4}$$

where the fitted value $\hat{y}_{ijh}$ is obtained from a given fitted model. The matrix form of Eq. (2.4) is equivalent to $\mathbf{r} = \mathbf{y} - \hat{\mathbf{y}}$, where $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} + \mathbf{Z}\hat{\mathbf{u}}$

2.3.2 Description of examples and procedures for variance estimation

We used four known published datasets on GL experiments: an α-design of 24 oats genotypes in three replicates each consisting on six blocks [John and Williams, 1995, p.146], a 5×5 square lattice for 25 soybean varieties [Cochran and Cox, 1957, p.406], a 3×4 rectangular lattice for 12 treatments [Cochran and Cox, 1957, p.418] and a 9×9 triple lattice with 81 rice varieties [Gomez and Gomez, 1984, p.55-56]. Datasets are available in Appendix A with a complete description of the design settings. Three versions of each of these datasets were considered (Table 2.1). First, as originally published, then, deleting some observations (as representing missing observations) and finally generating outliers.

To assess the similarity of the outputs of a PlabStat analysis and a REML-based software also in the case of missing observations, we used the original datasets (Examples 1.1, 2.1, 3.1 and 4.1) and the sets with missing observations (Examples 1.2, 2.2, 3.2 and 4.2) and compared the divergence between the variance component estimates and the difference in the number of the outliers detected. We used the MIXED procedure of SAS, which allows to perform a REML analysis and to switch to ANOVA; thus, we were able to exactly resemble the PlabStat procedure by using the ANOVA method in SAS, hereafter denoted as SAS-ANOVA, and compare PlabStat with the same procedure but using REML, hereafter denoted as SAS-REML.

Table 2.1: Labels and description of the examples used for analysis.

Label	Description
Example 1.1	α-Design original published [John and Williams, 1995]
Example 1.2	α-Design with three missing observations
Example 1.3	α-Design with three outlying observations
Example 2.1	Triple lattice original published [Gomez and Gomez, 1984]
Example 2.2	Triple lattice with three missing observations
Example 2.3	Triple lattice with three outlying observations
Example 3.1	Square lattice original published [Cochran and Cox, 1957]
Example 3.2	Square lattice with three missing observations
Example 3.3	Square lattice with three outlying observations
Example 4.1	Rectangular lattice original published [Cochran and Cox, 1957]
Example 4.2	Rectangular lattice with three missing observations
Example 4.3	Rectangular lattice with three outlying observations
Example 5.1	Rye MET

2.3.3 Outlier detection methods

We used the datasets with artificial outliers (Examples 1.3, 2.3, 3.3 and 4.3) to implement the outlier detection method of PlabStat under a REML-based framework and to compare this output with other methods described in the following.

In a first approach of all the methods, we computed residuals using model (2.1), i.e., assuming random incomplete block effects $[b_{ij} \sim N(0, \sigma_b^2)]$. In this case, residuals $\mathbf{r}$ depend on other random effects. In a strict sense, residuals are confounded with the random effects $\mathbf{u}$, since estimating $\mathbf{r}$ requires an estimate of $\mathbf{Zu}$ [Nobre and Singer, 2007]; hence, the residuals themselves are biased. Therefore, assuming fixed incomplete block effects and using OLSE to obtain an unbiased error variance estimate remains as a potential alternative to scrutinize the datasets and identify outliers [Schützenmeister and Piepho, 2012]. In a second approach, we implemented all methods using the same model (2.1) but assuming incomplete blocks as fixed effects. We labelled the methods that considered incomplete blocks as random with *r* and the ones that considered blocks as fixed with *f* (Table 2.2). Hereafter we refer to the former methods as "r-methods" and to the latter as "f-methods".

Table 2.2: Labels and short description of outlier detection methods.

Label		Method	Incomplete blocks
M1r	PS	PlabStat	random
M1f	PS	PlabStat	fixed
M2r	BH-SR	Bonferroni-Holm with studentized residuals	random
M2f	BH-SR	Bonferroni-Holm with studentized residuals	fixed
M3r	SRR	Studentized residual razor	random
M3f	SRR	Studentized residual razor	fixed
M4r	BH-MADR	Bonferroni-Holm with re-scaled MAD standardized residuals	random
M4f	BH-MADR	Bonferroni-Holm with re-scaled MAD standardized residuals	fixed
M5r	BH-STRO	Bonferroni-Holm with robust studentized residuals	random
M5f	BH-STRO	Bonferroni-Holm with robust studentized residuals	fixed

Method M1: Outlier detection in PlabStat (PS)

Although the judging process for outlier detection can be purely subjective and difficult for the non-specialist, in routine analysis, when a large bulk of data or many similar smaller sets need to be analyzed,

the user may also use a rejection rule for outliers in order to be protected against adverse effects of spurious readings [Anscombe, 1960]. In a classical paper, Anscombe [1960] compares a "householder's fire insurance policy" to a rejection rule of residuals in the sense that both involve three key concepts: a payable *premium* for using the insurance policy or the rejection rule, a protection level in case of the event and the real danger of happening. Given the fact that fires or spurious readings *do* occur, one should worry more about the *premium* and the protection level rather than the danger. "The *premium* may be taken as the percentage increase in the variance of estimation errors due to using the rejection rule, when in fact all the observations come from a homogeneous normal source; the protection given is the reduction in variance (or mean square error) when spurious readings are present" [Anscombe, 1960]. The generalization of the threshold for the rejection rule was published a couple of years later [Anscombe and Tukey, 1963] and adopted (with some modifications) for plant breeding applications in the PlabStat software as default outlier detection method [Utz, 2003].

In METs, the outlier detection process of PlabStat comprises one step at the trial level and, depending on the user's needs, another step at the across-environments level. In general, a raw residual (r_{ijh}) is computed for each observation. From these, the median absolute deviation (MAD) among raw residuals is calculated to later define a threshold for outlier identification. The observations flagged as outliers are the ones whose raw residuals exceed the threshold, which in turn depends on the residual degrees of freedom, df_e, and the number of observations, n (See description below).

Standardizing raw residuals is useful to define a variant of residual that is independent of the scale and, thus, easier to judge [Cook and Weisberg, 1982, p.17]. There are different kinds of standardization, i.e., division by an estimate of the residual's standard deviation (studentized residuals), division by the standard deviation of y_{ijh} (Pearson residuals), division by a robust scale estimate or, as is used in the output display of PlabStat, by the square root of the effective error mean square, which is yet to be explained and which is only computed to judge the relative magnitude of the suspicious observation.

Notice that PlabStat reports raw residuals to flag outliers, but effectively, it uses residuals standardized by a robust scale estimate (s^r) to identify outliers since this estimate is used in the computation of the threshold to flag raw residuals. Additionally, residuals standardized by the square root of the effective error mean square are printed in the output of the software. The outlier detection method of PlabStat is described in detail below:

1. Compute raw residuals, $r_{ijh} = y_{ijh} - \hat{y}_{ijh}$, where $\hat{y}_{ijh}$ is given in Eq. (2.2).

2. Compute the effective error mean square (MSE_{Eff}) as $\text{MSE}_{\text{Eff}} = \text{m.v.d.} * rep/2$, where m.v.d. is the mean of the variances of the difference between all pairs of adjusted means and rep stands for the number of replicates.

3. Compute the median of absolute deviation from the median, MAD [Iglewicz, 2000, p.408], using the raw residuals, r_{ijh}.

$$\text{MAD} = \text{median}\ \{|r_{ijh} - \text{median}\ \{r_{ijh}\}|\}.$$

4. Calculate the re-scaled MAD, s^r, as $s^r = \text{MAD} \times 1.4826$. The re-scaled MAD is a robust estimate of the standard deviation. 1.4826 is a scaling factor for the normal distribution [Ruppert, 2011, p.118]. This is an approximation because the scale assumes independent identically distributed (i.i.d.) residuals but r_{ijh} are not independent or homoscedastic.

5. Compute the threshold $[-s^r CP, s^r CP]$, where s^r is the estimated robust standard deviation calculated in the previous step, C is a given constant [Anscombe and Tukey, 1963] and $P = 1.15$, a constant defined based on research experience of the second author.

   To calculate the value of C, we use the approximative formula by Anscombe and Tukey [1963]:

$$C = K \left\{ 1 - \frac{K^2 - 2}{4df_e} \right\} \sqrt{\frac{df_e}{n}}, \tag{2.5}$$

   where df_e denotes the degrees of freedom of the error, n is the total number of residuals and

$$K = 1.40 + 0.85N, \tag{2.6}$$

   with N, the value of the normal quantile that can be calculated by solving

$$premium = 100 \frac{n}{df_e} \Phi(-N) \quad \text{per cent} \tag{2.7}$$

   for N, where Φ stands for the standard normal cumulative density and *premium* is the penalty (charged as increase in percentage of the variance) to be paid due to the use of a protection. In practice, for a *premium* of 2%, "the chance that the spurious observation will escape rejection is of the order of 0.02", and "how much *premium* one is willing to pay depends on how greatly we fear spurious observations [Anscombe, 1960]".

   The default *premium* in PlabStat is 0.5%; hence, if n is large and df_e is close to n, $n/df_e \sim 1$, a

premium of 0.5% will lead to an N of about 2.5758 and hence to

$$K = 1.40 + 0.85 \times 2.5758 = 3.5895$$

$$C = 3.5895 \left\{1 - \tfrac{2.72}{df_e}\right\}.$$

If the value of C computed in this way is smaller than 1.5, its value is set to 1.5.

6. Flag as outliers all the observations whose raw residual is greater than the threshold ($|r_{ijh}| > s^r CP$). If the square root of the effective error mean square ($\sqrt{\text{MSE}_{\text{Eff}}}$) is greater than the threshold, use the square root of the effective error mean square as threshold.

7. For the output report compute standardized residuals, r^s_{ijh}, as the ratio of the raw residual and the square root of the effective error mean square: $r^s_{ijh} = r_{ijh}/\sqrt{\text{MSE}_{\text{Eff}}}$. These r^s_{ijh} are only used for descriptive purposes in PlabStat, but not for outlier identification.

The calculation of the error degrees of freedom (df_e) is, in any case, straightforward. Residual degrees of freedom correspond to $n - \text{rank}(\mathbf{X}|\mathbf{Z})$, where n is the number of observations and $\mathbf{X}$ and $\mathbf{Z}$ are the design matrices for fixed and random effects, respectively, as defined for model (2.3).

Method M2: using Bonferroni-Holm test to judge studentized residuals as outliers (BH-SR)

The problem of outliers can be treated by one of several significance tests of a non-outlier null hypothesis against an alternative hypothesis [Hampel, 1985]. This simultaneous testing of several hypotheses implies the so-called multiple testing problem, whence the probability of finding at least one significant but spurious outlier by chance alone may be inappropriately large [Hochberg and Tamhane, 1987, p.7].

The general problem of multiple testing (not specifically in the context of outlier detection) was first approached by using the Boole inequality within multiple inference theory (known as classical Bonferroni test), which basically states that having a family-wise significance level α to test all n hypotheses, each individual test should be performed at an individual significance level of α/n [Holm, 1979]. This correction has been widely used but it is also criticized for being conservative.

The sequentially rejective Bonferroni test, also known as Bonferroni-Holm technique, is an improvement over the classical Bonferroni test in the sense that the test gains power as long as many hypotheses are completely wrong. Thus, Bonferroni-Holm test gives the same protection against Type I errors (to

falsely declare a null effect to be real or non-zero) but also reduces the probability of Type II errors (failing to declare a real effect) compared to the classical Bonferroni test [Holm, 1979].

For our first Bonferroni-Holm procedure for testing outliers, we used a studentized version of the residuals proposed by Nobre and Singer [2007] that does not depend on the scale. The authors present a thorough motivation on the use of this type of residuals to test if the ijhth observation is an outlier. The procedure for testing outliers is described in the following:

1. Compute absolute values of studentized residuals, r_{ijh}^{stu}. A studentized residual is defined as

$$r_{ijh}^{stu} = \frac{r_{ijh}}{s\sqrt{\hat{q}_{ijh,ijh}}},$$

   where r_{ijh} is the raw residual for the ijhth observation, s is the estimate of the error standard deviation (i.e., square root of the error variance estimate, $\sqrt{\hat{\sigma}_e^2}$) and $\hat{q}_{ijh,ijh}$ is an estimate of $q_{ijh,ijh}$, the ijhth diagonal element of a matrix $\mathbf{Q}$, whose diagonal elements are a function of the joint leverage of fixed and random effects [Nobre and Singer, 2007; Schützenmeister and Piepho, 2012]. Matrix $\mathbf{Q}$ can be obtained by decomposing $Var(\hat{\mathbf{e}})$ as $\sigma_e^2\Sigma\mathbf{Q}\Sigma$ (See appendix of Nobre and Singer [2007] for details).

2. Let H_o^{ijh} be the null hypothesis testing the absolute value of the ijhth studentized residual $|r_{ijh}^{stu}|$, i.e., H_o^{ijh} : $|r_{ijh}^{stu}|$ does not correspond to an outlying observation. Compute the p-value for each $|r_{ijh}^{stu}|$ assuming that the studentized residuals have an approximate standard normal distribution. The approximate p-value of r_{ijh}^{stu} equals $2\Phi(-|r_{ijh}^{stu}|)$, where Φ stands for the cumulative distribution function of the standard normal.

3. Test the no-outlier null hypothesis for each residual using the Bonferroni-Holm method (refer to Hochberg and Tamhane, 1987, p.57 for details).

The Bonferroni-Holm procedure can be easily implemented in SAS using the PROC MULTTEST options *inpvalues holm*. Studentized residuals are printed in the MIXED procedure by adding the option *residuals outp= file_name* to the *model* statement. The codes are presented in Appendix A.2.

Method M3: Studentized residual razor (SRR)

Another common practice that is widely used is to plot the studentized residuals against the fitted values and inspect the observations whose studentized residuals are located beyond a fixed threshold. The threshold is defined by the researcher and is based on experience; it may vary according to the percentage

of outlying observations identified. In practice, breeders sometimes try several fixed thresholds till the residual plots look acceptable and till no more than a certain percentage, e.g., 5%, of the observations are flagged. The choice of the threshold can be guided. Making the assumption that errors follow a normal distribution with zero mean and unit variance, if we use threshold $[-t_{SRR}, t_{SRR}]$, where t_{SRR} is the $(1-\alpha/2)$-quantile of the standard normal distribution, then we expect a proportion of α falsely flagged residuals when all observations meet the assumptions. Hereafter, we often use an exemplary threshold for $\alpha = 0.005$, which corresponds to a $t_{SRR} = 2.8$ (threshold $[-2.8, 2.8]$).

Method M4: using Bonferroni-Holm test to judge residuals standardized by the re-scaled MAD (BH-MADR)

Combining the strengths of the PlabStat procedure and the Bonferroni-Holm test may lead to an improved outlier detection method. Here, we suggest to use a robust estimate of the standard deviation, the re-scaled MAD, to effectively standardize the residuals and then use the Bonferroni-Holm test to decide for each observation whether it should be flagged as outlier or not. The BH-MADR method operates as follows:

1. Standardize raw residuals using re-scaled MAD (s^r).

$$r^M_{ijh} = \frac{r_{ijh}}{s^r}.$$

2. Proceed as for Method M2 (BH-SR) numerals 2 and 3 using r^M_{ijh}.

Method M5: using Bonferroni-Holm test to judge studentized residuals using a robust scale estimate (BH-STRO)

By this method we combine the principle of studentization with a robust scale estimate to standardize residuals and judge them independently using the Bonferroni-Holm test. Hereafter we refer to robust studentized residuals (r^{rs}_{ijh}) when a robust scale estimate, i.e., s^r, instead of the error standard deviation estimate (s) is used for studentization.

The method is described as follows:

1. Compute robust studentized residuals (r^{rs}_{ijh}) as:

$$r^{rs}_{ijh} = \frac{r_{ijh}}{s^r\sqrt{\hat{q}_{ijh,ijh}}},$$

where r_{ijh} is the raw residual for the ijhth observation, s^r is the re-scaled MAD and $\hat{q}_{ijh,ijh}$ is an estimate of the ijhth diagonal element of a matrix $\mathbf{Q}$, whose diagonal elements are a function of the joint leverage of fixed and random effects [Nobre and Singer, 2007; Schützenmeister and Piepho, 2012].

2. Proceed as for Method M2 (BH-SR) numerals 2 and 3 using r^{rs}_{ijh}.

2.3.4 Comparison of methods: $Premium$ vs. α_B and vs. t_{SRR}

The outlier detection method implemented in PlabStat does not have an explicit α, and although the *premium* is meant to protect against "bad observations", the *premium* is a completely different rate, defined as the percentage increase in estimation variance one is willing to pay if all residuals are from the same homogeneous distribution. Comparing the PlabStat procedure with the sequential Bonferroni-Holm may not be possible in terms of equating both thresholds as the stepwise Bonferroni-Holm method has a threshold changing with each step of the procedure. But for comparative purposes, equating the threshold of PlabStat (a function of the *premium*, the df_e and the total number of observations n) to the threshold of a classical Bonferroni test (a function of $\alpha_B = \alpha/n$) may give us an approximation of the values that the *premium* can take to produce the same threshold of a classical Bonferroni test. We propose to solve the *premium* for the α_B to which the classical Bonferroni test would correspond. A detailed description of the comparison extended to include the SRR threshold is provided in Appendix A.3.

2.3.5 ROC curves

A way to compare the methods that display different Type I error rates at the same given threshold is to plot the so-called receiver-operating characteristic (ROC) curves, which are useful for assessing the accuracy of binary predictions. ROC curves typically plot the true positive rate (TPR) against false positive rate (FPR) resulting from continuously varying the decision thresholds. On the y-axis is the true positive rate (TPR), which correspond to the proportion of the number of true outlying observations correctly declared as outliers, out of the total number of true outlying observations. On the x-axis is the false positive rate (FPR), defined as the proportion of the number of non-outlying observations falsely declared as outliers, out of the total number of non-outlying observations. A test with perfect discrimination has a ROC plot that passes through the upper left corner, where the TPR is 1.0, and the FPR is 0. The theoretical plot for a test with no discrimination (identical distributions of results for the true positives and the false positives) is a 45° diagonal line from the lower left corner to the upper right

corner. Qualitatively, the closer the ROC curve is to the upper left corner, the higher the overall accuracy of the test [Zweig and Campbell, 1993].

We used the original dataset of the triple lattice (Example 2.1) to illustrate the construction of the ROC curves for low-, medium- and high-contamination scenarios (Scenarios 1, 2 and 3, respectively). In each scenario we simulated pure shift outliers at one side [Lourenço and Pires, 2014; Rocke and Woodruff, 1996]. The observations were standardized, so that the mean of the response variable (yield) was fixed at $\mu = 0$. Outliers were generated from a $N(0, 1)$ distribution and then shifted d units. We used $d = 4, 7, 10$ and a contamination percentage of $2\%, 5\%$ and 10% . All the combinations of d and contamination percentage were generated but only representative scenarios were selected to show the results. For Scenario 1 we used $d = 4$ and contamination of 2%, for Scenario 2 $d = 7$ and contamination of 5%, and for Scenario 3 $d = 10$ and contamination of 10%. Each scenario was repeated 100 times and the average FPR and TPR by threshold point was computed and used to plot the ROC curves.

To plot the ROC curve for the PlabStat methods (M1r and M1f) we varied the threshold $\left(premium * \frac{df_e}{n}\right)$ between values very close to 0 and close to 1, i.e., starting from 10^{-15} and multiplying each step by a factor of $10/7$ until reaching 0.01 and from there progressing towards 1 with increments of 0.01, and released the restrictions of the method on the minimum value of C and the minimum threshold, which cannot be less than the effective error mean square. We used the threshold $\left(premium * \frac{df_e}{n}\right)$ to ensure that $N = -\Phi^{-1}\left[premium * \frac{df_e}{n}\right]$ can be computed (See method M1 and Appendix A.3 numeral A.3.1). For the methods using Bonferroni-Holm test (M2, M4 and M5), we varied the α_B from 0.01 to 0.99 with 0.01 increments. And for SRR (M3r and M3f), we used threshold values $|t_{SRR}|$ ranking from 0 to 5 increasing by 0.1. The area under the curve (AUC) was computed using the trapezoids approach, i.e., the area of one trapezoid was calculated as the distance between two consecutive false positive rates multiplied by the average of the corresponding true positive rates, and then all trapezoid areas were added up. For comparison among the methods we pinpointed the values that the threshold takes at FPR=5% and TPR=95%.

Additionally, knowing the threshold values at fixed rates, we produced again 100 simulations of outliers in the three scenarios and compared one rate at the fixed level of the other across methods.

2.3.6 Special case: genomic prediction of a rye multi-environment trial using different outlier detection methods

To evaluate all the outlier detection methods using a large empirical dataset, we used a rye MET (Example 5.1) and validated the results through genomic prediction analysis by comparing the final predictive abilities of each method. The rye MET was carried out during years 2009 to 2011 at several locations of

Germany and Poland. The aim of the MET was the selection of promising rye genotypes. The trials were laid out as α-designs. In the first year, sets of 320 genotypes were evaluated in each trial and testcrossed with two different testers most of the trials in different locations. At least one location in each country contained a set of genotypes testcrossed with the two testers. Series of trials were conducted at each location and were connected through common checks. Approximately 10% of the evaluated genotypes in the first year were forwarded to a second test in the next year together with more entries. The same testers were used and some locations were shared with the previous year tests. A further selection of 10% of the genotypes was performed in the third year using four different testers and more locations. Thus, the first two years were connected through testers, locations and genotypes and the third year was connected with the other years through genotypes and locations. A total of 908 genotypes were evaluated across the three years and 826 had molecular markers information (a more extensive analysis of this dataset is presented in Bernal-Vasquez et al. [2014]).

The available marker information obtained using a 16K Infinium iSelect HD Custom BeadChip of the selected genotypes was used to perform GP. We performed a stage-wise analysis, with a pre-processing step, where we flagged and dropped detected outliers at the trial level using model (2.1). We used the five methods defined before under the GP approach plus the complete dataset without removing outliers as control (labeled as Complete set), and another dataset where we removed manually only the observations reported by the breeders as having problems in the field (labeled as Manual).

Subsequently, in stage 1, genotypic adjusted means by location by year ($m^{(1)}_{hrsv}$) are computed using the model

$$y_{hijkv} = \mu + (gt)_{hv} + s_k + \gamma_{ik} + b_{ijk} + e_{hijkv}, \tag{2.8}$$

where y_{hijkv} represents the observation of the hvth genotype-tester combination of the jth block within the ith replicate of the kth trial, μ is the general mean, $(gt)_{hv}$ is the effect of the hth genotype testcrossed with the vth tester, s_k is the effect of the kth trial, γ_{ik} the effect of the ith replicate within the kth trial, b_{ijk} the effect of the jth block within the ith replicate of the kth trial and e_{hijkv} the error associated with the observation y_{hijkv}. For simplicity, we omit the subscripts for year and location in model (2.8), but it should be understood that all terms of the model are indexed by location (s) and year (r).

In stage 2, we used the adjusted means of the previous stage to be analyzed across locations and years using the model

$$m^{(1)}_{hrsv} = \mu + g_h + l_s + (at)_{rv} + (ga)_{hr} + (gat)_{hrv} + (gl)_{hs} + (la)_{rs} + (lat)_{rsv} + (gla)_{hrv} + (glat)_{hrsv} + e_{hrsv}, \tag{2.9}$$

where $m_{hrsv}^{(1)}$ represents the adjusted mean for the hth genotype with the vth tester within the rth year in the sth location. The model contains the general mean μ, the main effects of genotypes (g_h), testers within years $[(at)_{rv}]$ and locations (l_s), the two-way and three-way interaction effects and the error e_{hrsv}, which is confounded with the three-way interaction. To overcome this loss of information due to the confounding, we weighted the adjusted means from the first stage by the diagonal elements of the inverse of the variance-covariance matrix of the adjusted means of the first stage [Smith et al., 2001]. In the second stage we computed one adjusted mean for each genotype.

The last stage was the implementation of GP using the model

$$\mathbf{m^{(2)}} = \mathbf{1_n}\mu + \mathbf{Zu} + \mathbf{e}, \tag{2.10}$$

where $\mathbf{m^{(2)}}$ represents a $n \times 1$ vector of adjusted means for genotypes across locations and years, $\mathbf{1}_n$ is a $n \times 1$ vector of ones, μ is the overall mean, $\mathbf{Z}$ is the markers matrix for random effects, $\mathbf{u}$ is the vector of random effects, i.e., the SNP effects. It is assumed that $\mathbf{u}$ has a normal distribution with zero mean and variance matrix $\mathbf{I}\sigma_u^2$ $[\mathbf{u} \sim N(0, \mathbf{I}\sigma_u^2)]$ and the error has a normal distribution with zero mean and variance matrix $\mathbf{R}$ $[\mathbf{e} \sim N(\mathbf{0}, \mathbf{R})]$. $\mathbf{R}$ is a diagonal matrix with diagonal elements equal to the inverses of the diagonal elements of the inverse of the original variance-covariance matrix of the adjusted means of the second stage [Smith et al., 2001]. Specifically, the GEBV of the hth genotype corresponds to the estimate of the genotype effect $\hat{g}_h = \sum_{p=1}^{W} z_{hp}\hat{u}_p$, with $p = 1, \cdots, W$ where W is the number of markers, u_p is the effect of the pth marker and z_{hp} is the SNP genotype of the pth marker for the hth genotype.

The predictive ability of GP was assessed using 10 replications of fivefold cross validation (CV), where the dataset was randomly split in 5 subsets 10 times. In each time, i.e., each replicate, groups of 4 subsets were used to estimate the parameters of the model and predict the observation of the fifth subset, so that each subset is predicted using the other 4 subsets. Thus, in the end, we had 50 prediction runs. The predictive ability of each of the ten repeats of the fivefold CV was computed as the Pearson correlation coefficient between observed values and the predicted GEBV. The estimate of the predictive ability of the method (ρ_{GP}) corresponds to the average of the correlation coefficients of the ten repeats.

2.4 Results

2.4.1 Comparison of variance estimates for PlabStat and REML-based analysis with all-cells-filled data of the published GL examples

The datasets where all observations were available are referred to as all-cells-filled data. A cell is defined by the intersection of a genotype by replicate classification [Searle, 1987, p.8]; thus, the data used in this case correspond to the original datasets of the examples and the datasets using three artificial outlying observations each where all cells of the replicate-by-genotype classification were filled (Examples 1.1, 1.3, 2.1, 2.3, 3.1, 3.3, 4.1 and 4.3). A comparison of the variance component estimates obtained for each example is given in Table 2.3. The variance components obtained by SAS-ANOVA were identical to the values obtained by PlabStat.

Table 2.3: Comparison of variance component estimates for SAS-REML and PlabStat for datasets with all cells of the replicate-by-genotype classification filled (cases where all observations were available).

	SAS-REML		PlabStat[a]	
Example	σ_b^2	σ_e^2	σ_b^2	σ_e^2
1.1	0.062	0.085	0.059	0.084
1.3	0.000	7.458[b]	−1.191	8.493
2.1	0.040	0.265	0.040	0.265
2.3	0.050	2.491	0.051	2.491
3.1	19.630	13.655	19.630	13.655
3.3	21.405	31.293	21.405	31.293
4.1	~ 0	~ 0	10.472	10^{-4}
4.3	2.621	13.332	3.054	12.938

[a] Results are identical to SAS-ANOVA
[b] Using the NOBOUND option in PROC MIXED to allow negative variance estimates, $\sigma_b^2 = -0.5419$ and $\sigma_e^2 = 7.9150$

Variance component estimates via ANOVA (PlabStat and SAS-ANOVA) and REML (SAS-REML) were expected to differ because of the planned unbalancedness of the GL designs, but they were also expected to be very close. In the case of balanced designs yielding positive variance components estimates, both approaches (ANOVA and REML) coincide. When all cells of the replicate-by-genotype classification were filled, the biggest differences occurred where a negative or a zero value of the block variance component was estimated. The negative estimates can only occur in SAS-REML if the user allows estimation of negative components. This can be done by using the *nobound* option of the MIXED procedure (See Appendix A.2).

In the α-design with three outliers (Example 1.3), we found $\hat{\sigma}_b^2 = 0$ in SAS unless the *nobound* option was used to allow negative variance estimates in which case $\hat{\sigma}_b^2 = -0.5419$; whereas in PlabStat $\hat{\sigma}_b^2 = -1.1911$. The zero block variance means that there is no correction due to incomplete blocks, i.e., the BLUPs of block effects are zero ($\hat{b}_{ij} = 0 \; \forall \; i, j$) as if the design were a randomized complete block design (RCBD) [Cochran and Cox, 1957], with complete blocks corresponding to replicates of the α-design. The original data from the rectangular lattice (Example 4.1) were generated to force the error variance estimate to be exactly zero, which makes the likelihood equal infinity; therefore, in PlabStat, $\hat{\sigma}_e^2$ is set to 0.0001, making the analysis feasible. In the rectangular lattice with three outliers (Example 4.3), PlabStat estimated the variance component of replicates as -0.1310, and because of the unbalancedness, estimators for blocks and error are dissimilar. Despite these differences, the raw residuals, the block effect estimates and the treatment adjusted means were highly correlated between both procedures (correlation ≈ 0.98).

The negative variance estimates in some of the ANOVAs arose due to the presence of the three outliers in these small datasets. In practice, negative variance estimates are a warning of an incorrect model or statistical noise obscuring the underlying analysis [Thompson, 1962]. Using REML, estimates are constrained to be non-negative and set to zero when iterations reach the boundary. Consequently, an automatic solution is to use REML. Zero estimates with REML can be handled, e.g., dropping the corresponding effect from the model and re-estimating the others [Searle et al., 1992], providing starting values of variances as low boundary constrains [Littell et al., 2006], or by regularization of the estimates using a Bayesian approach [Barnett and Lewis, 2000].

2.4.2 Comparison of variance estimates between PlabStat, ANOVA and REML-based analysis using data with missing observations of the published GL examples

Since PlabStat imputes the missing values, the analyses of the example with missing observations are expected to yield different results and more dissimilar outputs. For those datasets (Examples 1.2, 2.2, 3.2 and 4.2) the results were slightly different among the three procedures, i.e., PlabStat, SAS-ANOVA and SAS-REML. Despite the differences, the Pearson correlation coefficients for outputs of residuals, BLUPs of block effects and adjusted treatment means between PlabStat and SAS-REML were high (Table 2.4), reflecting the great similarity between the two approaches.

In Example 4.2 (rectangular lattice with three missing observations), PlabStat imputes the missing observations with the predicted observations fitting the model as if incomplete blocks were a fixed effect. The analysis with this value imputation results in a zero or very close to zero estimate of the error variance as in the original data (Example 4.1). This case differs the most between PlabStat and SAS-ANOVA,

Table 2.4: Comparison of variance component estimates for PlabStat (PS), SAS-REML (REML) and SAS-ANOVA (ANOVA) for datasets with missing observations.

	PS		REML		ANOVA		Correlations PS - REML		
Example	σ_b^2	σ_e^2	σ_b^2	σ_e^2	σ_b^2	σ_e^2	Residuals	Adjusted treatment means	BLUPs of block effects
1.2	0.061	0.084	0.066	0.084	0.063	0.084	0.998 13	0.999 48	0.986 54
2.2	0.043	0.268	0.042	0.268	0.042	0.268	0.999 51	0.984 18	0.999 94
3.2	15.860	8.244	15.543	8.198	15.917	8.244	0.990 48	0.998 15	0.998 96
4.2	10.470	10^{-4}	10.166	10^{-8}	8.707	10^{-12}	–[a]	0.999 99	0.999 92

[a] Non-estimable because PlabStat residuals are set to zero

which does no imputation. The consequence of the zero residual variance in the REML approach is again that the likelihood equals infinity and the analysis is not feasible unless the error variance is set to a very low positive non-zero value. In PlabStat this value is automatically set to 0.0001 and in SAS the parameter must be held at a tiny value, e.g., 10^{-8}. The BLUPs of the block effects and the adjusted means of PlabStat and SAS-REML are almost identical. The residuals cannot be correlated because PlabStat produces zero values and SAS very low values for the residuals close to zero at around $\pm 10^{-10}$.

Moreover, for all the examples with missing data, the flagged outliers using SAS-REML were the same as the ones PlabStat identified. The comparison of the re-scaled MAD and the thresholds obtained for all the examples with missing observations using PlabStat, SAS-ANOVA and SAS-REML are presented in Table A.5.

2.4.3 Comparison of outlier detection methods using data with artificial outliers of the published GL examples

The scatter plots of the different types of residuals against predictions according to the method of outlier detection are depicted in Figs. 2.1, A.2, A.3 and A.4. For the purpose of comparing all the methods graphically, solid reference lines represent the procedures with fixed thresholds (M1, M3) and dashed reference lines the procedures with varying thresholds (M2, M4, M5) showing the threshold for the largest residual. For each method we plotted raw residuals against predictions to be able to visually compare the methods.

The number of detected outliers varied between methods and between approaches using incomplete blocks as random or fixed effects. Within methods using random blocks, M5r selected always the greater number of possible outliers, followed by M1f. M4r selected an intermediate number of outliers and

M2r and M3r produced the fewest detections. This tendency was similar for the methods that used fixed blocks and in general "f-methods" identified fewer residuals than "r-methods".

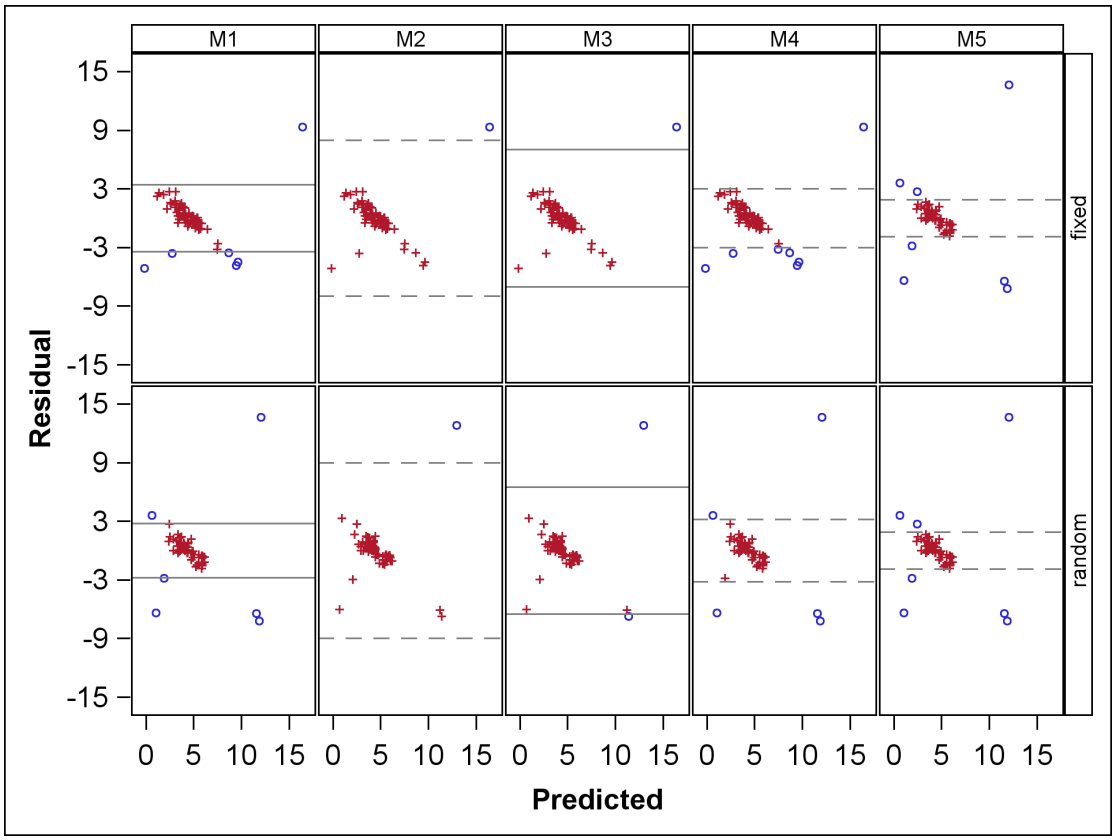


Figure 2.1: Scatter plots of raw residuals vs. predictions for the α-design with three outlying observations (Example 1.3) using PlabStat outlier detection method (M1), Bonferroni-Holm test using studentized residuals (M2), Studentized residual razor (M3), Bonferroni-Holm test using re-scaled MAD to standardize residuals (M4), and Bonferroni-Holm test using the robust studentized residuals (M5). In the *first row*, methods used fixed incomplete block effects and in the second row methods used random incomplete block effects. *Solid reference lines* are used for methods with fixed thresholds and *dashed reference lines* for methods with varying thresholds representing the threshold calculated for the largest residual. Flagged outliers are indicated with an *empty circle* and non-suspicious observations with a *cross*.

2.4.4 Comparison of ROC curves

Figs. 2.2, A.5 and A.6 show the ROC curves under scenarios 3, 2 and 1, respectively. Some ROC curves were not intact because the number of simulated outliers is finite and relatively small (i.e., for 10% contamination a maximum of 25 outliers are simulated); therefore, some of the computed thresholds do not cover the full possible range of values of FPR and TPR. For this reason, we think that comparing the AUC among curves may not be fair. Scenario 3 produced the highest number of complete curves (but still with some incomplete curves), where the AUC had no trend between fixed and random methods

(Appendix A.3 Table A.6). We added cut points at fixed FPR and TPR for comparison purposes and displayed the threshold value in each case. For FPR$= 5\%$, "r-methods" had slightly larger TPR than "f-methods". Similarly, for TPR$= 95\%$, "r-methods" had smaller FPR. This could be also observed in the simulations using fixed rates (Figs. A.7, A.8) especially for scenario 1 and 3. A substantial difference among the methods within scenarios is not immediately perceptible but a difference between scenarios can be easily appreciated.

2.4.5 Comparison of outlier detection methods for a genomic prediction analysis using a rye MET

The complete dataset was composed of a total of 25632 observations on 908 genotypes. The residual pattern of one trial looked quite unusual due to some observations that had a different yield, leading to segregated residuals. In Fig. 2.3 we depict the residual plots using the five methods using random and fixed block effects and the "manual" method (displayed with the fixed block effect methods only for ease of graphical comparison), which shows the residuals removed manually. Either the two clearly separated clouds suggest that there is a systematic effect in the data that has not been accounted for, or some severe bias is present, perhaps due to a problem with the part of the plots of that particular trial. Indeed, the breeders reported a herbicide drift problem in the outer row of the field (Fig. A.9). These observations were the ones removed manually.

To minimize the comparison error among outlier detection methods, we kept the seed of the random number generator of the CV procedures fixed so that the exact same stream of random numbers was used for each CV. Predictive abilities were calculated for each dataset resulting from the outlier detection methods, for the dataset with manual removals and for the complete set (Table 2.5). Additionally, a paired t-test using the least significant difference (LSD, $\alpha = 5\%$) was carried out to compare the predictive abilities of all datasets. We used a randomized complete block model considering each repetition of the CV as a block, thus accounting for the dependence among observations from the same sample. Statistically significant differences were detected across the methods. M4r and the manual outlier removal had the highest predictive abilities and M2f, M3r and M3f were not significantly different from the control, i.e., the complete set. The other methods performed slightly better than the control, although the improvement was not dramatic. The methods using random block effects were classified better than their homologous method with fixed block effects and, interestingly, the same methods with fixed or random always were neighbors in the ordered list (Table 2.5).

We could not identify a trend between the number of removed observations (flagged outliers) and the predictive abilities of the methods, i.e. method M4r removed 99 observations and was classified as

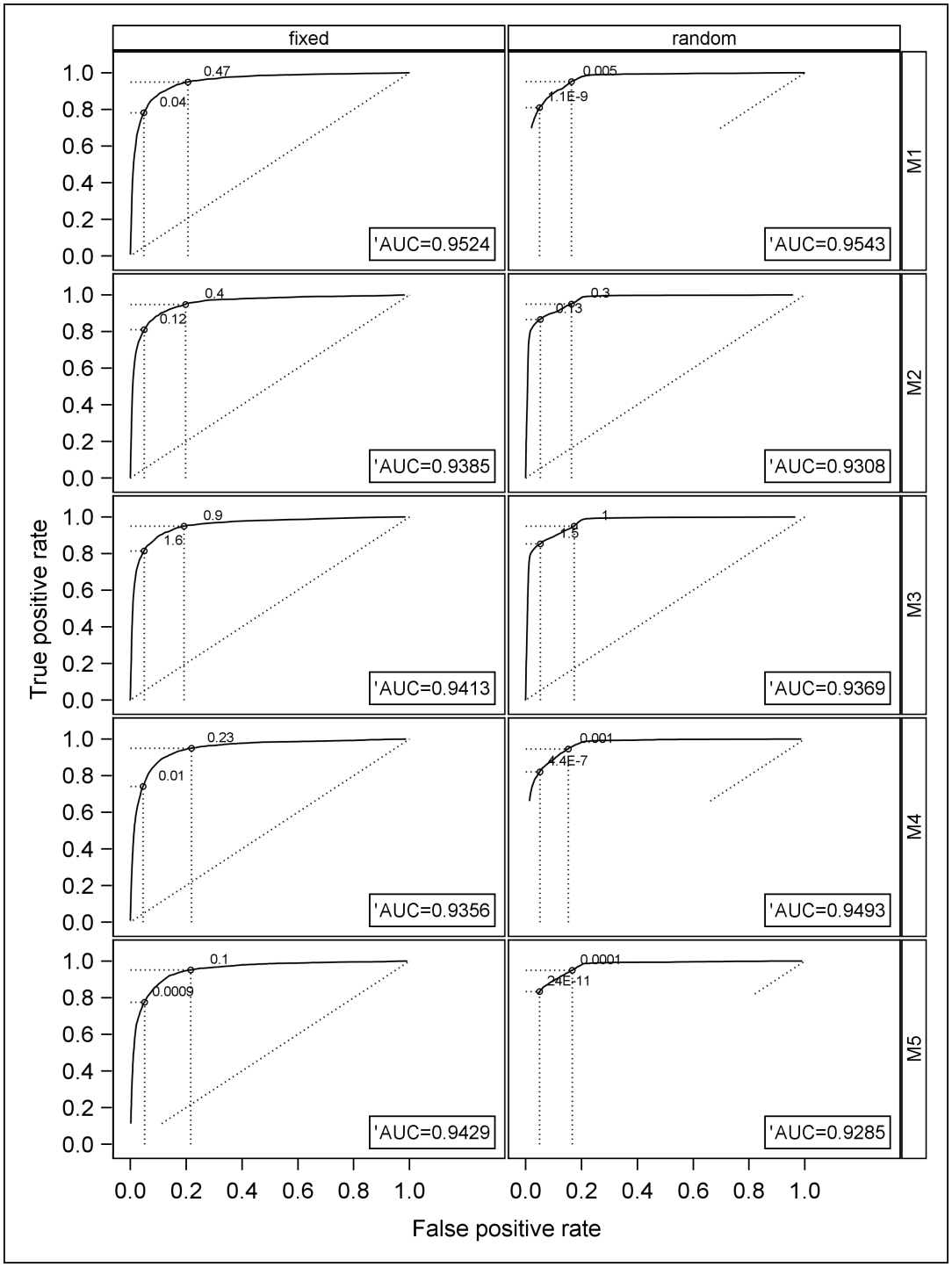


Figure 2.2: ROC curves of all methods assuming fixed (*first column*) and random (*second column*) incomplete block effects under a scenario with 10% contamination and 10 deviation units from the mean (Scenario 3). Methods used were PlabStat (M1) outlier detection method, Bonferroni-Holm test using studentized residuals (M2), Studentized residual razor (M3), Bonferroni-Holm test using re-scaled MAD to standardize residuals (M4) and Bonferroni-Holm test using robust studentized residuals (M5).

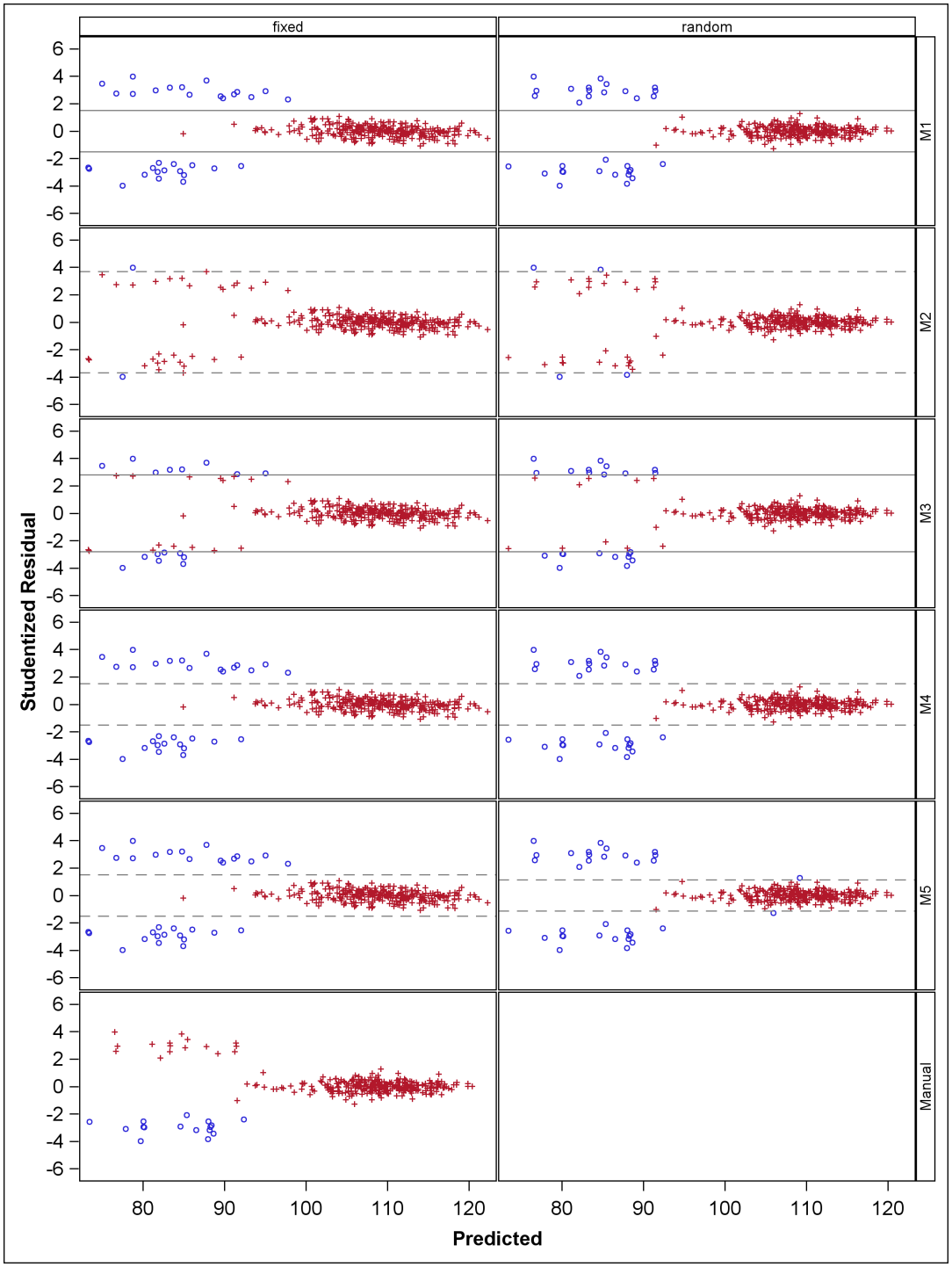


Figure 2.3: Scatter plots of studentized residuals vs. predictions for one unusual trial of the rye MET. Methods from the *first column* of the panel considered incomplete blocks as fixed effects and in the *second column* methods that considered incomplete blocks as random effects. Methods used were PlabStat (M1) outlier detection method, Bonferroni-Holm test using studentized residuals (M2), Studentized residual razor (M3), Bonferroni-Holm test using re-scaled MAD to standardize residuals (M4), Bonferroni-Holm test using robust studentized residuals (M5) and Manual removal, which is displayed with the fixed block effect methods only for graphical comparison purposes. *Solid reference lines* are used for methods with fixed thresholds and *dashed reference lines* for methods with varying thresholds representing the threshold calculated for the largest residual. Flagged outliers are indicated with an *empty circle* and non-suspicious observations with a *cross*.

accurate as the manual outlier removal, which took away only 20 observations (Table 2.5). Therefore, a general overview of the flagged outliers may help to understand the results (Fig 2.4). M1 and M5 identified almost all the peripheral observations and included quite a few observations that belong to the main cloud. M2 identified the smallest number of peripheral observations and M3 did not flag all peripheral observations but some observations from the main cloud. M4 showed a consistent identification of residuals from the periphery and not affecting the main cloud.

Table 2.5: Predictive abilities (ρ_{GP}) in the GP stage and number of outliers removed using the entire dataset (Complete set), the dataset with manually removed observations (Manual) and the methods of outlier detection: PlabStat with fixed and random block effects (M1f, M1r), Bonferroni-Holm using studentized residuals with fixed and random block effects (M2f, M2r), studentized residual razor with fixed and random block effects (M3f, M3r), Bonferroni-Holm using re-scaled MAD with fixed and random block effects (M4f, M4r) and Bonferroni-Holm using robust studentized residuals with fixed and random block effects (M5f, M5r). Correlations followed by a common letter are not significantly different ($\alpha = 5\%$) according to the LSD test.

Method	ρ_{GP}	Number of outliers removed
M4r	0.6124^{a}	99
Manual	0.6115^{a}	20
M4f	0.6103^{b}	93
M5r	0.6098^{b}	557
M5f	0.6079^{c}	702
M1r	0.6072^{cd}	422
M1f	0.6063^{de}	440
M2r	0.6058^{e}	67
M2f	0.6046^{f}	64
Complete set	0.6037^{f}	0
M3r	0.6036^{f}	234
M3f	0.6036^{f}	219

2.5 Discussion

In this work, we have demonstrated through some examples that the outputs of an analysis via ANOVA are not far away from what is obtained using REML. We also reviewed the underlying theory of both approaches so that the reader can appreciate where any differences come from. We used a common model to analyze all the datasets, since our goal was to resemble the analysis implemented in PlabStat and not to evaluate which model fitted best. In general, the results obtained using PlabStat based on ANOVA estimation were very similar to the ones obtained using SAS based on REML estimation. Other authors

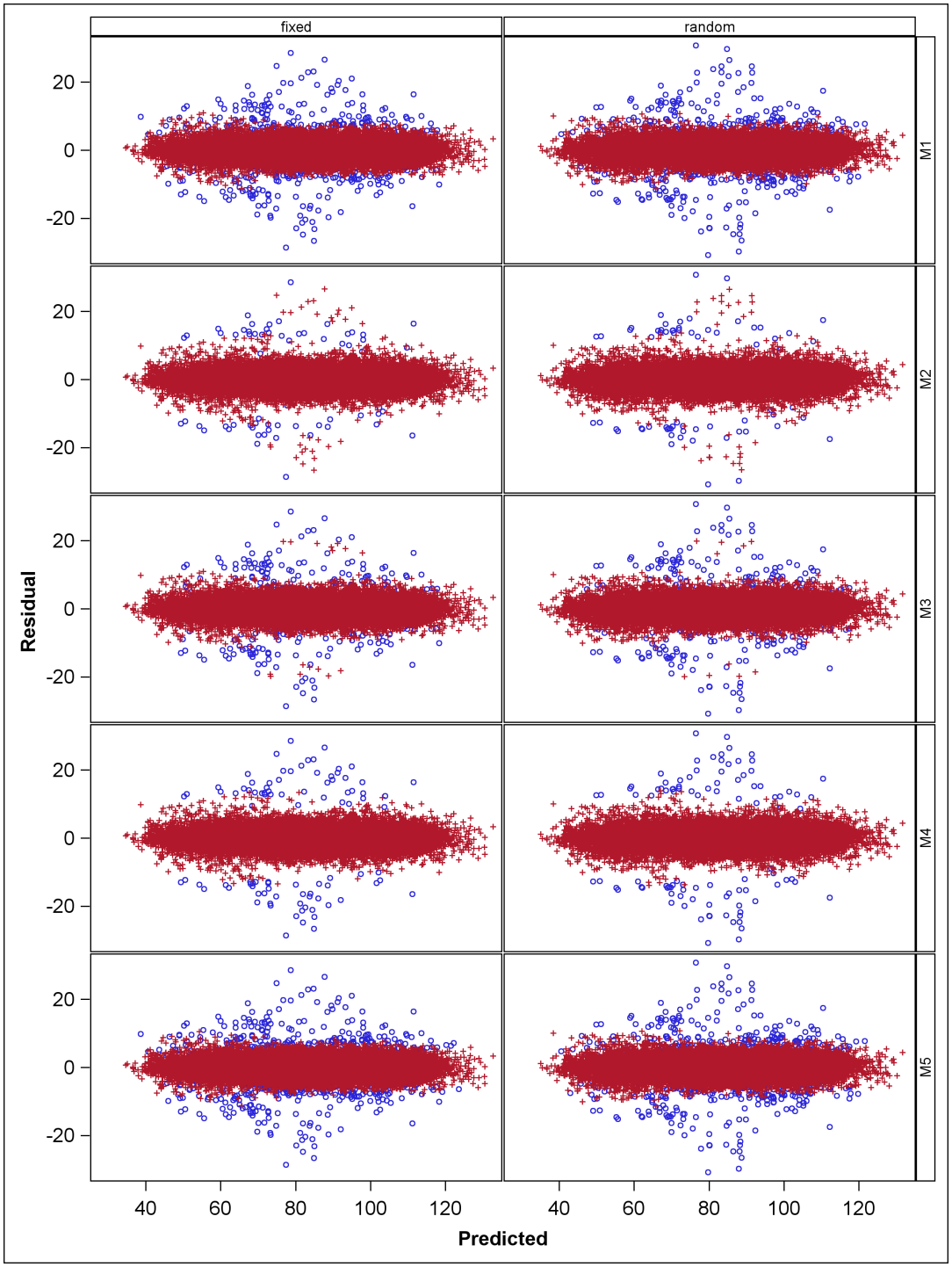


Figure 2.4: Overview of flagged outliers across all the dataset. Methods from the *first column* of the panel considered incomplete blocks as fixed effects and in the *second column* methods considered incomplete blocks as random effects. Methods used were PlabStat (M1) outlier detection method, Bonferroni-Holm test using studentized residuals (M2), Studentized residual razor (M3), Bonferroni-Holm test using re-scaled MAD to standardize residuals (M4) and Bonferroni-Holm test using robust studentized residuals (M5). Flagged outliers are indicated with an *empty circle* and non-suspicious observations with a *cross*.

[Wensch et al., 2013; Wulff, 2008] have also reviewed the properties of ANOVA and REML estimators under unbalanced data and showed their similarities. A noticeable difference between PlabStat and REML-based packages is that in PlabStat missing values are imputed, whereas other packages do not do imputation.

Searle et al. [1992] and Littell [2002] present extensive reviews comparing variance estimation methods. In most of the cases, they prefer REML over ANOVA. The analysis using REML has some technical advantages: REML has no problems with missing observations, so long they are missing at random, while PlabStat, using ANOVA, employs data imputation; goodness-of-fit measures are available (using the likelihood) using REML. Furthermore, REML can handle correlated random effects such as time-varying effects using a variance-covariance matrix that accounts for the serial correlations [Searle et al., 1992]. A special case relevant for breeders is the possibility of modelling genotypic or environmental correlated effects. The ANOVA method only allows estimating simple random effects models, whereas REML allows a more flexible variance-covariance structure to accommodate heterogeneity of variances and genetic correlations among environments, e.g., factor-analytic (FA) models [Meyer, 2009]. More recently, modelling genetic correlations using pedigree and marker relationship matrices under REML frames has been shown to lead to more accurate genomic prediction models [Burgueño et al., 2012; Lopez-Cruz et al., 2015]. We assessed the outlier detection method implemented in PlabStat, but instead of using the ANOVA approach as PlabStat does, we used REML and obtained the same flagged residuals. We showed then that the PlabStat outlier detection implementation produces the same results (or very similar) within a REML-based environment. For breeders whose experience with PlabStat has certified the quality of its analyses, the fact that we demonstrate that results using REML lead to the same conclusions, hopefully gives some confidence that the transition from PlabStat to REML-based packages can be made safely. Given the mentioned advantages from REML over ANOVA estimates, we recommend using REML approach.

The rationale behind the outlier detection methods varies in terms of the error rates controlled. The method used in PlabStat (M1) uses the *premium*, which is "charged" to the error variance to protect against bad observations, the Bonferroni-Holm methods, i.e., M2, M4 and M5, use the adaptive adjustment for each p-value of each residual (in turn standardized differently) in order to deal with the multiple testing problem, and method M2, the Studentized residual razor (SRR), uses a fixed threshold derived from the standard normal distribution. Likewise, the outlier detection method implemented in PlabStat depends on the ratio df_e/n, the Bonferroni-Holm methods account for the sample size n, and SRR relies only on the studentization of the residuals to detect a expected proportion of outlying observations. M4 and M5 involve a mixture of the features from PlabStat, from studentization and from the Bonferroni-Holm test. Despite these differences among the methods, we were able to show how a family-wise

significance level $\alpha_B = \alpha/n$ from the classical Bonferroni test could be adjusted to correspond to a given *premium* from the PlabStat method (Fig. A.1a) in order to flag the same outlying observations.

The SRR method could also be expressed as a function of the *premium* by finding the corresponding threshold t_{SRR} for a given α_B. In Fig. A.1b the SRR method would select more outliers than a classical Bonferroni threshold and, compared with PlabStat using *premium*$= 0.005$, a higher t_{SRR} would be needed. This behavior may be attributed to the differences between the re-scaled MAD and the studentization denominator. These two standardization approaches surely play a big role in outlier identification since they lead to somewhat different standardized residuals, accounting for the differences among the outlier detection methods. On the one hand, studentized residuals are suitable to check for outlying observations, homoscedasticity and normality of residual errors [Schützenmeister and Piepho, 2012] and the use of the leverage in the studentization approach makes the studentized residual reflect the change of the ijhth fitted value with respect to the ijhth observed value [Nobre and Singer, 2011]. This property is advantageous in the case the covariance matrices are correctly specified. This was probably the reason why SRR showed good TPR and low FPR in scenarios with low and medium outliers contamination. On the other hand, using re-scaled MAD produces larger residuals avoiding the inflation of the estimated standard deviation caused if there are outliers in the data [Swallow and Kianifard, 1996]; thus, robust standardized residuals in combination with the Bonferroni-Holm test allows only exceptionally large residuals to be judged as outliers.

The ROC curves within "f-methods" and within "r-methods" did not provide enough evidence that one method was better than the others. This is, in a sense, advantageous, because all methods may have a similar potential to identify outliers. Outlier simulations fixing the false positive and true positive rates showed that "f-methods" had slightly lower TPR than "r-methods", but markedly higher FPR specifically for scenarios 1 and 3. This behavior may be due to the fact that residuals calculated under models that contain other random effects (different than the error) may have a *supernormal* distribution [Nobre and Singer, 2007; Schützenmeister and Piepho, 2012]. The optimal threshold for each method depends on how large a TPR and FPR we want to admit. We provide cut points along the ROC curves for all methods, showing that the same or very similar results (in terms of number of outlier identified) can be achieved by any method by adjusting the parameters controlling the thresholds. Nevertheless, we can not recommend any particular value for any of the thresholds because these values are specific for each case. When possible, ROC curves can be used to exactly define the threshold parameters that lead to a desired FPR and TPR. Since our objective is to propose an automatic method to identify possible outlying observations, the methods may be compared with the default parameter settings, i.e., *premium* in PlabStat of 0.05%, Bonferroni-Holm α of 0.05 and t_{SRR} of 2.8. Considering the results of the ROC curves (Figs. 2.2, A.5, A.5) and the simulations with fixed FPR and TPR (Figs. A.7, A.8) where no marked or clear

differences can be observed among methods within each scenario, we do not have enough evidence to recommend a specific method. Methods that take into account a protection against multiple testing error (M2, M4 and M5) may have a theoretically founded benefit over the other methods.

The statistical differences observed in the GP analysis showed that manually removing observations for biological reasons (as in the Manual method) yielded the highest predictive ability statistically equivalent to the predictive ability of M4r, also classified as the best method. These results are surprising given the variable number of outliers identified (99 for M4r and 20 for Manual). The overview of the flagged/removed observations by all methods (Fig. 2.4) indicate that methods M4 and M5 (independent of the fixed or random block effects) led to removing the observations detaching the main cloud, however M5 picked more observations within the main cloud. Now, the fact that "r-methods" had higher ρ_{GP} than "f-methods" may be explained by efficiency gained from using random incomplete blocks over fixed incomplete blocks when the number of blocks is higher than 10 [Cochran and Cox, 1957]. The predictive abilities for "r-methods" vs. "f-methods" are in accordance with the observations derived from the ROC curves and simulations with fixed FPR and TPR. M4r clearly had a higher predictive ability followed by M5r, whereas M3 methods were penalized with the lowest predictive ability. The reason may be that the protection against multiple testing error benefited the performance of methods M4 and M5.

The consequence of dropping one observation has stronger effects in small datasets than in big datasets in terms of increasing the unbalancedness. PlabStat seemingly avoids generating more unbalancedness by imputing the dropped data points with the prediction of the observation, taking the block effect as fixed but one must bear in mind that imputed observations are not equivalent to observed data. There are now better methods to deal with missing data than imputation, i.e., using REML. Another option for detecting outliers in small experiments may be a Bayesian approach, e.g., using previous experiments to derive *a priori* information to get plausible values of the variance estimates of the model [Barnett and Lewis, 2000]. Other approaches entail, for example, tetrads [Bradu and Hawkins, 1982] or Bootstrapping [Marubini and Orenti, 2014]. For regression models, robust approaches have been used successfully for outlier identification [Cerioli et al., 2013; Lourenço and Pires, 2014; Marubini and Orenti, 2014; Swallow and Kianifard, 1996] since they can be more efficient on reducing the masking and swamping effect of the outliers in the residuals.

Outlier detection methods for LMM such as VSOM [Gumedze et al., 2010] have been successfully implemented and used in different fields [Babadi et al., 2014; Gumedze and Chatora, 2014; Gumedze and Jackson, 2011]. This method is advantageous for cases where powerful computing resources are available. We note that for n observations the method requires fitting of n mixed models, one for

each observation in turn, which poses higher demands on computing time than the simpler methods we consider. We did not consider the VSOM method since we focused more on a simple and easy-to-use outlier detection approach that raises a warning on possible spurious observations.

The numerical differences across methods were small in the rye example studied, even when no outliers were removed or when too many observations were removed. One of the reasons could have been that the dataset we used was huge (25,632 observations), thus removing 20 or 99 observations did not have a strong impact on the predictive abilities. An additional test using two smaller subsets of the complete rye MET, i.e., only one country in one year (7,680 observations each) and comparing only "r-methods", revealed more sensitivity among predictive abilities depending more on the type of outliers rather than on the number of observations detected as outliers. For the Polish dataset, where there was no trial with manifest outliers (thus no manual removal), predictive abilities ranged from 0.5579 to 0.5665 and the number of identified outlying observations from 1 to 69. These numerical differences are not practically relevant. By contrast, for the German dataset, where breeders reported the trial with the herbicide drift problem, predictive abilities of methods that removed all those spurious observations (M1r and M5r) were the highest with predictive abilities of 0.5922 and 0.5896 with 121 and 180 identified outlying observations, respectively. Methods that kept all (Complete set) or some of those observations (M2r and M3r) produced the lowest predictive abilities (0.5549, 0.5613, 0.5655 with 0, 20 and 82 observations detected as outliers, respectively), and the Manual method and M4r yielded intermediate predictive abilities of 0.5781 and 0.5844 with 20 and 54 identified outlying observations, respectively. Estaghvirou et al. [2014] found that a single outlier can have a marked effect on the estimation of accuracy and heritability of genomic prediction. They advise to check and eliminate outliers whenever possible to maximize the phenotypic variance to be captured. We therefore encourage the use of an outlier detection method that helps with the identification of spurious observations. We recommend to have in mind the strengths and weaknesses of the selected methods depending on the size of the dataset and the purpose of the analysis. In this work we reviewed several outlier detection methods, we dropped the flagged observations and studied what the consequences were. In practice, the deletion/replacement of an observation should be supported by subject matter knowledge about the trials.

Considering that errors do occur, using an outlier identification procedure in routine analysis is an insurance policy that helps to detect obvious outlying observations that may escape eye scrutiny. In our view, the PlabStat method, with the default threshold, is very powerful and sensitive in detecting conspicuous observations whose residuals are slightly detached from the main cloud. The Bonferroni-Holm based approaches are simple, easy to program and implement and theoretically well founded. The SRR approach based on empirical experience demonstrated to be also helpful as outlier identification method, although the decision on which threshold should be used is always somewhat arbitrary and the

method does not account for the multiple testing problem. All the methods showed similar performance in terms of false and true positive rates across simulated scenarios containing outliers and thus we do not recommend one specific method. One of our favorites is BH-MADR (M4), produced as the combination of a robust scale estimate to standardize residuals and a test that deals with the multiple testing problem. This method displayed good properties as outlier detection method under a genomic prediction application.

Chapter 3

The importance of phenotypic data analysis for genomic prediction - a case study comparing different spatial models in rye[2]

Angela-Maria Bernal-Vasquez[a], Jens Möhring[a], Malthe Schmidt[b], Manfred Schönleben[c], Chris-Carolin Schön[c] and Hans-Peter Piepho[a]

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

[b] KWS-LOCHOW GMBH, Ferdinand-von-Lochow-Strasse 5, 29303 Bergen, Germany

[c] Plant Breeding, Technische Universität München, Liesel-Beckmann-Strasse 2, 85354 Freising, Germany

3.1 Abstract

Background: Genomic prediction is becoming a daily tool for plant breeders. It makes use of genotypic information to make predictions used for selection decisions. The accuracy of the predictions depends on the number of genotypes used in the calibration; hence, there is a need of combining data across years. A proper phenotypic analysis is a crucial prerequisite for accurate calibration of genomic prediction procedures. We compared stage-wise approaches to analyse a real dataset of a multi-environment trial

[2] A version of this chapter is published as:
Bernal-Vasquez, A.-M., Möhring, J., Schmidt, M., Schönleben, M., Schön, C.-C., & Piepho, H.-P. (2014). The importance of phenotypic data analysis for genomic prediction - a case study comparing different spatial models in rye. *BMC Genomics*, 15:646.

(MET) in rye, which was connected between years only through one check, and used different spatial models to obtain better estimates, and thus, improved predictive abilities for genomic prediction. The aims of this study were to assess the advantage of using spatial models for the predictive abilities of genomic prediction, to identify suitable procedures to analyse a MET weakly connected across years using different stage-wise approaches, and to explore genomic prediction as a tool for selection of models for phenotypic data analysis.

Results: Using complex spatial models did not significantly improve the predictive ability of genomic prediction, but using row and column effects yielded the highest predictive abilities of all models. In the case of MET poorly connected between years, analysing each year separately and fitting year as a fixed effect in the genomic prediction stage yielded the most realistic predictive abilities. Predictive abilities can also be used to select models for phenotypic data analysis. The trend of the predictive abilities was not the same as the traditionally used Akaike information criterion, but favoured in the end the same models.

Conclusions: Making predictions using weakly linked datasets is of utmost interest for plant breeders. We provide an example with suggestions on how to handle such cases. Rather than relying on checks we show how to use year means across all entries for integrating data across years. It is further shown that fitting of row and column effects captures most of the heterogeneity in the field trials analysed.

Keywords: Stage-wise analysis, Genomic prediction, Cross validation, Spatial models, Multi-environment trials (MET), Restricted maximum likelihood (REML).

3.2 Background

Genomic prediction (GP) was first introduced in 2001 [Meuwissen et al., 2001] as a method that allows the prediction of genomic estimated breeding values (GEBV) for plants and animals by using information of genetic markers. In plant breeding, GP has been adopted as another stage of the breeding scheme [Schulz-Streeck et al., 2013b], not diminishing the importance of the phenotypic analysis usually carried out in several environments. Merging the phenotype and the genotype analyses has been addressed through the so-called stage-wise analysis [Piepho et al., 2012a]. In the first stage environments are analysed separately and genotype means are computed, to then in the GP stage predict GEBV based on dense genetic markers such as single nucleotide polymorphisms (SNPs).

In plant breeding, assessing genotypic adaptability and stability, and predicting breeding values of the genotypes in other environments and other years, makes use of multi-environment trials (METs), which aim to evaluate as many genotypes as possible in as many as possible locations [Burgueño et al.,

2011; Crossa et al., 2006; Piepho et al., 2008a; Smith et al., 2001]. These METs are typically laid out as generalised lattice designs testing a large number of different genotypes per trial. The number of tested genotypes is limited by factors such as seed production, production cycle length and availability of physical resources, e.g. land and budget [Besag and Kempton, 1986].

Within years, genotypes are tested in series of trials, which are connected by checks. Checks are lines grown in every trial as controls because their performance is known and/or they are already commercial material. Checks can be also used to connect years. In the rye breeding program considered in this paper, a completely different set of genotypes is tested in each year, but these genotypes are from the same breeding population. The accuracy of a genomic prediction model depends on the number of genotypes used for calibration. So there is definitely a need to combine data across years. Low connectivity across years is a challenge when trying to combine data across years, and this is one main motivation for this paper. Furthermore, the unbalancedness due to the design layout and the different and large number of evaluated genotypes increases the heterogeneity introducing high complexity to the variance-covariance structure among adjusted genotype means [Piepho et al., 2012a].

Analysis of METs could be done as single-stage analysis, modelling the complete observed data at the level of individual plots, or using a stage-wise approach, where experiments are analysed first at the level of environments (or trials), obtaining adjusted means per genotype, which are then summarised across environments (or trials) in the next stage [Piepho et al., 2012a]. A single-stage analysis accounts entirely for the variance-covariance structure of the recorded observations [Smith et al., 2001], therefore it is regarded as the gold standard. However, it has been shown that in a stage-wise analysis, a loss of information occurring in the transition through stages can be minimized by an appropriate weighting scheme [Möhring and Piepho, 2009].

If feasible, a single-stage approach is preferable to a stage-wise analysis [Cullis et al., 1998]. Nevertheless, the latter is acceptable for GP, since it is simple, computationally more efficient and also allows to easily account for any specifics of randomisation layout and error modelling for each environment Piepho et al. [2012a]. It should be stressed, however, that in a stage-wise analysis the weights are chosen to approximate the variance-covariance matrix of adjusted means from previous stages. We used here a three-stage approach and compared different spatial correlation structures in the first stage to correct field heterogeneity at the trial level.

Spatial error models may provide more accurate estimates of genotype effects than models not accounting for spatial adjustment [Duarte and Vencovsky, 2005; Zimmerman and Harville, 1991] but they are computationally more demanding and convergence may be difficult to reach. Any effort in terms of improving the genomic predictions would include checking if these improved estimates have

an effect on the predictive ability when markers are added to the model. The performance of alternative spatial models can be assessed by *k*-fold cross validation (CV).

Similarly, the merits of different spatial models used to compute adjusted means in the first stage can be compared by the same CV procedure, if the same GP procedure is used for each analysis. This suggests that genomic prediction-cross validation (GP-CV) can be used to identify the best-fitting mixed model in stage one. The common method of model selection makes use of information criteria based on the log likelihood, e.g. the Akaike information criterion (AIC) or the Bayesian information criterion (BIC) [Spilke et al., 2010]. When the restricted maximum likelihood (REML) method is used, models can only be compared by information criteria if they have the same fixed effects; otherwise, the maximum likelihood (ML) method should be used [Spilke et al., 2010]. CV is, in this sense, not used to tune parameters as in many penalization methods (e.g. adaptive Lasso, SCAD (Smoothly Clipped Absolute Deviation), machine learning methods) but only as a tool to compare models that use REML. REML is considered the best available method of variance parameter estimation, preferable to ML [Searle et al., 1992]. Consequently, it is of interest to devise model selection procedures that can use REML and also can compare models with different fixed effects. GP-CV has already been used to judge environments in order to optimise the accuracy in GP [Heslot et al., 2013b]. We used this tool here as model selection method in comparison to the traditional use of AIC.

The aims of this work were: i) to assess the advantage for the predictive ability when using a spatial model for phenotypic analysis, ii) to compare stage-wise approaches for GP when the data are weakly connected across years, and iii) to compare AIC and GP-CV as methods of selection of models for phenotypic data analysis towards GP in rye.

3.3 Methods

3.3.1 Field layout and data set

A commercial rye breeding program by KWS-LOCHOW established in Poland and Germany aims to develop superior hybrid varieties for the seed market. The implementation of GP within the breeding program makes use of the measurements of hybrid performance of the first cycles of phenotypic evaluation of the material (Cycle1). Selections made in Cycle1 are intensively evaluated in further cycles, aiming to double-check the selection decisions. For our purposes, these additional cycles do not add much useful information. Hence, we used only the first cycles of the program. The populations tested in each year consist of S_2 genotypes, which display genetic relatedness and population stratification due to complex genealogical history [Kang et al., 2008].

Besides the phenotypic data, a 16K Infinium iSelect HD Custom BeadChip was used to characterise 1610 individuals from Cycle1-2009 and Cycle1-2010 and 6 checks. Several traits were evaluated during this project: grain dry matter yield, plant height and thousand kernel weight, as well as ordinal scores of rust, mildew and lodging among others. In this work we used grain dry matter yield measurements of the phases of selection Cycle1-2009, Cycle1-2010 and Cycle1-2012, and marker information for the genotypes of 2009 and 2010. Although no marker information of year 2012 was available, it makes sense to use this dataset to observe the trend in one additional year and in this way, support the results of the phenotypic analysis of previous years.

A Cycle1 experiment consists of subsets of 320 genotypes from the S_2 populations tested in several locations within each of the two countries involving two testers (Tables 3.1 and 3.2). We define a trial as the physical unit within a location, where a subset of genotypes that were testcrossed to the same tester is evaluated. Trials at a location were laid out as α-designs with two replicates. Each trial was randomized independently from the others using the software CycDesign (VSN International; http://www.vsni.co.uk/). (However, we are aware that some breeders tend to use the same randomization layout in several locations. Ideally, each trial should have a different randomization). In our notation, trials of a Cycle1 experiment are labelled as S1, S2, ..., S24. Row and column coordinates of the plots to account for spatial variation are available.

Table 3.1: **General representation of the testers by locations (Loc) by years classification of Cycle1 year 2009 and 2010 in Germany (G-L1, $\cdots$, G-L8) and Poland (P-L1, $\cdots$, P-L4).**

Loc	**Cycle1-2009**						**Cycle1-2010**					
	Tester1			Tester2			Tester3			Tester4		
G-L1	S1	S2	S3				S10	S11	S12			
G-L2	S1	S2	S3					S11		S10		
G-L3	S1	S2	S3									
G-L4	S1	S2	S3	S1	S2	S3	S10	S11	S12	S10	S11	S12
G-L5				S1	S2	S3				S10	S11	S12
G-L6				S1	S2	S3				S10	S11	S12
G-L7				S1	S2	S3					S11	S12
G-L8							S10	S11	S12			
P-L1	S7	S8	S9	S7	S8	S9	S13	S14	S15	S13	S14	S15
P-L2	S7	S8	S9	S7	S8	S9	S13	S14	S15	S13	S14	S15
P-L3	S7	S8	S9	S7	S8	S9	S13	S14	S15	S13	S14	S15
P-L4	S7	S8	S9	S7	S8	S9	S13	S14	S15	S13	S14	S15

Series of trials are represented with the labels S1, S2, $\cdots$ S15.

Table 3.2: **General representation of the testers by locations (Loc) classification of Cycle1 year 2012 in Germany (G-L4, · · · , G-L11) and Poland (P-L1, · · · , P-L6).**

Loc	Cycle1-2012											
	Tester5						Tester6					
G-L4	S16		S17		S18							
G-L5							S16		S17		S18	
G-L6							S16		S17		S18	
G-L7							S16		S17		S18	
G-L8			S17		S18							
G-L9	S16		S17		S18							
G-L10	S16						S16		S17		S18	
G-L11	S16		S17		S18							
P-L1	S19	S20	S21	S22	S23	S24	S19		S21		S23	
P-L2	S19	S20	S21	S22	S23	S24	S19	S20	S21	S22	S23	S24
P-L3	S19	S20	S21	S22	S23	S24	S19	S20	S21	S22	S23	S24
P-L4		S20		S22		S24	S19	S20	S21	S22	S23	S24
P-L5	S31		S33		S35							
P-L6								S20		S22		S24

Series of trials are represented with the labels S16, S17, · · · S24.

Normally throughout the program, only a single tester was used per location and year, but in some locations, some subsets of genotypes were testcrossed with the two available testers. This is the case, for example, for location G-L4 in Cycle1-2009, where the genotypes evaluated in the trials S1, S2 and S3 were testcrossed with both Tester1 and Tester2, and it is also the case of locations P-L1, P-L2, P-L3 and P-L4 evaluating genotypes of trials S7, S8 and S9 with both testers. In each year, four common checks were testcrossed with the testers and grown twice in each trial. Over the years 2009 and 2010 one check was in common and none was shared with 2012 (Table 3.3).

The field layout of some trials was not perfectly rectangular. Some trials at a given location and year had fewer blocks but larger size, i.e., there were two different block sizes within a few trials. Blocks were nested within rows of the field layout.

In the genetic dataset, homozygous marker genotypes were coded as -1 and 1, and the heterozygous type, missing values and technical failures were coded as 0. 58.7% of the markers corresponded to homozygous alleles and 16.1% were heterozygous. Only a 0.03% of the markers were recorded as missing values or technical failures; therefore, an imputation method would not have a strong impact on the subsequent analyses. Monomorphic markers and markers with minor allele frequency (MAF) less

Table 3.3: **Year x Check classification in Germany (G) and Poland (P).**

	2009		2010		2012	
	G	P	G	P	G	P
Check1	x	x				
Check2	x	x				
Check3	x	x	x	x		
Check4	x	x				
Check5			x	x		
Check6			x	x		
Check7			x	x		
Check8					x	x
Check9					x	x
Check10					x	x
Check11					x	x
Check12					x	x
Check13						x

than 1% or missing information of more than 10% per marker were dropped. A total of 11285 markers passed the quality test and were used for GP.

3.3.2 Models

In this chapter we present the models used in the first stage of the analysis and the models of the approaches followed to adjust the year effect either in the second or the third stage. Figures 3.1 and 3.2 depict a general scheme that helps visualizing the methodology.

First stage

In the first stage we computed adjusted genotype means by location and year. The factors used for the analysis were genotypes (G), testers (T), trials (S), replicates (R) nested within trials and blocks (B) nested within replicates. We defined a baseline model as

$$Y_{hijkv} = (GT)_{hv} + S_i + R_{ij} + B_{ijk} + e_{hijkv}, \tag{3.1}$$

where Y_{hijkv} is the observed grain dry matter yield of the h-th genotype testcrossed with the v-th tester in the k-th block within the j-th replicate of the i-th trial, $(GT)_{hv}$ is the effect of the h-th genotype

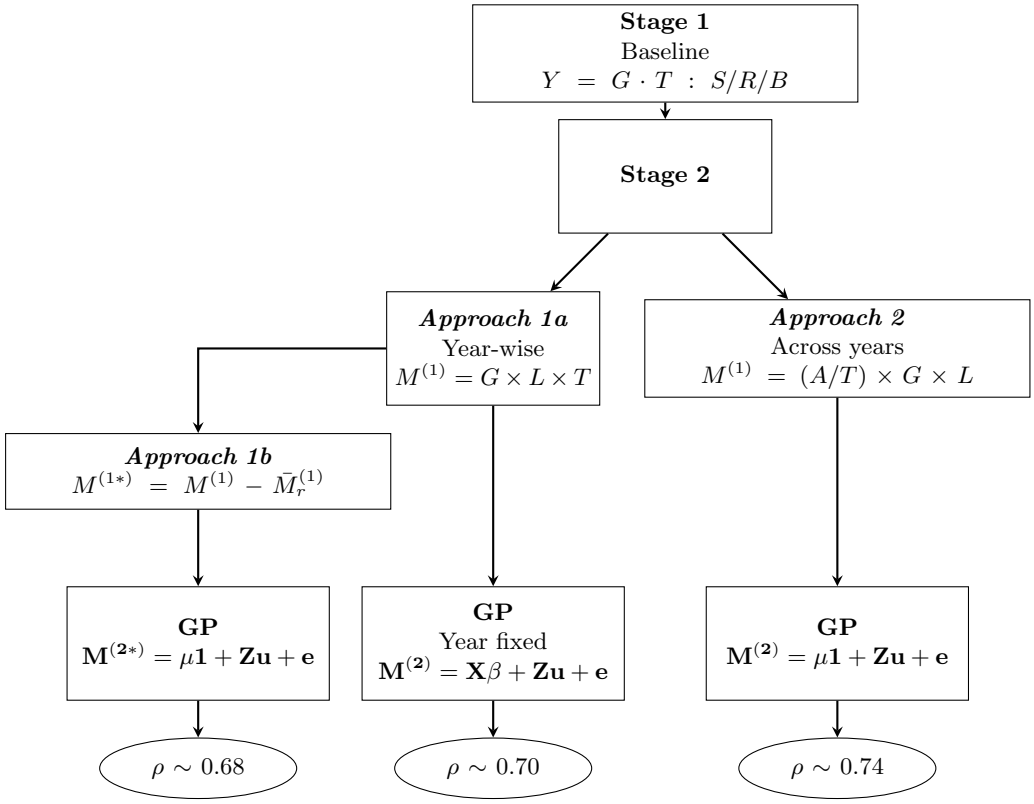


Figure 3.1: General representation of stage-wise approaches to compare year-effect adjustment. Factors were genotype (G), tester (T), location (L), year (A), trial (S), replicate (R) and block (B). Grain dry matter yield (Y) is the response variable in the first stage, $M^{(1)}$ is the adjusted mean of genotypes across locations used in the second stage, $M^{(1*)}$ is the year effect-corrected genotype adjusted mean, $\bar{M}_r^{(1)}$ represents the simple mean of genotypes of the r-th year. In the genomic prediction (GP) stage, $\mathbf{M^{(2)}}$ is the $n \times 1$ vector of adjusted means of genotypes by year for *Approach 1a* and across years for *Approach 2*, $\mathbf{M^{(2*)}}$ is the $n \times 1$ vector of adjusted means of year effect-corrected genotypes in *Approach 1b*, $\mathbf{X}$ and β are respectively the design matrix and parameter vector of fixed effects, $\mathbf{Z}$ is the $n \times p$ marker matrix, $\mathbf{u}$ is the p-dimensional vector of SNP effects and $\mathbf{e}$ the error vector. $Y = G \cdot T : S/R/B$ is the shorthand notation of the model eq. (1) in the text: $Y_{hijkv} = (GT)_{hv} + S_i + R_{ij} + B_{ijk} + e_{hijkv}$, $M^{(1)} = G \times L \times T$ stands for the model eq. (2) in the text: $M^{(1)}_{hsv} = G_h + L_s + T_v + (GL)_{hs} + (GT)_{hv} + (LT)_{sv} + (GLT)_{hsv} + e_{hsv}$, and $M^{(1)} = (A/T) \times G \times L$ represents the extended model eq. (4) in the text: $M^{(1)}_{hrsv} = G_h + L_s + (AT)_{rv} + (GA)_{hr} + (GAT)_{hrv} + (GL)_{hs} + (LA)_{rs} + (LAT)_{rsv} + (GLA)_{hrs} + (GLAT)_{hrsv} + e_{hrsv}$. The final predictive abilities (ρ) are presented in the ellipses.

testcrossed with the v-th tester, S_i is the effect of the i-th trial [$S_i \sim N(0, \sigma^2_S)$], R_{ij} is the effect of the j-th replicate nested within the i-th trial [$R_{ij} \sim N(0, \sigma^2_R)$], B_{ijk} is the effect of the k-th block nested within the j-th replicate of the i-th trial [$B_{ijk} \sim N(0, \sigma^2_B)$] and e_{hijkv} is the plot error associated with the Y_{hijkv} observation [$e_{hijkv} \sim N(0, \sigma^2_e)$]. In model equation (3.1) we assumed genotypes crossed with testers as a fixed effect to be able to compute genotype adjusted means per tester, whereas the other effects were considered as random effects due to the nested design structure [Piepho et al., 2003].

Table 3.4 summarises the further models. Some SAS code to fit the first stage models is provided

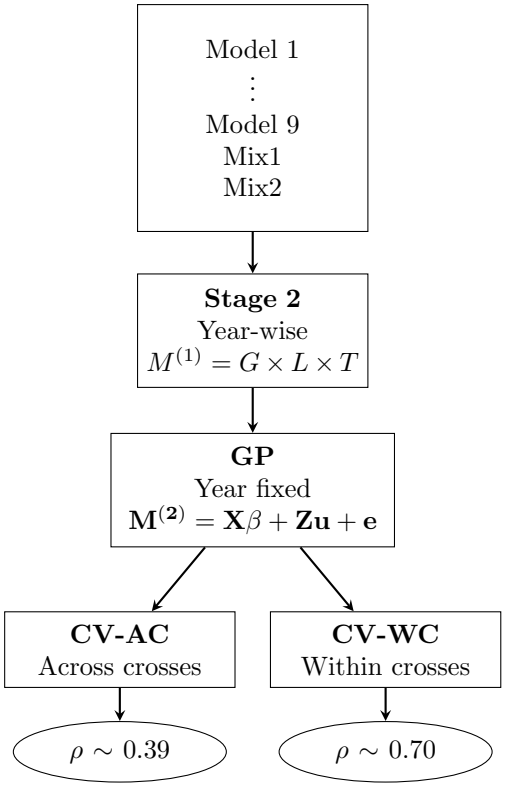


Figure 3.2: General representation of model comparison through all the stages of the analysis. Datasets generated from 9 spatial and non-spatial models plus two mixed datasets generated from best models given the Akaike information criterion (Mix1) and the predictive abilities (Mix2). Factors in second stage were genotype (G), location (L) and tester (T). $M^{(1)}$ represents the adjusted mean of genotypes across locations and years. $M^{(1)} = G \times L \times T$ is the shorthand notation for $M^{(1)}_{hsv} = G_h + L_s + T_v + (GL)_{hs} + (GT)_{hv} + (LT)_{sv} + (GLT)_{hsv} + e_{hsv}$. In the genomic prediction (GP) stage $\mathbf{M^{(2)}}$ is the adjusted mean of genotypes across locations, $\mathbf{X}$ and β are respectively the design matrix and parameter vector of fixed effects, $\mathbf{Z}$ is the $n \times p$ marker matrix, $\mathbf{u}$ is the p-dimensional vector of SNP effects and $\mathbf{e}$ the error vector. Sampling methods in cross validation (CV) were across crosses (AC) and within crosses (WC). The final predictive abilities (ρ) are presented in the ellipses.

in the supplementary material (Appendix B). The first model (M1) will be referred to as the baseline model because it was the simplest model and represented the randomisation structure. In the second model (M2) we considered additionally the effects of the o-th row (W_{ijo}) and the q-th column (V_{ijq}) both within the j-th replicate of the i-th trial. Subsequently, we added a spatially correlated residual plot effect different from the baseline model, which uses the independent model (ID) with homogeneous variances. We fitted one- and two-dimensional spatial models with and without the so-called nugget, a geostatistical term to designate an independent error effect. As one-dimensional models we used the autoregressive AR(1) variance-covariance nested within blocks without nugget (M3) and with nugget (M7), and linear variance LV within blocks with nugget (M4). In the AR(1) we accounted for the correlation between plots in the same block assuming an exponential decay of correlation with distance,

whereas by using LV, it is assumed that the covariance among plots in the same block decays linearly with spatial distance [Piepho et al., 2008b; Williams, 1986]. The most common extension of the spatial model in two dimensions is the direct product structure AR(1) $\times$ AR(1), which assumes that an AR(1) model holds both along rows and along columns [Gilmour et al., 1997]. The two-dimensional models were fitted along rows and columns within replicates without nugget (M5), with nugget (M8), adding rows and columns as effects without nugget (M6) and with nugget (M9). The LV model can also be extended in two dimensions [Piepho and Williams, 2010]; however, for METs, where the arrangement of the plots might not be perfectly rectangular, this LV $\times$ LV model was cumbersome to fit with the software we used, thus we did not consider this model.

Table 3.4: **Spatial and non-spatial models used for the first stage.**

Label	Model	Variance-covariance structure for error
M1	$Y_{hijkv} = (GT)_{hv} + S_i + R_{ij} + B_{ijk} + e_{hijkv}$	ID
M2	$Y_{hijkoqv} = (GT)_{hv} + S_i + R_{ij} + B_{ijk} + W_{ijo} + V_{ijq} + e_{hijkoqv}$	ID
M3	$Y_{hijkv} = (GT)_{hv} + S_i + R_{ij} + B_{ijk} + e_{hijkv}$	AR(1) within B
M4	$Y_{hijkv} = (GT)_{hv} + S_i + R_{ij} + B_{ijk} + e_{hijkv}$	LV within B + nugget
M5	$Y_{hijkoqv} = (GT)_{hv} + S_i + R_{ij} + B_{ijk} + W_{ijo} + V_{ijq} + e_{hijkoqv}$	AR(1) $\times$ AR(1) within R
M6	$Y_{hijkv} = (GT)_{hv} + S_i + R_{ij} + B_{ijk} + e_{hijkv}$	AR(1) $\times$ AR(1) within R
M7	$Y_{hijkv} = (GT)_{hv} + S_i + R_{ij} + B_{ijk} + e_{hijkv}$	Model 3 + nugget
M8	$Y_{hijkv} = (GT)_{hv} + S_i + R_{ij} + B_{ijk} + e_{hijkv}$	Model 5 + nugget
M9	$Y_{hijkoqv} = (GT)_{hv} + S_i + R_{ij} + B_{ijk} + W_{ijo} + V_{ijq} + e_{hijkoqv}$	Model 6 + nugget

Y_{hijkv} is the observed dry matter yield of the h-th genotype testcrossed with the v-th tester in the k-th block within the j-th replicate of the i-th trial, $(GT)_{hv}$ is the effect of the h-th genotype testcrossed with the v-th tester, S_i is the effect of the i-th trial [$S_i \sim N(0, \sigma^2_S)$], R_{ij} is the effect of the j-th replicate nested within the i-th trial [$R_{ij} \sim N(0, \sigma^2_R)$], B_{ijk} is the effect of the k-th block nested within the j-th replicate of the i-th trial [$B_{ijk} \sim N(0, \sigma^2_B)$] and e_{hijkv} is the plot error associated with the Y_{hijkv} observation [$e_{hijkv} \sim N(0, \sigma^2_e)$]. In the models including row and column effects, W_{ijo} is the effect of the o-th row within the j-th replicate of the i-th trial [$W_{ijo} \sim N(0, \sigma^2_W)$] and V_{ijq} is the effect of the q-th column within the j-th replicate of the i-th trial [$V_{ijq} \sim N(0, \sigma^2_V)$]. Spatial variance-covariance structure were independent (ID), autoregressive in one direction (AR(1)), one-dimension linear variance (LV) and two-dimension autoregressive [AR(1)$\times$ AR(1)].

Note that we use $(GT)_{hv}$ as fixed effect, which is necessary to obtain the genotype by tester means. The purpose is also to recover the information of the entries that are grown in the same locations but using different testers (e.g. in Cycle1 location G-L4 and the Polish locations P-L1 to P-L4), so that we captured the effect of the tester in the shared locations.

Second stage

In the second stage we computed genotype means across locations and testers. This was done either separately for each year (*Approach 1*) or also averaging across years (*Approach 2*). The years 2009 and 2010, where molecular marker data were available, were connected through only one check. The resulting fundamental question is then how to fit the year effect. Either the year effect is estimated by the mean of all tested entries *(Approach 1)* or we rely on the adjustment by the one single check *(Approach 2)*. We assume that genotypes tested in each year can be regarded as a random sample from the same parent population. Based on the structure of the breeding program, this is a realistic assumption that motivates the approaches described in the following.

Both approaches were compared using the $M^{(1)}$ resulting from the analysis of the baseline model in the first stage.

Approach 1: Year-wise analysis

Each year was analysed in the second stage using a three-way interaction model of genotypes (G), locations (L) and testers (T) as factors to obtain adjusted genotypes means of each year. The model was

$$M^{(1)}_{hsv} = G_h + L_s + T_v + (GL)_{hs} + (GT)_{hv} + (LT)_{sv} + (GLT)_{hsv} + e_{hsv}, \tag{3.2}$$

where $M^{(1)}_{hsv}$ represents the adjusted mean of grain dry matter yield of the h-th genotype, testcrossed with the v-th tester in the s-th location, G_h, L_s and T_v are the main effects of the h-th genotype, the s-th location and the v-th tester, respectively, $(GL)_{hs}$, $(GT)_{hv}$ and $(LT)_{sv}$ are the two-way interaction effects, $(GLT)_{hsv}$ is the effect of the three-way interaction and e_{hsv} is the residual error associated with $M^{(1)}_{hsv}$ $[e_{hsv} \sim N(0, \sigma^2_{e_{[hsv]}})]$, with $\sigma^2_{e_{[hsv]}}$ the variance of the hsv-th adjusted mean $(M^{(1)}_{hsv})$ obtained in the first stage.

Location was considered as random effect $[L_s \sim N(0, \sigma^2_L)]$ and hence, all the interactions containing this factor are random [Piepho et al., 2003]. The crossed effect of genotypes and testers $[(GT)_{hv}]$ could have been a fixed effect since genotypes and testers are taken as fixed factors in this stage. However, the crossed effects that include G were taken as random here because the factor genotype was used as random in the GP stage. But note that in the first and the second stage we needed to take genotype main effects as fixed in order to compute adjusted means [Piepho et al., 2012a]. Besides, since not every

genotype was tested with every tester (e.g. in Cycle1 locations G-L1 to G-L3 and G-L5 to G-L8), we needed to take $(GT)_{hv}$ random to be able to estimate genotype means across levels of testers.

In this approach, the year effect was adjusted in two ways, hereafter referred as to *Approach 1a* and *Approach 1b*. *Approach 1a* used years as fixed factors in the GP stage and *Approach 1b* used a manual adjustment after the second stage by simply calculating the mean of the genotypes by year $(\bar{M}_r^{(1)})$ and subtracting it to each genotype adjusted mean of the corresponding year (Figure 3.1). The rationale behind the latter approach is the assumption that the correction for the year effect is better represented by the simple mean of the complete sample of genotypes per year than by just a few checks. The resulting year effect-corrected genotype means $(M_{hsv}^{(1*)})$ are forwarded to the GP stage, and through CV are evaluated as predictors.

As in the transition from the first to the second stage, there is a loss of information in passing on from the second to the third stage because the $(GLT)_{hsv}$ effect is confounded with the residual error term. This loss can be minimized by weighting the adjusted means [Piepho et al., 2012a]. We used the Smith et al. [2001] scheme, where adjusted means are weighted by the diagonal elements of the inverse of their variance-covariance matrix computed in the first stage.

At this stage, we computed the heritability for each year using the *ad hoc* method described in Piepho and Möhring [2007] as

$$\bar{H}^2 = \frac{\sigma_G^2}{\sigma_G^2 + \bar{v}/2} \quad , \tag{3.3}$$

where σ_G^2 is the genetic variance and $\bar{v}$ is the mean variance of a difference of two adjusted genotype means, corresponding to the best linear unbiased estimators (BLUE). Even though this is not the best method to estimate heritability [Estaghvirou et al., 2013], the square root of this heritability estimate gives a rough idea of an upper limit for the predictive abilities.

Approach 2: Across years analysis

The model to account for the year effect in the second stage through the shared check was

$$\begin{aligned} M_{hrsv}^{(1)} = {} & G_h + L_s + D_{rv} + (GD)_{hrv} + (GL)_{hs} + (LD)_{rsv} \\ & + (GLD)_{hrsv} + e_{hrsv}, \end{aligned} \tag{3.4}$$

where $M_{hrsv}^{(1)}$ represents the adjusted mean of grain dry matter yield of the h-th genotype, testcrossed with the v-th tester, in the s-th location and r-th year, G_h is the main effect of the h-th genotype, L_s is

the main effect of the s-th location and D_{rv} the main effect of the v-th tester within the r-th year, which can be extended as $D_{rv} = A_r + (AT)_{rv}$, with A_r the effect of the year and T denoting the tester [Piepho et al., 2003]. $(GD)_{hrv}$, $(GL)_{hs}$ and $(LD)_{rsv}$ are the two-way interaction effects, $(GLD)_{hrsv}$ is the effect of the three-way interaction and e_{hrsv} is the residual error associated to $M^{(1)}_{hrsv}$ $[e_{hrsv} \sim N(0, \sigma^2_{e_{[hrsv]}})]$, with $\sigma^2_{e_{[hrsv]}}$ the variance of the $hrsv$-th adjusted mean $(M^{(1)}_{hrsv})$ obtained in the first stage. The effects containing D_{rv} can be extended as $(GD)_{hrv} = (GA)_{hr} + (GAT)_{hrv}$, $(LD)_{rsv} = (LA)_{rs} + (LAT)_{rsv}$ and $(GLD)_{hrsv} = (GLA)_{hrs} + (GLAT)_{hrsv}$.

We considered genotypes and testers as fixed factors and location and year as random factors $[L_s \sim N(0, \sigma^2_L)$ and $A_r \sim N(0, \sigma^2_A)]$. All effects involving A_r are random except $(AT)_{rv}$ because we do not want to recover inter-year information since there are only two years and the year by tester classification is very disconnected (years do not share testers). Moreover, the $(AT)_{rv}$ term is analogous to a block factor in an incomplete block design because it is free of G_h; therefore, due to the unbalancedness and the small number of years, we can use it as a fixed effect. Furthermore, the main year effect (A_r) can be dropped considering that the adjustment of the genotype means is the same for $A_r + (AT)_{rv}$ as for only $(AT)_{rv}$.

Including all the effects, the final model (3.4) is

$$\begin{aligned} M^{(1)}_{hrsv} = {} & G_h + L_s + (AT)_{rv} \\ & + (GA)_{hr} + (GAT)_{hrv} + (GL)_{hs} + (LA)_{rs} + (LAT)_{rsv} \\ & + (GLA)_{hrs} + (GLAT)_{hrsv} + e_{hrsv}, \end{aligned}$$

To minimise the loss of information in the transition to the GP stage, we weighted the adjusted means using the inverse of the squared standard errors, which is also appropriate since we are not fitting random block effects [Möhring and Piepho, 2009].

Third stage: Genomic prediction

At the third stage, the dataset of p markers was merged with the n grain dry matter yield adjusted means by years of evaluated models. GP was performed using ridge-regression best linear unbiased prediction (RR-BLUP), where the genotypic values are predicted using the marker information by regressing each SNP on the phenotype [Piepho, 2009b].

The model was

$$\mathbf{M^{(2)}} = \mathbf{X}\beta + \mathbf{Zu} + \mathbf{e} \tag{3.5}$$

where, $\mathbf{M^{(2)}}$ is the $n \times 1$ vector of phenotypic records, here, containing the adjusted means calculated from the second stage, $\mathbf{X}$ and β are, respectively, the design matrix and parameter vector of fixed effects, $\mathbf{Z}$ is the $n \times p$ marker matrix, whose elements z_{hm} represent the SNP genotype of the m-th marker of the h-th genotype entry and take the values -1, 0, or $+1$ for the aa, Aa, and AA genotypes Piepho [2009b], $\mathbf{u}$ is the p-dimensional vector of SNP effects and $\mathbf{e}$ is the error vector. The term $\mathbf{Zu}$ is interpreted as the genetic effect and its estimate $\mathbf{Z\hat{u}}$ as the GEBV. The GEBV of the h-th genotype corresponds to $GEBV_h = \sum_{m=1}^{p} \hat{u}_m z_{hm}$, with $m = 1, \cdots, p$ the number of markers, $\hat{u}_m$ is the estimated effect of the m-th marker and z_{hm} the SNP genotype of the m-th marker for the h-th genotype entry. The assumptions of the model are that the error is normally distributed with zero mean and variance $\mathbf{R}$ [$\mathbf{e} \sim N(0, \mathbf{R})$] and that $\mathbf{u}$ has a normal distribution with zero mean and variance $\mathbf{I}_p\sigma_u^2$ [$\mathbf{u} \sim N(0, \mathbf{I}_p\sigma_u^2)$]. $\mathbf{R}$ is a diagonal matrix with diagonal elements equal to the inverses of the diagonal elements of the inverse of the original variance-covariance matrix of the adjusted means of the second stage Smith et al. [2001]. $\mathbf{I}_p$ is the p-dimensional identity matrix and σ_u^2 represents the proportion of the genetic variance contributed by each individual SNP.

Under the model equation (3.5) the variance of the observed data is $var(\mathbf{M^{(2)}}) = \mathbf{\Gamma}\sigma_u^2 + \mathbf{R}$, in which $\mathbf{\Gamma} = \mathbf{ZZ^T}$ and $\mathbf{Z}^T$ denotes the transpose of $\mathbf{Z}$ [Piepho, 2009b]. To speed up the computation, $\mathbf{\Gamma}$ was rescaled by replacing $\mathbf{Z}$ with $\mathbf{Z}/\sqrt{p}$, with p the number of markers [Piepho et al., 2012b].

In the year-wise analysis (*Approach 1a*), the genotype adjusted means by year are merged in the $\mathbf{M^{(2)}}$ vector, and vector β contains the intercept and the year effect. In the across-years analysis (*Approach 2*), where year effect was already accounted for, $\mathbf{M^{(2)}}$ contains the genotype adjusted means and vector β contains only the intercept. In the year-wise analysis correcting genotype adjusted means for year effects (*Approach 1b*), the model used did not include a fixed year factor (since we had already adjusted for it) but a common intercept, thus the model was the same as for across-years analysis.

To measure the influence of the relationship among the genotypes on the predictions, we used the adjusted means obtained in the second stage and the pedigree information of the entries in a mixed model testing genotypes and crosses as random effects, so that the variances of both effects would give us an estimation of how much the variation is attributed to the pedigree, e.g. the crosses. The model was

$$M_{ah}^{(2)} = G_h + C_a + e_{ah} \tag{3.6}$$

where $M_{ah}^{(2)}$ is the adjusted mean of the h-th genotype obtained in the second stage, G_h is the effect of

the h-th genotype, C_a is the effect of the a-th grand parent (gp) cross, e.g. (gp1 $\times$ gp2) $\times$ (gp3 $\times$ gp4), and e_{ah} the associated error. Additionally, we plotted the relationship heat-map of estimated coefficients of relatedness for individuals based on marker data computed according to VanRaden [2008].

Cross validation for model comparison

To evaluate model performance, k-fold CV was carried out. In CV, the data is split into k subsets t times. $k-1$ subsets are used as the training set (TS) and the one other subset is the validation set (VS). The TS is used to estimate the parameters that then are used to predict the observations in the VS. The performance of the model was assessed by the Pearson correlation coefficient between the predicted GEBV and the corresponding observations of the VS. This correlation is referred to as predictive ability [Estaghvirou et al., 2013]. As in the first stage, the predictive ability was not adjusted by the square root of the heritability. Although breeding programs are most of the time operating with closely related genotypes, breeders are also interested in knowing the results in a scenario with more distantly related genotypes, for example, using genotypes that share the same grandparents either in the TS or in the VS but not in both. Hence, we wanted to check if accounting for the effect of population structure in the randomisation of CV would make the spatial error models improve the predictive abilities. We chose two scenarios given the relatedness level of the entries and followed the suggested sampling schemes from Albrecht et al. [2011], which takes into consideration this fact in the CV procedure. In the first sampling scheme, hereafter called "within crosses" (WC), random sampling is done using all genotypes in the dataset; in the second scheme, hereafter referred to as "across crosses" (AC), genotypes were clustered by cross, so that complete cross-groups were used randomly either in the VS or the TS. There were 349 crosses of different sizes, sharing none, one or two grand parents. The general overview of the methodology is depicted in Figure 3.2.

3.3.3 Model selection

Two strategies for selecting the best phenotypic model were used in the first stage. In strategy one the best model for all locations is selected, that is, there is no model selection per location but across locations. In strategy two, model selection is location-specific (Figure 3.3). For both strategies we computed the AIC and performed genomic prediction-cross validation (GP-CV), both per location-year combination. To accomplish the GP-CV approach, we used the adjusted means per location and year of all spatial and non-spatial models. Then, means of genotypes by year-location combination were joined with the molecular marker data to perform GP-CV, in which genetic values were regressed on markers and validation of the model was done using k-fold CV. Predictions of unobserved records and predictive

abilities of each model were obtained for each year-location combination. We assessed the predictive ability of the models using the Pearson correlation coefficient (ρ) between the predicted GEBV and the observed phenotypic value. Hereafter we denote this predictive ability as ρ-GP-CV. Predictive abilities were not adjusted with the square root of the heritability, as suggested by Dekkers [2007], since this adds an extra error due to heritability computation [Estaghvirou et al., 2013; Heslot et al., 2013b].

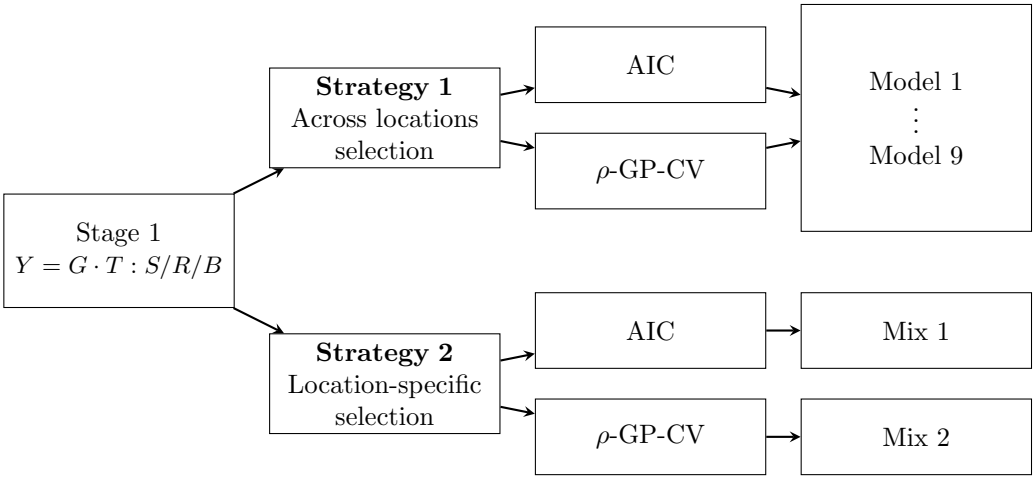


Figure 3.3: **General representation of strategies to compare model selection methods.** Factors were genotype (G), tester (T), trial (S), replicate (R) and block (B). Grain dry matter yield (Y) is the response variable in the first stage. $Y = G \cdot T : S/R/B$ is the shorthand notation for the model $Y_{hijkv} = (GT)_{hv} + S_i + R_{ij} + B_{ijk} + e_{hijkv}$. Datasets of 9 spatial and non spatial models plus one mixed dataset (Mix1) generated from best models given the Akaike information criterion (AIC) and another mixed dataset (Mix2) generated from best models given the predictive abilities (ρ-GP-CV).

For strategy one (across locations model selection), the number of locations with the best fits (either AIC or ρ-GP-CV) was counted, so that the model with the best fits in the majority of locations was identified as the best model. For strategy two (location-specific model selection), two datasets were built: "Mix 1", containing the adjusted means of the locations with the best fit according to the AIC and "Mix 2", containing the adjusted means of the locations with the highest ρ-GP-CV. Thus, after the first stage we had in total eleven data sets of adjusted means, nine corresponding to each tested model from strategy one, plus two more datasets from strategy two: A mixed data set (Mix 1) with the best models per location-year according to the AIC, and another mixed set (Mix 2) with best models per location-year according to the ρ-GP-CV.

3.3.4 Softwares

All analyses were performed using SAS. Stage 1 and 3 used the MIXED procedure and Stage 2 used PROC HPMIXED. Relationship matrix was calculated using the Symbreed Package [Wimmer et al.,

2012] for R 2.15.

3.4 Results

3.4.1 First stage - strategy 1: Model selection across locations

In the first stage - strategy 1, we did model selection across locations using AIC and predictive abilities (ρ-GP-CV) per location-year combination. According to the AIC, the results favoured the two-dimensional models (Table 3.5). To do a fair comparison between selection methods using AIC and ρ-GP-CV, we first describe AIC for years 2009 and 2010, for which ρ-GP-CV were also available and then, as additional information, for year 2012, for which ρ-GP-CV was not available since the marker information was missing.

For years 2009 and 2010, M9 and M8 had the majority of best fits across locations. M9 (Baseline + row + column and AR(1) $\times$ AR(1) + nugget) resulted in 12 out of 22 cases as the best model. M8 (Baseline and AR(1) $\times$ AR(1) + nugget) was best in seven out of 22 cases. The baseline model + row + column (M2) fitted the best 9% of the times and M6 5% of the times.

A similar tendency was observed in 2012, where 43% of the times (six out of 14) M9 had the best fit and M8 was best 29% of the times. For this year 2012, models M7, M8 and M9 could not be fitted in some locations. Another third of the times (29%), M2 had best fits. Interestingly, M2 had the best fits in the locations that had convergence problems for models M8 and M9. M1, M3, M4, M5 and M7 never had best fits in any of both groups of years.

The predictive abilities (ρ-GP-CV) per location-year combination showed a rather different pattern for best models within locations; however, the two-dimensional models were also more frequently selected than one-dimensional models (Table 3.6). M8 (Baseline and AR(1) $\times$ AR(1) + nugget) showed in seven of 22 settings the highest ρ-GP-CV per location-year combination followed by M9 (Baseline + row + column and AR(1) $\times$ AR(1) + nugget) with six out of 22 times. The baseline model + row + column (M2) was selected twice and models M3, M4 and M6 had also one, three and three selections out of 22, respectively. M1, M5 and M7 had no best fits at all.

One location of 2009 (P-L3) produced a negative predictive ability for all models. We did not consider this location in the counting of best fits, since a higher negative number is actually a worse fit in regard to predictions, but low or high negative are both interpreted as zero prediction. Despite the negative correlations, this location was included in the mixed datasets produced from the site-specific model selection. We used the adjusted means produced from the baseline model. Another location

Table 3.5: **Akaike information criterion (AIC) of models at first stage (M1, $\cdots$, M9) by year and location (L) for grain dry matter yield (Y).**

Year	L	M1	M2	M3	M4	M5	M6	M7	M8	M9
2009	G-L1	101.7	84.3	45.5	47.2	20.4	6.9	45.6	**0**	1.7
2009	G-L2	83.1	67.5	50.9	38.5	31.4	20.7	40.7	0.5	**0**
2009	G-L3	45.7	30.4	41.5	31.1	40.1	26.9	31.2	1.0	**0**
2009	G-L4	125.0	19.1	125.1	114.9	90.3	19.6	115.5	65.0	**0**
2009	G-L5	29.1	8.0	18.1	24.5	15.3	1.2	–	12.3	**0**
2009	G-L6	51.6	47.6	37.7	29.5	41.7	35.4	29.4	**0**	1.2
2009	G-L7	81.5	56.1	55.3	62.8	36.5	11.0	55.5	5.1	**0**
2009	P-L1	126.4	115.6	121.6	116.3	109.5	108.8	116.2	**0**	1.9
2009	P-L2	62.3	45.4	62.4	54.6	57.3	47.2	54.9	1.5	**0**
2009	P-L3	120.9	65.9	116.1	105.5	99.7	49.6	105.5	17.3	**0**
2009	P-L4	145.9	98.6	132.8	126.4	126.4	80.1	126.4	0.4	**0**
2010	G-L1	35.5	4.9	35.6	31.5	12.3	**0**	32.0	12.3	1.8
2010	G-L2	25.0	7.2	27.0	21.7	29.7	11.9	19.7	**0**	*-3.2*
2010	G-L4	141.4	74.2	128.7	117.1	130.2	57.4	118.4	5.0	**0**
2010	G-L5	21.6	**0**	23.4	22.9	21.9	3.3	22.9	22.1	2.8
2010	G-L6	80.9	60.0	72.8	59.8	55.4	41.5	61.1	**0**	0.6
2010	G-L7	69.5	22.3	56.2	47.8	37.2	23.6	48.1	2.6	**0**
2010	G-L8	40.8	24.7	32.1	22.6	27.7	19.6	23.1	**0**	1.4
2010	P-L1	38.8	5.7	38.8	38.8	39.4	9.4	40.8	39.1	**0**
2010	P-L2	40.0	0.7	41.6	36.1	39.8	4.1	36.9	4.3	**0**
2010	P-L3	66.4	**0**	68.4	67.2	69.5	3.7	70.4	71.5	5.7
2010	P-L4	95.0	80.4	90.5	79.1	87.0	66.7	79.4	**0**	3.2
	Counts	0	2	0	0	0	1	0	7	**12**
		0%	9%	0%	0%	0%	5%	0.00	32%	**55%**
2012	G-L4	35.3	**0**	35.3	36.2	26.0	0.6	35.3	24.2	–
2012	G-L5	66.3	2.6	67.0	66.3	42.1	5.9	–	21.5	**0**
2012	G-L6	148.4	131.4	93.8	93.7	18.7	18.7	89.9	**0**	**0**
2012	G-L7	38.3	4.5	40.3	38.3	36.3	**0**	42.3	–	1.9
2012	G-L8	45.3	39.8	37.7	33.5	35.6	37.3	33.9	1.9	**0**
2012	G-L9	402.3	321.5	200.9	181.7	81.9	81.9	191.6	**0**	**0**
2012	G-L10	39.7	**0**	41.5	41.4	22.1	3.5	43.5	6.7	1.1
2012	G-L11	18.0	**0**	19.7	18.0	8.4	1.2	21.6	3.7	–
2012	P-L1	189.5	168.8	158.9	148.9	146.3	137.8	149.1	**0**	1.7
2012	P-L2	127.4	49.3	129.1	122.6	129.7	49.9	123.9	5.9	**0**
2012	P-L3	107.8	55.3	103.1	95.0	101.0	49.3	96.1	7.9	**0**
2012	P-L4	226.3	0.2	226.3	222.1	226.3	**0**	226.3	226.3	2.0
2012	P-L5	13.2	**0**	13.2	13.2	11.9	1.5	13.2	13.9	3.5
2012	P-L6	79.0	54.8	70.4	66.9	65.8	37.9	67.0	**0**	1.7
	Counts	0	4	0	0	0	2	0	4	**6**
		0%	29%	0%	0%	0%	14%	0%	29%	**43%**

Table shows ΔAIC relative to the best model.
Boldfaced entries in the table indicate best model (fit) within location. Empty cells (–) correspond to locations where the model did not converge. In italics, we report the models that converged but the Hessian matrix was not positive definite.

Table 3.6: **Predictive abilities of observed and predicted values of a 5-fold-CV by year-location combination of models at first stage (M1, $\cdots$, M9) for grain dry matter yield (Y), and repeatability (R) of the trait by location.**

Year	Loc	M1	M2	M3	M4	M5	M6	M7	M8	M9	R
2009	G-L1	0.469	0.473	0.462	0.481	0.448	0.455	0.474	**0.481**	0.478	0.376
2009	G-L2	0.271	0.272	0.279	0.280	0.282	**0.288**	0.282	0.270	0.269	0.177
2009	G-L3	0.347	0.344	0.351	0.350	0.345	0.339	0.350	**0.355**§	0.355	0.264
2009	G-L4	0.595	0.593	0.597	**0.602**§	0.592	0.594	0.602	0.592	0.598	0.440
2009	G-L5	0.495	0.514	0.506	0.505	0.519	0.527	–	0.514	**0.529**	0.303
2009	G-L6	0.393	**0.398**	0.357	0.372	0.359	0.360	0.369	0.372	0.378	0.077
2009	G-L7	0.596	0.594	0.586	**0.599**	0.578	0.565	0.591	0.584	0.577	0.299
2009	P-L1	0.127	0.118	0.132	0.138	0.116	0.114	0.138	**0.174**	0.167	0.225
2009	P-L2	0.301	0.306	0.303	0.310	0.307	0.309	0.310	0.323	**0.323**§	0.338
2009	P-L3	-0.154	-0.165	-0.153	-0.154	-0.169	-0.172	-0.154	-0.158	-0.175	0.247
2009	P-L4	0.520	0.518	0.527	0.525	0.520	0.522	0.525	**0.558**	0.555	0.362
2010	G-L1	0.428	0.471	0.426	0.432	0.464	**0.478**	0.431	0.466	0.475	0.263
2010	G-L2	0.394	0.392	0.399	**0.407**	0.400	0.398	0.406	0.401	*0.400*	0.248
2010	G-L4	0.470	0.472	**0.477**§	0.476	0.478	0.477	0.477	0.404	0.424	0.326
2010	G-L5	0.469	0.485	0.471	0.469	0.476	0.486	0.469	0.479	**0.487**	0.407
2010	G-L6	0.576	0.583	0.601	0.612	0.601	0.608	0.611	**0.619**	0.618	0.310
2010	G-L7	0.520	0.552	0.557	0.564	0.541	0.556	0.565	**0.579**	0.574	0.298
2010	G-L8	0.589	0.600	0.599	0.597	0.605	0.605	0.598	0.603	**0.607**	0.540
2010	P-L1	0.327	0.334	0.327	0.327	0.326	0.333	0.327	0.327	**0.337**	0.439
2010	P-L2	0.277	0.310	0.275	0.266	0.275	0.309	0.268	**0.311**	0.307	0.436
2010	P-L3	0.461	0.466	0.461	0.462	0.459	**0.467**	0.461	0.459	**0.467**	0.416
2010	P-L4	0.314	**0.322**	0.317	0.316	0.315	0.317	0.317	0.317	0.315	0.360
	Counts	0	2	1	3	0	3	0	**7**	**6**	
		0%	9%	5%	14%	0%	14%	0%	32%	27%	

Boldfaced entries in the table indicate best model (fit) within location. Empty cells correspond to locations where the model did not converge. In italics, we report the models that converged but the Hessian matrix was not positive definite.
§ Better than second best model at forth decimal place.

(G-L1 2009) showed way lower predictive abilities than the rest of the locations. To understand these two situations, we calculated the repeatability of the trait in each location for the baseline model. The repeatability R is defined as the ratio of the between-individual component to the total phenotypic variance [Falconer and Mackay, 1996], which in our case, and following the methodology described by Nakagawa and Schielzeth [2010], corresponds to

$$R = \frac{\sigma^2_{GT}}{\sigma^2_{GT} + \sigma^2_S + \sigma^2_R + \sigma^2_B + \sigma^2_e} \tag{3.7}$$

where σ^2_{GT} is the between-groups variance and corresponds to the variance of the effect $(GT)_{hv}$ fitted as

random effect, and in the denominator, the total phenotypic variance given by the sum of the between-groups variance σ^2_{GT} and the within-groups variances, i.e. replicates within trials $(\sigma^2_S + \sigma^2_R)$ and blocks within replicates (σ^2_B) plus the residual variance (σ^2_e). The interpretation of this repeatability strictly refers to the expected within-group correlations among measurements, i.e. the agreement among measurements; thus, the gist of the definition of repeatability is related to the reproducibility of the absolute values of measurements. A slightly higher repeatability in Cycle1-2009 was observed for location G-L4 (Table 3.6), which involved more trials, i.e. more genotypes, in comparison with other locations in Germany. The trend in Cycle1-2010 was in favour of the Polish locations, which overall had more homogeneous and higher repeatabilities. We discuss the relation between repeatabilities and predictive abilities in the next section.

3.4.2 Second stage: Fitting genotypes by year vs. across years

From a methodological point of view, fitting the year effect in the GP stage was easier and more direct than accounting for the year effect in the second stage, in the sense that the model for the latter approach became too complex and the variance covariance matrix of adjusted means was not possible to be produced using the procedure HPMIXED of SAS given the high computer power required. Instead, we computed the adjusted means with corresponding standard errors, which were then used to do the weighting to pass on from the second to the third stage.

The adjusted means obtained from the across-years analysis (*Approach 2*) were plotted against the year effect-corrected genotype adjusted means (from *Approach 1b*) to compare the difference of adjustments, in the former case based on one single check against the adjustment given the simple mean of the genotypes in each year (Figure 3.4). Below the two principal lines, an observation corresponding to the shared check across years stood out from the others, reflecting the year adjustment. At first glance, it is clear that the check was the only observation pulled down implying that the year adjustment of this check was not strong enough to pull down the observations of the whole year. Both approaches were examined later using the predictive abilities obtained in the GP stage.

3.4.3 Third stage: Genomic prediction

The predictive abilities of the GP stage were taken as the definitive decision criterion for identifying the best strategy for model selection, the best model, and the most reliable approach to account for year effects, and to identify the consequences of population stratification in GP. We start by presenting results of the comparison of the approaches used for fitting the year effect, since with these we only used the

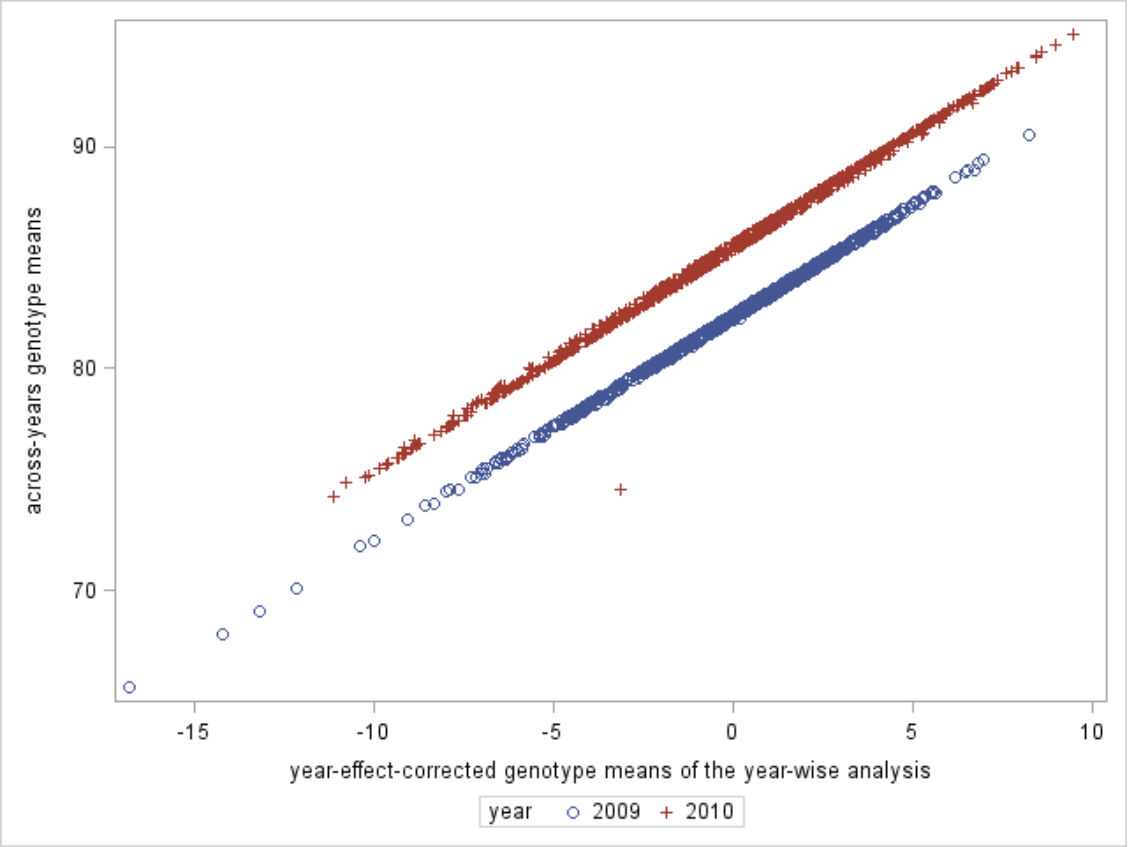


Figure 3.4: **Comparison of approaches for year adjustment.** In the x-axis, the genotype adjusted means across-year analysis are plotted. In the y-axis, the year-effect-corrected adjusted means from the year-wise analysis are depicted.

baseline model. Then we present the differences between sampling methods for CV together with the comparison of the models and the model selection strategies.

Comparison of approaches to account for year effect in GP

The GP-CV for the approach using the year as a fixed term in the third stage (*Approach 1a*) yielded a predictive ability of 0.70 (Table 3.7), whereas predictive ability for the approach accounting for a fixed year effect in the second stage (*Approach 2*) was 0.74. The predictive ability reached 0.68, using the year-effect-corrected adjusted means in the GP-CV (*Approach 1b*). The scatter plots of GEBV ($\mathbf{Z\hat{u}}$) against the observed phenotypic values (adjusted means) in the three cases are depicted in Figure 3.5. In *Approach 1a*, we plotted the GEBV against the corrected observed phenotypic values, calculated as $\mathbf{M^{(2)}} - \mathbf{X}\hat{\beta}$, where $\mathbf{M^{(2)}}$ is the vector of genotype adjusted means obtained in the second stage and $\mathbf{X}\hat{\beta}$ the predicted year effect (Figure 3.5A). For *Approach 2*, the observed phenotypic values $\mathbf{M^{(2)}}$ against $\mathbf{Z\hat{u}}$ are shown (Figure 3.5B). For *Approach 1b*, $\mathbf{M^{(2*)}}$ against $\mathbf{Z\hat{u}}$ are plotted, with $\mathbf{M^{(2*)}}$ the year-effect-corrected adjusted means of genotypes (Figure 3.5C).

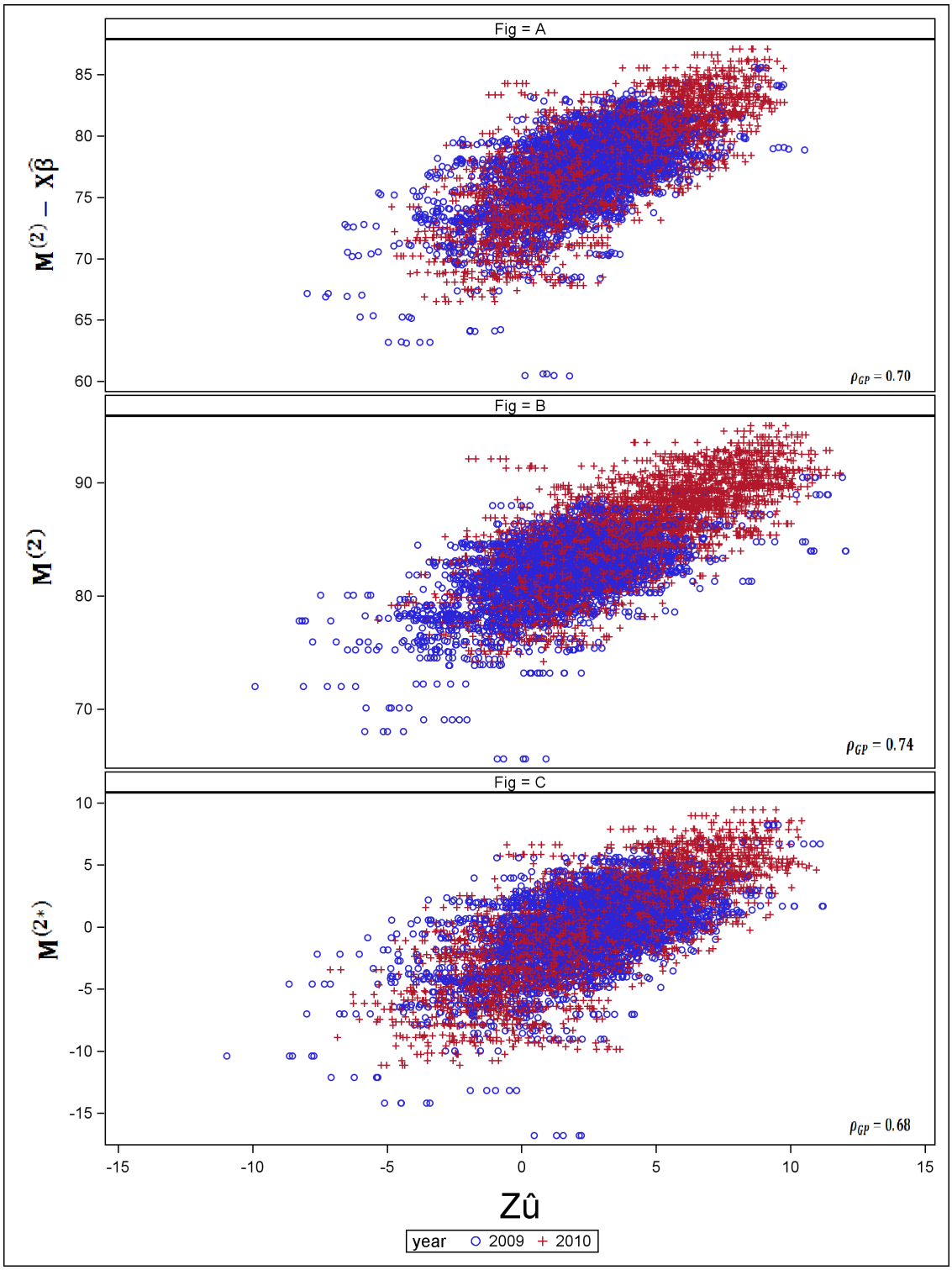


Figure 3.5: **Comparison between approaches to fit the year effect.** The y-axis represents the genotype adjusted means [$\mathbf{M^{(2)}} - \mathbf{X}\hat{\beta}$ in **(A)**, $\mathbf{M^{(2)}}$ in **(B)** and $\mathbf{M^{(2*)}}$ in **(C)**] and the x-axis represents the GEBV ($\mathbf{Z\hat{u}}$). (A) Year-wise analysis (*Approach 1a*), fitting year as fixed effect in the GP stage, (B) Across-years analysis (*Approach 2*), using year in the second stage and (C) year-wise analysis using the year effect-corrected genotype means (*Approach 1b*). ρ_{GP} represents the predictive ability.

Table 3.7: **Predictive abilities between observed and predicted values for 9 spatial and non-spatial models (M1, $\cdots$, M9) and mixed datasets using the best locations given the AIC (Mix1) and the ρ-GP-CV per location-year combination (Mix2).**

	M1	M2	M3	M4	M5	M6	M7	M8	M9	Mix1	Mix 2
WC	0.700	0.694	0.691	0.679	0.692	0.692	0.691	0.694	0.689	0.689	0.690
	a	ab	ab	c	ab	ab	ab	ab	abc	bc	abc
AC	0.395	0.398	0.390	0.395	0.391	0.389	0.389	0.395	0.391	0.391	0.390
	b	a	cd	de	c	e	de	b	c	c	cd

Same letters within rows indicate no significant differences ($\alpha = 5\%$) according to a paired t-test. Sampling strategies were: Within crosses (WC) and across crosses (AC).

Comparison of model selection strategies using different sampling methods in cross validation

Fitting model (3.6) to measure the influence of the relationship among genotypes on predictions yielded variance components for genotypes, crosses and error for year 2009 of 4.03, 3.67 and 1.66, respectively, and for year 2010 of 4.72, 10.70 and 1.32, respectively. Thus, the cross effect in 2009 is contributing in about 40% and in the next year more than 60% to the total variation explained by the data.

The marker-based relationship heat-map (Figure 3.6) shows some clusters among genotypes of the same cross indicating genetic relatedness. The predictive abilities using five times fivefold CV of datasets resulting from first stage analysis of all spatial and non-spatial models plus the mixed datasets were in general very similar within sampling strategies (Table 3.7). For the across-crosses (AC) sampling scheme, the predictive abilities were lower than the ones obtained with the within-crosses (WC) sampling scheme. In the AC sampling, we fixed the initial seed of the random number generator used for randomization in the CV procedure at the same value for all models to be able to compare the models when the same crosses were used in the training set.

We compared the models and the sampling methods using a paired t-test ($\alpha = 5\%$) by resembling a randomized complete block design, where the predictive ability of each repetition of the CV was taken as a block, thus accounting for the dependence among observations from the same samples (Table 3.7). For the first sampling method (WC), three groups were identified with some overlaps, but showing not much of a difference among models. From the across-crosses sampling strategy (AC), five groups were distinguished with some overlaps: M2 had the highest predictive ability and models M4, M6 and M7 had the worst predictive abilities.

Potential bias of GP is another important element that could be used to compare models. We

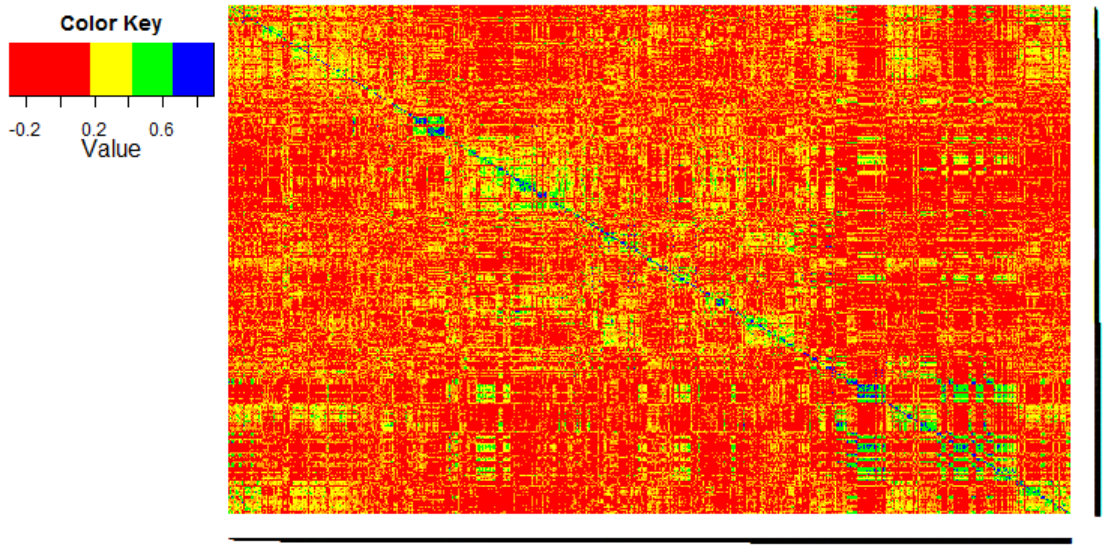


Figure 3.6: **Marker-based relationship heat-map.** Visualised are pairwise relationship coefficients estimated from the maker data for genotypes of years 2009 and 2010. Higher values represent a stronger relationship.

computed the bias as suggested by Le Roy et al. [2012] and Wang et al. [2012]. The comparison of the biases of all models followed a rather similar trend as the predictive abilities showed in Table 3.7. We present the analysis of bias as supplementary material (Appendix B.2).

The heritability (square root of heritability) for the baseline model was estimated as 0.68 (0.82) for year 2009, 0.73 (0.85) for year 2010 and 0.69 (0.83) for 2012 using the equation (3.3). In principle, the *ad hoc* method may approximate the true value of heritability but making the unrealistic assumption of uncorrelated genotypes [Estaghvirou et al., 2013]. We computed the heritability to have a rough idea of how much could we expect from the predictive abilities. The predictive ability divided by square root of heritability is an estimate of the accuracy of GP [Estaghvirou et al., 2013], and the square root of the heritability provides the upper bound for the predictive ability [Falconer and Mackay, 1996], thus one expects that the predictive abilities are not very far from the square root of heritability. In this case, the square roots of the heritabilities are somewhat larger than the corresponding predictive abilities, indicating that the predictions are not sufficiently accurate due to limited data size, thus not exhausting completely the genetic variance. To explore in which extend could have our models explained the variance not captured by the markers, we fitted an additional component accounting for the polygenic effect in the GP stage [Piepho, 2009b]. The baseline model (M1) yielded a genotypic variance of 2.99; when we incorporated the polygenic effect, the genotypic variance was 2.72 and polygenic variance was 0.36, indicating that about 88% of the total genetic variance was captured by the RR-BLUP model.

3.5 Discussion

Selecting the models at the first stage produced different results than assessing them in the third stage. AIC had better scores for the models that used row and column effects, e.g. Models M9, M6 and M2 (Table 3.5) or M8 that had a two-dimensional variance-covariance error structure. ρ-GP-CV also picked M8 and M9 (Table 3.6) but the choices were more spread over the models covering even the baseline model. In general, in the first stage, both AIC and ρ-GP-CV produced better scores for the two-dimensional models, whereas in the third stage the baseline and one-dimensional models seemed to be better than the more complex models (Table 3.7). The explanation of this pattern may be related to the second stage, where the interaction genotype $\times$ location played a role. The two-dimensional models performed very well in modelling heterogeneity within field, but when the means were integrated across the whole experiment, including all locations and years, the two-dimensional spatial error models seemed to over-adjust the means, yielding a poorer predictive ability in the GP stage. The one-dimensional spatial error models and the two-dimensional model without spatial error structure were sufficient to estimate appropriately adjusted means. This corroborates Piepho and Williams [2010] who concluded that for small portions of a field, a particular spatial model may hold well but if fitted all across the field it may fail. In a wheat experiment, Lado et al. [2013] found that using moving averages as covariable significantly improved the predictive abilities of GP. They recognised strong heterogeneous patterns of irrigation in the field, that were not controlled with a single blocking system.

Models M1, M3 and M7 were never selected as having the best fits either by AIC or ρ-GP-CV. These models had in common that none of them used rows and columns as additional factors, strengthening the conclusion that row-column designs may have the potential to correctly control field heterogeneity and thus enhance predictive ability of genomic prediction.

Fitting a location-specific error model did not have an advantage over fitting a common model across locations. Neither did the dataset composed of means computed using models have best AIC fits (Mix 1) nor the second dataset containing the means computed using models with highest ρ-GP-CV (Mix 2) produce better predictive abilities in the GP stage.

The models with nugget had better fits than the corresponding baseline model without the nugget. The drawback was that fitting those models was not straightforward, since almost every location required a separate coding specifying initial values and lower boundary constraints on the covariance parameters. Good statistical and biological reasons have been presented of why including a nugget to analysis of field experiment is beneficial [Wilkinson et al., 1983].

If we ignore the two-dimensional spatial models (M5, M6, M8 and M9), the AIC privileges M2 and

ρ-GP-CV yields more diverse results with the majority of choices for M2 and M4. In fact, when the spatial component of a resolvable row-column design based on linear variance (LV) does not lead to an improved fit, returning to classical row-column design provides randomisation protection [Williams et al., 2006].

Williams and Luckett [1988] performed studies aiming to find the optimal plot size, the optimal plot arrangements and the best spatial model (the so-called uniformity trials) and showed that in cotton and barley row and column designs are well suited for variety testing in plant breeding trials. Moreover, recent simulation studies from Möhring et al. [2014] showed that designs including rows and columns outperformed one-dimensional blocking. In the same work, the authors mention that blocking in the direction of plots with common long sides is preferable, which is common in cereal breeding [Patterson and Hunter, 1983].

We cannot affirm that ρ-GP-CV was better than AIC for model selection or vice versa, nor that the results showed the same trend; but if we would have used either of these two strategies to select the best model, we would have selected the M9 with AIC or M8 with ρ-GP-CV. The GP predictive ability obtained by M2 (Table 3.7) was slightly better than M8 and M9 (specifically AC sampling method); however, this model (M2) was not highlighted by either of the two selection criteria (AIC or ρ-GP-CV).

In practice, the fact that there were no large statistical differences is good news for the breeders because the baseline model (M1), or even better, the simplest model with row-column adjustment (M2), are appropriate for phenotypic analysis towards GP.

As a model selection method, GP-CV is of interest because it may allow to compare models with different fixed effects, even when REML is used for estimating the variance parameters. No simple recommendation has been reported concerning the best model selection criterion in the case of spatial models [Lee and Ghosh, 2009; Spilke et al., 2010]. Predictive abilities have been used between environments as similarity measure and then to join similar environments into clusters [Heslot et al., 2013b]. Thus, in a sense ρ-GP-CV allows giving an interpretation to the environment under scrutiny and the displayed trend do not depart far from the classical AIC. The repeatabilities (R) presented in parallel to the ρ-GP-CV (Table 3.6) show a low correlation ($\rho = 0.36$, p-value = 0.0965) with the predictive abilities from the baseline model. In fact, we expected that for location P-L3 of 2009, which had a negative predictability, the R was very low almost zero, but this was not the case; hence we could not conclude that the low predictive ability is mainly due to environmental effects. Riedelsheimer et al. [2013] also reported negative predictive accuracies when testing unrelated crosses in the CV procedure and observed that using unrelated crosses could have provided a negative prediction signal due to opposite linkage phases with important QTL displayed in the TS, suggesting that the negative predictive accuracies are

associated with the marker pattern.

In this study we explored three ways to adjust the year effect given the weak connectivity across years. Using the single check (*Approach 2*) to make the year adjustment was not a better choice than adjusting by the simple year mean (*Approach 1b*) or accounting for the year effect in the GP stage (*Approach 1a*), even though the estimated predictive ability was the highest. The "year clouds" produced using *Approach 2* (Figure 3.5B) did not overlap perfectly, from which we concluded that the correction was not appropriate and generated an over-fitting of the markers in the GP-CV procedure due to the fact that markers also predicted the year effect and not the SNP-effects alone. Using the year-mean correction for adjusted means in the second stage (*Approach 1b*) produced a lower ρ-GP-CV, that, given the overlay of the clouds of predicted vs. observed values, seems to be more realistic. However, fitting the year effect manually, i.e. using ordinary least squares estimation (OLSE) vs. fitting it as a fixed effect in the GP stage, i.e. using generalised least squares estimation (GLSE) can definitively yield a more precise estimate. Indeed, the residual variance in *Approach 1b* using year effect-corrected adjusted means was around 3.9 (in average for the five replicates) and in *Approach 1a* using the year fixed effect in the GP stage yielded residual variance of 3.0 (in average for the five replicates). In *Approach 1a*, where we fitted the year in the GP stage, we removed the year effect from the observed adjusted means derived from the second stage $(\mathbf{M^{(2)}} - \mathbf{X}\hat{\beta})$ to avoid bias of the predictive abilities; however, there would still be some bias because the subtracted year effect was not the true effect but an estimate of the year effect.

Models were eventually assessed and compared using the ρ-GP-CV in the third stage. The two sampling scenarios to perform the CV procedure aimed to recreate the cases where the material was genetically close, with some individuals coming from the same parental cross, and more distantly related to avoid individuals from the same parental cross in the randomisation procedure of CV. This more distantly related material shows some identical-by-state (IBS) similarity, therefore it was not unrelated in the theoretical sense of population genetics. This more distantly related scenario may be seen also as a case where one tries to predict a scenario whose linking information is weak or lacking, e.g. different genotypes and/or locations in the TS and VS [Burgueño et al., 2012; Schulz-Streeck et al., 2013a; Windhausen et al., 2012].

The predictive abilities obtained for GP using WC sampling were located in the middle-high range and using AC sampling, predictive abilities were placed in the middle range. The predictive ability of the AC sampling was significantly lower than WC, as expected for GP of a dataset showing population structure. Riedelsheimer et al. [2013] drew similar conclusions using unrelated biparental maize families. They concluded that predictive accuracy could be increased by adding crosses (families) sharing both parents to the TS. In this respect, the use of pedigree and marker information to borrow information from

both sources is suggested [Burgueño et al., 2012].

3.6 Conclusions

The main conclusions of this study are: (i) Fitting a traditional model including row and column factors across all locations was good enough to account for field heterogeneity in the first stage under GP frame. This also suggests that row-column designs may be preferable to designs with a single blocking factor; (ii) AIC and ρ-GP-CV did not have the same trend in selecting across models, but both favoured in the end models M8 and M9; however, none of the methods picked the model with highest predictive ability. Fitting a location-specific error model did not produce an advantage over fitting a common model across locations; (iii) the baseline model (M1) and the simplest row-column adjustment (M2) had in overall the best results, which is very good news since in routine analysis complex models may require much programming expertise and powerful computers; (iv) in a dataset weakly connected across years, a more reasonable model-wise structure is to account for the year factor in the genomic prediction stage rather than in a previous stage, to ensure that the effect is not confounded with the markers adjustment, and (v) datasets of distantly related genotypes may have a poor performance for GP purposes; however, increasing the size of the crosses may be an opportunity to enhance predictive ability in these cases of disconnected datasets on related sets of genotypes.

Chapter 4

Genomic prediction in early selection stages using multi-year data in a hybrid rye breeding program[3]

Angela-Maria Bernal-Vasquez[a], Andres Gordillo[b], Malthe Schmidt[b], Hans-Peter Piepho[a]

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

[b] KWS LOCHOW GMBH, Ferdinand-von-Lochow-Strasse 5, 29303 Bergen, Germany

4.1 Abstract

Background

The use of multiple genetic backgrounds across years is appealing for genomic prediction (GP) because past years' data provide valuable information on marker effects. Nonetheless, single-year GP models are less complex and computationally less demanding than multi-year GP models. In devising a suitable analysis strategy for multi-year data, we may exploit the fact that even if there is no replication of genotypes across years, there is sufficient replication at the level of marker loci. Our principal aim was to evaluate different GP approaches to simultaneously model genotype-by-year (GY) effects and breeding values using multi-year data in terms of predictive ability. The models were evaluated under different scenarios reflecting common practice in plant breeding programs, such as different degrees of relatedness between training and validation sets, and using a selected fraction of genotypes in the training set. We used empirical grain yield data of a rye hybrid breeding program. A detailed description of the

[3] A version of this chapter is published as:
Bernal-Vasquez, A.-M., Gordillo, A., Schmidt, M. and Piepho, H.-P. Genomic prediction in early selection stages using multi-year data in a hybrid rye breeding program. *BMC Genetics* (2017) 18:51.

prediction approaches highlighting the use of kinship for modeling GY is presented.

Results Using the kinship to model GY was advantageous in particular for datasets disconnected across years. On average, predictive abilities were 5% higher for models using kinship to model GY over model without kinship. We confirmed that using data from multiple selection stages provides valuable GY information and helps increasing predictive ability. This increase is on average 30% higher when the predicted genotypes are closely related with the genotypes in the training set. A selection of top-yielding genotypes together with the use of kinship to model GY improves the predictive ability in datasets composed of single years of several selection cycles.

Conclusions Our results clearly demonstrate that the use of multi-year data and appropriate modeling is beneficial for GP because it allows dissecting GY effects from genomic estimated breeding values. The model choice, as well as ensuring that the predicted candidates are sufficiently related to the genotypes in the training set, are crucial.

Keywords: Multi-year data, Genomic prediction, Genotype-by-year interaction, Hybrid rye breeding

4.2 Background

Genomic prediction (GP) is a tool for predicting genomic estimated breeding values (GEBV) of selection candidates based on marker information. A reference set of individuals, called training set (TS), is phenotyped and genotyped to train a model, which can be used to predict GEBV of another set of individuals that has only been genotyped but not phenotyped, the so-called prediction or validation set (VS) [Meuwissen et al., 2001]. Prediction performance of GP procedures can be assessed through cross validation (GP-CV). In GP-CV the datasets are divided into k folds, where *k-1* folds are used for model training and the remaining fold for model validation. This process is repeated using each of the k folds in turn as validation set and then repeating the process several times. An alternative method to evaluate prediction performance is genomic prediction - forward validation (GP-FV), which makes use of data from previous years for training the model to predict genotypes tested in later years and in this way validate the model. GP-FV mimics the ultimate goal in plant breeding, where new genotypes in new environments are to be predicted.

One of the factors determining the accuracy of the predictions is the size of the training and the validation set [Auinger et al., 2016; Rutkoski et al., 2015; Schmidt et al., 2016; Schulz-Streeck et al., 2013b]; thus using multi-year data is an attractive approach to train GP procedures because it allows increasing the TS-size, thereby potentially increasing prediction performance. But using multi-year data

is challenging because different cycles (in different years) are physically disconnected, that is, there are no genotypes in common across cycles; therefore, genotype-by-year effects (GY) and genotype main effects will be confounded. The only connection across years is genetic, i.e., through the relatedness within the material, which we expect, since the data comes from a breeding program. The genetic connectivity has been difficult to exploit with standard phenotypic models. Multi-location field trial data in breeding programs are often analyzed by year and not over years because: (i) it is simpler and faster, and (ii) it is difficult to accurately estimate variation across years, partly because few if any genotypes are common between breeding cycles. If GY effects are not properly modeled, the genomic prediction procedure will divert part of the marker information into prediction of the GY interaction effects rather than the GEBV. This situation poses the main challenge when combining data across years.

Several authors have proposed an extension of the GP model to predict genotype-by-environment interaction effects by incorporating environmental data and crop modeling [Heslot et al., 2014; Jarquín et al., 2014] or assuming a covariance matrix composed of a genotype-related and an environment-related component [Lado et al., 2016; Malosetti et al., 2016]. In these studies, environment is understood as the conditions of a given location in a given year, i.e., the conditions in a year-location combination, and no attempt is made to differentiate the effects of locations and years. Hence, year-location combinations are represented by a single factor for "environment". In the structure of the present hybrid rye breeding program, however, it is crucial to separate the location and year effects, since the program runs in the same locations across years and the interest of the breeders is in predicting the GEBV free of GY and genotype-by-location (GL) effects. Most procedures used for GP do not include model terms that dissect genotype effects, including GEBV and GY, mainly because of the lack of overlapping genotypes across years (selection cycles in the TS).

We hypothesize that in a multi-year dataset of a breeding program, where there are no common genotypes across years, GEBV can be dissected from GY based on the genetic correlation between genotypes via the kinship matrix. Further, genotypes from the same breeding cycle evaluated in multiple years in the TS will enhance the separation between GEBV and GY effects. In light of this, our principal objective was to evaluate the merit of different models accounting for the GY effect. In order to put the different models to a realistic test, we evaluated them under scenarios representing common practice in breeding programs, i.e., in different relatedness scenarios and top-yield selection scenarios, where different fractions of genotypes with top-yield performance in the TS were selected. The top-yield selection scenarios are interesting to breeders because considering only subsets of the best genotypes would allow reducing the effect of genotypes with confounded yield- and non-yield-QTL effects, i.e., genotypes whose grain yield is susceptible to be affected by diseases or lodging or other - environmentally triggered - threshold traits.

4.3 Materials and Methods

4.3.1 Phenotypic data structure

A first stage of the present hybrid rye program consists of selfing single plants and selecting for line *per se* performance in the subsequent selfing generations. After line *per se* evaluation, selected lines are crossed to one or more single crosses from the opposite gene pool. The testcross progenies are evaluated in multi-location trials [Geiger and Miedaner, 2009] to assess their general combining ability (GCA). In the first year of testcross evaluations, S_2 lines are evaluated, from which a selected fraction is subjected to a more intensive evaluation in the following year (GCA2), across a larger number of environments. Again, a selected fraction of genotypes is carried forward to a third selection stage (GCA3), where genotypes are evaluated in more environments and with more testers (See Figure C.1 for a complete selection cycle description). The minimum generation interval comprises five years, which is the time from initial crossing to GCA1. In Figure 4.1, we depict the breeding program structure to define the different GP-FV scenarios.

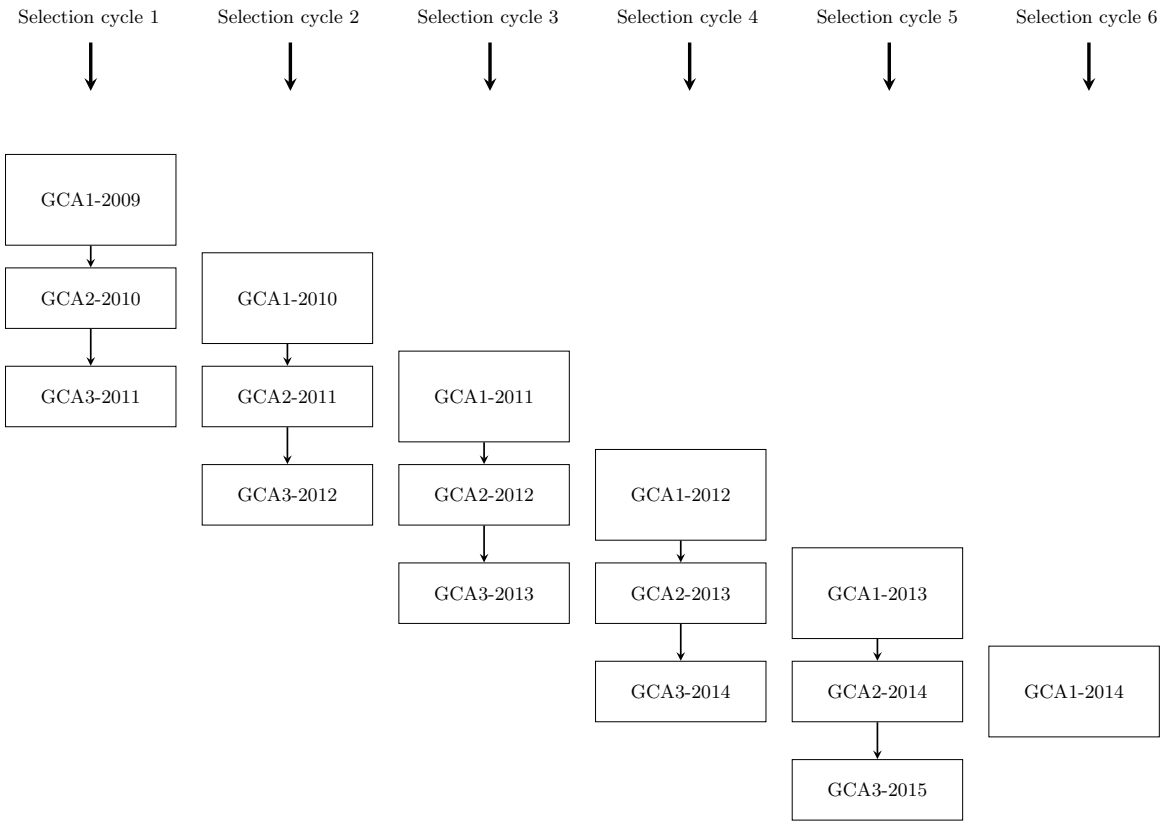


Figure 4.1: **Selection cycles structure in the rye hybrid breeding program.**

New GCA1 experiments are carried out each year with new testers from the opposite gene pool,

whereas testers remain the same across GCA1 and GCA2 experiments within the same selection cycle. At KWS-LOCHOW, a selected fraction of genotypes are test-crossed for GCA3 in combination with a different set of testers compared to GCA1 and GCA2, whereas the candidates are a selected fraction of the candidates in GCA1 and GCA2. GCA1 experiments of different selection cycles (e.g. GCA1-2009, GCA1-2010, GCA1-2011) do not normally share any genotype or check entry. Further, a GCA experiment consists of multi-environment trials (METs), where subsets of genotypes are evaluated in series of trials allocated in several locations (in one year). Within a year, trials are connected through common genotypes and check entries. Trials are laid out as α-designs with two replicates and 32 incomplete blocks of size 12 to 16.

We analyzed grain yield data from two rye hybrid breeding programs located in Germany and Poland of KWS-LOCHOW. Three datasets were formed, i.e., the German (GER) dataset, with only German lines, the Polish (PL) dataset, with only Polish lines, and the pooled dataset with German and Polish lines (GER&PL). The datasets were screened for outliers at the trial level using the method BH-MADR developed in Bernal-Vasquez et al. [2016]. The genotype sets evaluated at the GCA1 level differ between the two breeding programs. When selected candidates reach the GCA2 and GCA3 stage, they are evaluated in one common trial series across locations. We used a GP-FV approach, where GEBV of a VS with genotypes not included in the TS are predicted. We considered three scenarios that differ in the composition of their TS, different relatedness scenarios between TS and VS, and additionally, two different selection fractions for the set of top-yielding genotypes. To assess prediction performance we computed the predictive abilities of each scenario in the three datasets, i.e., GER, PL and GER&PL. Predictive abilities are defined in Subsection *Predictive abilities* of this Section.

In the scenarios described in the following, the use of GCA1, GCA2 and GCA3 data may indirectly increase the proportion of segregating first-degree relatives in the TS in comparison to a control TS composed of only GCA1 data. Each scenario is composed of three VS, one complete TS and a control TS (Figures C.2-C.4). The VS were: VS_1: GCA1-2012, VS_2: GCA1-2013 and VS_3: GCA1-2014. The control TS scenarios do not include the GCA2 and GCA3 trials. In the control TS, GCA1 data do not share common genotypes at all, thus we can evaluate if using kinship to model GY indeed helps to dissect GY from GEBV, thus allowing a more accurate predictive ability. Complete TS make use of all data in the cycle in order to check whether having this additional information about some genotypes across the years also allows to better dissect GY from GEBV with or without the use of kinship to model the GY effects. This comparison between control TS and complete TS is important because by using control TS we loose information of the common genotypes evaluated in additional years. In the complete TS, we exploit the information of those overlapping genotypes, which are very few in the end (approx. 1 to 2 % in GCA3 from the total in GCA1), but we can evaluate by cross validation whether they are sufficient to

improve the estimate of the GY effect. Since the minimum generation interval in the breeding scheme from crossing to GCA1 is five years, one would need to have breeding cycles going back at least five years to include parental lines in the TS. Hence, it is assumed that, for example, genotypes selected in GCA1-2009 are most likely to be the parents of genotypes evaluated in GCA1-2014. Thus, GCA1-2014 is likely to be more closely related to GCA1-2009 than GCA1-2013 to GCA1-2009. This theoretical relatedness cannot always become true, as the parental lines can be renewed any time or kept longer in the program. With this in mind, many TS-VS combinations can be evaluated as interesting scenarios, some being more realistic than others. Keeping the TS fixed to evaluate different VS in different years is more convenient for comparing predictive abilities, acknowledging that some TS-VS scenarios may not seem entirely realistic in that prediction is backwards rather than forwards in time. We would hold, however, that temporal direction is not crucial when evaluating predictive accuracy of a model or method.

The first scenario comprises lines from one selection cycle and corresponds to data from GCA1-2009, GCA2-2010, GCA3-2011 as TS (TS_1) to predict VS_1, VS_2 and VS_3 (Figure C.2). The control set corresponds to GCA1-2009 (controlTS_1).

The second scenario comprises lines of two selection cycles with data from GCA1-2009, GCA2-2010 (from selection cycle 1), GCA1-2010 and GCA2-2011 (from selection cycle 2) as TS (TS_2) to predict VS_1, VS_2 and VS_3 (Figure C.3). As control TS we use GCA1-2009 and GCA1-2010 (controlTS_2).

The third scenario comprises lines of three selection cycles with data from GCA1-2009, GCA2-2010, GCA3-2011 (of selection cycle 1), GCA1-2010, GCA2-2011, GCA3-2012 (of selection cycle 2), and GCA1-2011, GCA2-2012, GCA3-2013 (of selection cycle 3) as TS (TS_3) to predict VS_1, VS_2, and VS_3 (Figure C.4). The control TS contains GCA1-2009, GCA1-2010 and GCA1-2011 (controlTS_3).

To verify our hypothesis that using the kinship matrix helps to separate the GEBV from GY effects, we evaluated four different models using the complete TS (explained in the following) plus two models using the control TS of each scenario. The models were evaluated in three relatedness situations for each of the above described scenarios: all available genotypes (All-scenario) and genotypes with no (0P-scenario) and with one (1P-scenario) parent in the TS. The TS-size remains fixed and the VS-size changes according the relatedness degree with the TS. To guarantee a fair comparison with VS of the same size for the All-, 0P- and 1P-scenarios, a simple random sampling was carried out to ensure VS-size of 100 genotypes. We ran 10 iterations for VS-size $= 100$ and computed the simple means and confidence intervals of the estimated predictive abilities. The scenarios for the GER dataset with VS_1 used VS-size = 90, since there were less than 100 available genotypes. Finally, different selection fractions of top-yielding genotypes in the TS were evaluated TS composed of the 100% (Top100%), 75% (Top75%) and 50% (Top50%) best yielding genotypes, i.e. TS-sizes vary and VS-sizes remain fixed including all

available genotypes with markers.

4.3.2 Genotypic data

The marker information was obtained using a 10K Infinium iSelect HD Custom BeadChip (Illumina, San Diego, CA, USA). Monomorphic markers and markers with minor allele frequency (MAF) less than 1% or missing information of more than 10% per marker were dropped. A total of 10,633 markers passed the quality test and were used for GP. Homozygous marker genotypes were coded as -1 and 1, and the heterozygous type, missing values and technical failures were coded as 0 [Estaghvirou et al., 2014; Piepho, 2009b; Schulz-Streeck et al., 2013a].

4.3.3 Statistical models for the training sets

Mixed models are widely used for multi-environment trial (MET) analysis and can be fitted either in a single stage or in multiple stages. A single-stage analysis models the entire observed data in one stage at the level of individual plots, whereas a stage-wise analysis splits the analysis into analyses at the level of factors that are hierarchically nested, e.g., first by environments and then across environments [Piepho et al., 2012a].

The single-stage model can be stated as

$$\gamma = T : G \times Y \times L + T \cdot (G \times Y \times L) + (Y \cdot L)/S/R/B + e, \tag{4.1}$$

where γ is the vector of observed genotype yields, G represents the genotypes, T the testers, Y the years, L the locations, S the trials within locations, R the replicates within trials, B the blocks within replicates, and e the error associated with the observation γ. In the statement of model (4.1), we have used the notation described in Piepho et al. [2003], where the dot operator $(\cdot)$ defines crossed effects $(A \cdot B)$, the crossing operator $(\times)$ defines a full factorial model $(A \times B = A + B + A \cdot B)$ and the nesting operator $(/)$ indicates that a factor B is nested within another factor A $(A/B = A + A \cdot B)$. The colon $(:)$ is used to separate fixed (first) from random effects (last). Our model (4.1) takes all factors except T as random. It is therefore resolved as

$$\gamma = T : G + Y + L + G \cdot Y + G \cdot L + Y \cdot L + G \cdot Y \cdot L + G \cdot T \quad (4.2)$$
$$+T \cdot Y + T \cdot L + G \cdot T \cdot Y + G \cdot T \cdot L + T \cdot Y \cdot L + G \cdot T \cdot Y \cdot L$$
$$+Y \cdot L \cdot S + Y \cdot L \cdot S \cdot R + Y \cdot L \cdot S \cdot R \cdot B + e.$$

In routine analysis of breeding trials, it is common to analyze the data in stages. For this reason, we here also consider different stage-wise approaches. The following models are stage-wise representations of the single-stage model (4.1). They differ in the number of stages and the assumptions to model GY. As will become apparent, there are several options for stage-wise analysis and it is not obvious which option is preferable regarding our main objective to dissect GY from GEBV effects, which is why we compare different approaches. In some models, we move G to the fixed part to enable estimation of genotype means, for example in the second stage, where we then submit the means to a third stage. It is stressed here that taking G as fixed during all stages except the last is just a technical requirement to render the stage-wise analysis equivalent to the single-stage analysis, and this does not change the status of the genotype factor as random in the full stage-wise analysis [Piepho et al., 2012a]. In the models where G is kept as fixed, we will have T and G in the fixed part of the model. The interaction $G \cdot T$ is taken as random because not all genotypes are testcrossed with the same testers and because, as just mentioned, G keeps its random status in the last stage.

Note the slightly different interpretations of the main effect G depending on the context. This effect refers in general to the genotypic main effect. In the GP stage, however, where it is modeled with the marker information (i.e. using kinship), the main effect G refers specifically to the pure additive genetic part of the genotypic effect, i.e. the GEBV.

Among the models used for the control and the complete datasets, some use kinship to model GY and others not. For clarity, we differentiate approaches used for the control TS (described first with labels A1 and A1K) from the approaches using complete TS (with labels A2, A3, A4 and A5). The distinction is to point out the difference in the connectivity between the control TS and the complete TS. The control TS do not share common genotypes across years, whereas the complete TS share a fraction of selected genotypes within selection cycles, i.e., across GCA1 + GCA2 + GCA3 of the same cycle. Approaches A2 and A3 are a two-stage version of model (4.1), whereas approaches A4 and A5 have three stages. In A2, A3 and A5 we use kinship to model GY, while for the A4 approach, kinship is not used to model GY. Table 4.1 summarizes the labels, the short notation (both used indistinctly to better link the approaches in the Discussion and the Figures) and a brief description with the key elements to

distinguish the approaches. A detailed explanation of the models A1 to A5 follows next.

Table 4.1: **Summary of GP-FV approaches.**

Label	Short notation	TS used	No. stages	Use of Kinship to model GY	Description
A1	Year-wise without kinship	controlTS$_1$, controlTS$_2$, controlTS$_3$	2 + GP	no	Year-wise model and GP with year as fixed effect
A1K	Year-wise with kinship	controlTS$_2$, controlTS$_3$	2 + GP	yes	Year-wise model and GP with year as fixed effect and GY modeled using kinship
A2	2-stg-Kin	TS$_1$, TS$_2$, TS$_3$	2	yes	Across years model with GP included in the 2nd stage and GY modeled using kinship
A3	2-stg-Kin-het	TS$_1$, TS$_2$, TS$_3$	2	yes	Across years model with GP included in the 2nd stage and GY modeled using kinship. Allows heterogeneous variance among years in the GY interaction effect
A4	3-stg-NoKin	TS$_1$, TS$_2$, TS$_3$	3	no	Across years model for the TS using no kinship to model GY. Third stage is GP
A5	3-stg-Kin	TS$_1$, TS$_2$, TS$_3$	3	yes	Across years model for the TS. Uses kinship in the 2nd stage of the TS to model GY. Third stage is GP

Year-wise approach without (A1) and with (A1K) kinship: modeling for the control sets

All the control TS are composed of independent GCA1 trials in one, two or three years (controlTS$_1$, controlTS$_2$ and controlTS$_3$, respectively). We denote them as independent because the GCA1 trials have no checks in common. Thus, one approach was to estimate adjusted genotype means for each year separately in a first step and then model a fixed year effect while obtaining GEBV for genotypes in the GP stage [Bernal-Vasquez et al., 2014]. This approach presumes that the mean of the genotypes evaluated in

one year is a better year effect estimate than the year effect estimate based on a few checks shared across years. The approach is based on the assumption that the genotypes evaluated in each year are a random sample of the breeding population. Hereafter, we refer to this method as the year-wise approach (A1). One disadvantage of this approach is that it disregards annual genetic gain (1 to 2%).

In the first stage, we model the plot data within locations and years as

$$\gamma = G \cdot T : S/R/B + e, \tag{4.3}$$

which is resolved as

$$\gamma = G \cdot T : S + S \cdot R + S \cdot R \cdot B + e, \tag{4.4}$$

where factors are defined as for model (4.1). Adjusted genotype-by-tester means ($\mathbf{m}^{(1)}$) are computed for each year-location combination and are submitted to the second stage, where adjusted genotype means ($\mathbf{m}^{(2)}$) are calculated , using a year-wise model defined as

$$\begin{aligned} \mathbf{m}^{(1)} =& G + T : G \cdot T + L \cdot (G \times T) + \epsilon_1 \\ =& G + T : G \cdot T + L \cdot G + L \cdot T + L \cdot G \cdot T + \epsilon_1. \end{aligned} \tag{4.5}$$

All terms are defined as for model (4.1), ϵ_1 is the vector of errors associated with the adjusted means $\mathbf{m^{(1)}}$ with $\epsilon_\mathbf{1} \sim N(\mathbf{0}, \mathbf{R_1})$ and $\mathbf{R_1}$ is a diagonal matrix whose diagonal elements are computed from the inverse of the variance-covariance matrix estimated in the first stage [Smith et al., 2001]. Hereafter, $\mathbf{m^{(x)}}$ always denotes the adjusted mean and $\mathbf{R_x}$ always denotes a diagonal matrix carrying over these diagonal weights computed in the x$-th$ stage. The model at the GP stage is then

$$\mathbf{m^{(2)}} = \mathbf{X}\beta + \mathbf{Z_g u_g} + \epsilon_\mathbf{2}, \tag{4.6}$$

where $\mathbf{m^{(2)}}$ is the vector of adjusted genotype means across years, $\mathbf{X}$ is the design matrix of the years, β is the vector of year effects, $\mathbf{Z_g}$ is the marker matrix for genotypes, and $\mathbf{u_g}$ the vector of marker effects. We assume that $\mathbf{u_g} \sim N(\mathbf{0}, \mathbf{I}\sigma^\mathbf{2}_\mathbf{u_g})$, and $var(\mathbf{Z_g u_g}) = \mathbf{Z_g Z_g^T}\sigma^\mathbf{2}_\mathbf{u_g}$. Furthermore, $\epsilon_\mathbf{2}$ is the vector of errors associated with the adjusted means $\mathbf{m^{(2)}}$ with $\epsilon_\mathbf{2} \sim N(\mathbf{0}, \mathbf{R_2})$.

The alternative approach is to additionally model the GY effects in the GP stage. Hereafter, we refer to this strategy as the year-wise with kinship approach (A1K). Given the disconnectedness of the genotypes across years in GCA1 trials, dissecting the genotype main effects G (the GEBV) and the GY

becomes difficult. If kinship information is included to model the genotypic correlation among relatives, it may be possible to dissect the G and GY effects, provided that genotypes tested in different years can be regarded as representative of the same breeding population, which is usually the case. A slight bias will be incurred though due to genetic progress, but this can be tolerated if more than outweighed by the improved precision of the year effect estimate. The key idea behind the use of kinship to dissect the GY effects is that, while there is no replication of genotypes across years, there is plenty of replication across years at the level of genes and their alleles.

The model for the GP is

$$\mathbf{m^{(2)}} = \mathbf{X}\beta + \mathbf{Z_g u_g} + \mathbf{Z_{gy} u_{gy}} + \epsilon_{\mathbf{2}}, \tag{4.7}$$

where $\mathbf{m^{(2)}}$, $\mathbf{X}\beta$ and $\mathbf{Z_g u_g}$ are defined as for model (4.6). The GY effects are modeled as $\mathbf{w} = \mathbf{Z_{gy} u_{gy}}$, with $\mathbf{Z_{gy}}$ as the marker matrix for genotypes-by-year effects and $\mathbf{u_{gy}}$ the vector of marker-by-year effects whose variance is $var(\mathbf{u_{gy}}) = \mathbf{I}\sigma^{\mathbf{2}}_{\mathbf{u_{gy}}}$, and hence $var(\mathbf{w}) = \mathbf{Z_{gy} Z^T_{gy}}\sigma^{\mathbf{2}}_{\mathbf{u_{gy}}}$.

In particular, $\mathbf{Z_{gy}}$ is a block-diagonal matrix with blocks given by the marker coefficient matrices of genotypes in a given year ($\mathbf{Z_{gy_r}}$), e.g.,

$$\mathbf{Z_{gy}} = \begin{pmatrix} \mathbf{Z_{gy_1}} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{Z_{gy_2}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{Z_{gy_3}} \end{pmatrix}.$$

Under the mixed model formulation of ridge regression, $\mathbf{Z_{gy} Z^T_{gy}}\sigma^{\mathbf{2}}_{\mathbf{gy}}$ represents the linear structure of the genotype-by-year variance-covariance matrix with covariance of two genotypes within the same year depending on the similarity in their marker profiles [Piepho, 2009b]. Note that the covariance among different years is zero. Any covariance between years is captured by the main effect for genotypes via the $\mathbf{Z_g}$ matrix.

Two-stage approach with kinship matrix: 2-stg-Kin (A2)

The single-stage model (4.1) can be estimated in a two-stage analysis, where adjusted genotype-tester means by locations and years are computed in the first stage, and then in the second stage, adjusted genotype means across locations and years are calculated. GP-FV can be incorporated in this second stage, allowing to compute GEBVs for a set of genotypes that belong to a new year, i.e. the VS.

The first stage remains as for the previous approaches and is described by model (4.3). The second-stage model is

$$\mathbf{m^{(1)}} = T : G \times Y \times L + T \cdot (G \times Y \times L) + \epsilon_1. \tag{4.8}$$

The model is fitted using the adjusted genotype-by-tester means $\mathbf{m^{(1)}}$ for the different year-location combinations computed in the first stage. The four-way factorial in model (4.8) is resolved as

$$\begin{aligned} T : G + Y + L + G \cdot Y + G \cdot L + Y \cdot L + T \cdot Y + T \cdot L + G \cdot T \\ + G \cdot Y \cdot L + G \cdot T \cdot Y + G \cdot T \cdot L + T \cdot Y \cdot L + G \cdot T \cdot Y \cdot L. \end{aligned} \tag{4.9}$$

Hence, the second-stage model (4.8) can be written as

$$\mathbf{m^{(1)}} = \mathbf{1}\mu + \mathbf{X}\beta + \mathbf{Z_g u_g} + \mathbf{Z_{gy} u_{gy}} + \mathbf{Z_b u_b} + \epsilon_\mathbf{1}, \tag{4.10}$$

where $\mathbf{m^{(1)}}$ is the vector of adjusted genotype-tester means obtained in the first stage [model (4.3)], $\mathbf{1}$ is a $m \times 1$ vector of ones with m the number of genotypes, μ is the intercept, $\mathbf{X}$ is the design matrix for fixed effects, β is the vector of fixed-effects parameters. The tester (T) is the only fixed effect in model (4.9). The GEBV G is equivalent to $\mathbf{v} = \mathbf{Z_g u_g}$, with $\mathbf{Z_g}$ the marker matrix for genotypes and $\mathbf{u_g}$ the vector of marker effects whose variance is $var(\mathbf{u_g}) = \mathbf{I}\sigma^\mathbf{2}_\mathbf{u_g}$, and hence $var(\mathbf{v}) = \mathbf{Z_g Z_g^T}\sigma^\mathbf{2}_\mathbf{u_g}$. Similarly, the genotype-by-year effect $G \cdot Y$ is equivalent to $\mathbf{w} = \mathbf{Z_{gy} u_{gy}}$, where $\mathbf{Z_{gy}}$ is the marker matrix for genotypes-by-year and $\mathbf{u_{gy}}$ is the vector of marker-by-year effects whose variance is assumed to be $var(\mathbf{u_{gy}}) = \mathbf{I}\sigma^\mathbf{2}_\mathbf{u_{gy}}$, then $var(\mathbf{w}) = \mathbf{Z_{gy} Z^T_{gy}}\sigma^\mathbf{2}_\mathbf{u_{gy}}$. $\mathbf{Z_b}$ is the design matrix for the other random effects between years and $\mathbf{u_b}$ is the vector of random effects between years, which includes the effects of $G \times Y \times L + T \cdot (G \times Y \times L)$ except G and $G \cdot Y$. Thus, $\mathbf{u_b} = (\mathbf{u^T_{b(1)}}, \mathbf{u^T_{b(2)}}, ..., \mathbf{u^T_{b(t)}})^\mathbf{T}$ with $\mathbf{u_{b(k)}}$ the vector of the k-th random effect between years, and $var(\mathbf{u_b}) = \mathbf{\Sigma_b} = \oplus^\mathbf{t}_\mathbf{k=1}\mathbf{\Sigma_{b(k)}}$ with $var(\mathbf{u_{b(k)}}) = \mathbf{\Sigma_{b(k)}} = \mathbf{I}\sigma^\mathbf{2}_\mathbf{b(k)}$. The symbol $\oplus$ denotes the direct sum of matrices and defines block diagonal matrices [Searle et al., 1992]. The vector of errors is $\epsilon_\mathbf{1}$ with $\epsilon_\mathbf{1} \sim N(\mathbf{0}, \mathbf{R_1})$.

Two-stage approach with kinship matrix and heterogeneous variance: 2-stg-Kin-het (A3)

In this approach, we allow heterogeneity among years in the variance of the interaction $G \cdot Y$. Thus, for model (4.10) we assume $var(\mathbf{u_{gy}}) = \mathbf{\Lambda} = \oplus^\mathbf{m}_\mathbf{r=1}\mathbf{I}\sigma^2_{u_{gy_{(r)}}}$, where $\sigma^2_{u_{gy_{(r)}}}$ is the genotype-by-year variance

in the r-th year with the genotype entries sorted by year. If $\mathbf{w} = \mathbf{Z_{gy}u_{gy}}$, then $var(\mathbf{w}) = \mathbf{Z_{gy}\Lambda Z_{gy}^T}$.

Three-stage approach without kinship: 3-stg-NoKin (A4)

A three-stage approach for GP-FV may alleviate the computational burden imposed by using a two-stage model. In practice, plant breeders often use the following three-stage approach: In the first stage adjusted genotype-tester means ($\mathbf{m^{(1)}}$) are estimated per year-location combination using model (4.3). In the second stage adjusted genotype means across years and locations ($\mathbf{m^{(2)}}$) are estimated using the model

$$\mathbf{m^{(1)}} = \mathbf{X}\beta + \mathbf{Z_b u_b} + \epsilon_\mathbf{1}, \tag{4.11}$$

where $\mathbf{X}$ is the design matrix for fixed effects β. We need G to be fitted as a fixed effect (together with T), since we are estimating adjusted genotype means. Except for overlapping genotypes across different selection stages (GCA1, GCA2, GCA3), within the same selection cycles, the $G \cdot Y$ variance component is completely confounded with that for G under this model. $\mathbf{Z_b}$ and $\mathbf{u_b}$ are the design matrix and vector for the random effects between years, respectively. The vector includes all random effects indicated in model (4.8) except G. $\mathbf{u_b}$ is equivalent to $(\mathbf{u_{b(1)}^T}, \mathbf{u_{b(2)}^T}, ..., \mathbf{u_{b(t)}^T})^\mathbf{T}$ with $\mathbf{u_{b(k)}}$ the vector of the k-th random between-year effects. The variance is $var(\mathbf{u_b}) = \mathbf{\Sigma_b} = \oplus_{\mathbf{k=1}}^{\mathbf{t}} \mathbf{\Sigma_{b(k)}}$ where $var(\mathbf{u_{b(k)}}) = \mathbf{\Sigma_{b(k)}} = \mathbf{I}\sigma^\mathbf{2}_\mathbf{b(k)}$. This means, $G \cdot Y$, for example, is synonymous with $\mathbf{Z_{b_{(1)}} u_{b_{(1)}}}$, where $\mathbf{Z_{b_{(1)}}}$ is the design matrix for genotype-by-year effects and $\mathbf{u_{b_{(1)}}}$ the vector of random genotype-by-year effects with $var(\mathbf{u_{b_{(1)}}}) = \mathbf{I}\sigma^\mathbf{2}_\mathbf{b_{(1)}}$. The vector of errors associated with the records of $\mathbf{m^{(1)}}$ is $\epsilon_\mathbf{1}$ with $\epsilon_\mathbf{1} \sim N(\mathbf{0}, \mathbf{R_1})$.

Finally, in the third stage, the GP model is implemented as

$$\mathbf{m^{(2)}} = \mathbf{1}\mu + \mathbf{Z_g u_g} + \epsilon_\mathbf{2}, \tag{4.12}$$

where $\mathbf{m^{(2)}}$ is the vector of adjusted genotype means across locations and years, $\mathbf{1}$ is a $m \times 1$ vector of ones, with m the number of genotypes, μ is the intercept, $\mathbf{Z_g}$ the marker matrix for genotypes, and $\mathbf{u_g}$ the vector of marker effects. We assume $\mathbf{u_g} \sim N(\mathbf{0}, \mathbf{I}\sigma^\mathbf{2}_\mathbf{u_g})$, thus $var(\mathbf{Z_g u_g}) = \mathbf{Z_g Z_g^T}\sigma^\mathbf{2}_\mathbf{u_g}$. The vector of errors is $\epsilon_\mathbf{2}$ with $\epsilon_\mathbf{2} \sim N(\mathbf{0}, \mathbf{R_2})$.

The difference between the two-stage (A2, and A3) and the three-stage (A4) approaches [using model (4.10) and model (4.12)] for GP-FV is the estimation of the GY effects, which in the first case makes use of the kinship matrix, whereas in the second case kinship is ignored.

Three-stage approach with kinship in the second stage: 3-stg-Kin (A5)

The three-stage approach can also make use of the kinship matrix in the second stage to dissect GY from G main effects.

The second-stage model is written as

$$\mathbf{m^{(1)}} = \mathbf{X}\beta + \mathbf{Z_{gy}u_{gy}} + \mathbf{Z_b u_b} + \epsilon_\mathbf{1}, \tag{4.13}$$

where $\mathbf{X}$ is the design matrix for fixed effects β. We keep G and T as fixed effects. $\mathbf{Z_b}$ is the design matrix and $\mathbf{u_b}$ is the vector of random effects between years for the random effects except the GY effects, for which we use $\mathbf{Z_{gy}u_{gy}}$, where $\mathbf{Z_{gy}}$ is the marker matrix for genotypes-by-year effects and $\mathbf{u_{gy}}$ is the vector of marker-by-year effects whose variance is $var(\mathbf{u_{gy}}) = \mathbf{I}\sigma^\mathbf{2}_\mathbf{u_{gy}}$, such that $var(\mathbf{w}) = \mathbf{Z_{gy}Z^T_{gy}}\sigma^\mathbf{2}_\mathbf{u_{gy}}$. The vector of errors associated with the records of $\mathbf{m^{(1)}}$ is $\epsilon_\mathbf{1}$ with $\epsilon_\mathbf{1} \sim N(\mathbf{0}, \mathbf{R_1})$. The third stage is the same as for the 3-stg-NoKin approach [model (4.12)] using the adjusted genotype means computed in the previous stage.

4.3.4 Calculation of predictive ability - models for validation sets

Predictive abilities (ρ_{GP}) were estimated as the Pearson correlation coefficient between the adjusted genotype means of the VS ($\mathbf{m}^{(2)}$) and the GEBV ($\hat{\mathbf{v}} = \mathbf{Z}\hat{\mathbf{u}}$). To estimate $\mathbf{m}^{(2)}$ (adjusted genotype means) of the VS, we used a two-stage analysis, with model (4.3) as first stage to obtain adjusted genotype-tester means ($\mathbf{m}^{(1)}$) across locations and years. In the second stage, the adjusted genotype means $\mathbf{m}^{(2)}$ were estimated for VS_1:GCA1-2012 and VS_3:GCA1-2014 using the model

$$\begin{aligned}\mathbf{m}^{(1)} =& G + T : G \cdot T + L \cdot (G \times T) + \epsilon_1 \\ =& G + T : G \cdot T + L + L \cdot G + L \cdot T + L \cdot G \cdot T + \epsilon_1,\end{aligned} \tag{4.14}$$

where all terms are defined as for model (4.1). For VS_2:GCA1-2013, we did not include a location L main effect or a genotype-by-location effect $G \cdot L$ because testers and locations were totally confounded, thus the effect $L \cdot T$ represents $L + L \cdot T$ and $G \cdot L \cdot T$ represents $G \cdot L + G \cdot L \cdot T$. The model is

$$\mathbf{m}^{(1)} = G + T : G \cdot T + L \cdot T + G \cdot T \cdot L + \epsilon_1. \tag{4.15}$$

Adjusted genotype means based on models (4.14) and (4.15) (corresponding to VS_1 and VS_3, and

VS_2, respectively) are computed using best linear unbiased estimation (BLUE). Hence, predictive ability in each scenario was the Pearson correlation coefficient between the GEBV ($\hat{\mathbf{v}}$) from models (4.6), (4.7), (4.10) or (4.12) and $\mathbf{m}^{(2)}$ of the VS from models (4.14) and (4.15), i.e.

$$\rho_{GP} = corr(\hat{\mathbf{v}}, \mathbf{m}^{(2)}). \tag{4.16}$$

4.4 Results

4.4.1 Structure of datasets and variance components

Variance components were estimated using the two-stage model (4.8) for all datasets (GER&PL, GER and PL), the three complete TS (TS_1 [one cycle data], TS_2 [two cycles data] and TS_3 [three cycles data]) and the three VS (VS_1:GCA1-2012, VS_2:GCA1-2013 and VS_3:GCA1-2014) (Table 4.2). The expected confounding of some effects due to the unbalancedness of the trials and the poor connectivity across cycles and between TS and VS is illustrated by the asymptotic correlation matrix for variance component estimates computed from the information matrix [Searle et al., 1992, p. 248], e.g. for the GER&PL dataset TS_1-VS_3 (Table C.2 lower diagonal).

The correlation between variance component estimates for G and GY is -0.8747, for L and YL it is -0.2556, for GL and GYL it is -0.9229, for GTL and $GTYL$ it is -0.9758 and between GT and GTY it is -0.9491. The confounding is also observed in the asymptotic correlation matrix for variance component estimates of the TS_1 scenarios (Tables C.3 and C.4). For the TS_2 (Tables C.5-C.7) and the TS_3 (Tables C.8-C.10) scenarios, the confounding is still visible, though in rather lower magnitudes.

An asymptotic correlation of $\simeq -1$ indicates ill-conditioning [Pinheiro and Bates, 2000, p156]. Confounding of effects is the limiting case of ill-conditioning when the asymptotic correlation between two effects is exactly -1. It is clear that the extreme unbalancedness of the datasets renders variance component estimates unstable, in the sense that a few genotypes in the analysis impact strongly on the relative contribution of each effect to the total variance.

Additionally, variance components for genotype main effects (G) in the PL dataset are most of the times estimated as zero as well as for GL interaction effects, reflecting the poor connectivity of the datasets. The asymptotic correlations between the variance component estimates of GL and genotype-by-year-by-location interaction (GYL) effects were marginally more negative for the Polish scenarios than for the German ones (Table 4.2). This could be due to a different trial allocation across years and locations in Poland than in Germany. The GER dataset has more locations per year that are not

Table 4.2: **Summary of variance component estimates in the three datasets.** German and Polish together (GER&PL), only German (GE) and only Polish (PL), for all the training set (TS) and validation set (VS) combinations. Reported effects use the factors: Genotypes (G), year (Y) and location (L). ac(GL,GYL) is the asymptotic correlation between variance component estimates of GL and GYL effects. na represents non-estimable values due to a zero value of a variance component.

Dataset	TS	VS	G	GY	L	GL	YL	GYL	ac(GL,GYL)
GER&PL	TS_1	VS_1	0.00	6.44	145.10	0.00	93.30	4.48	na
GER&PL	TS_1	VS_2	2.29	2.19	109.86	1.36	161.58	3.71	-0.89
GER&PL	TS_1	VS_3	6.45	2.72	166.48	2.41	117.31	5.08	-0.92
GER	TS_1	VS_1	6.75	0.58	143.57	1.11	92.65	3.83	-0.89
GER	TS_1	VS_2	3.74	1.04	113.46	1.08	169.73	4.03	-0.88
GER	TS_1	VS_3	4.55	0.93	173.53	1.41	108.66	4.68	-0.92
PL	TS_1	VS_1	0.00	5.68	160.05	0.00	85.39	4.28	na
PL	TS_1	VS_2	0.00	3.41	108.72	1.72	155.03	3.03	-0.90
PL	TS_1	VS_3	0.00	11.28	173.99	3.24	94.82	5.17	-0.98
GER&PL	TS_2	VS_1	5.85	1.77	132.51	0.80	89.24	3.17	-0.96
GER&PL	TS_2	VS_2	4.18	1.54	110.06	1.27	149.52	2.78	-0.96
GER&PL	TS_2	VS_3	7.42	1.56	166.22	1.60	108.97	3.92	-0.97
GER	TS_2	VS_1	8.00	0.29	142.97	1.15	89.21	3.06	-0.93
GER	TS_2	VS_2	5.98	0.44	112.15	1.49	161.93	2.92	-0.94
GER	TS_2	VS_3	6.89	0.13	172.96	1.62	109.00	3.44	-0.93
PL	TS_2	VS_1	0.00	6.12	135.17	0.00	84.60	4.17	na
PL	TS_2	VS_2	0.00	4.22	89.73	0.004	155.83	4.00	-0.97
PL	TS_2	VS_3	0.00	9.97	158.31	0.00	92.84	6.13	na
GER&PL	TS_3	VS_1	2.24	4.53	163.69	0.68	86.92	3.89	-0.87
GER&PL	TS_3	VS_2	5.09	1.51	159.44	1.11	93.36	4.07	-0.81
GER&PL	TS_3	VS_3	7.32	1.02	176.06	1.18	85.59	4.84	-0.86
GER	TS_3	VS_1	7.19	1.10	170.60	0.78	86.35	3.66	-0.80
GER	TS_3	VS_2	7.02	0.38	186.59	1.18	84.42	4.14	-0.80
GER	TS_3	VS_3	7.01	0.32	166.34	1.16	88.32	3.69	-0.76
PL	TS_3	VS_1	0.00	5.33	156.13	0.77	84.70	3.80	-0.94
PL	TS_3	VS_2	0.67	5.00	144.19	0.97	93.47	4.10	-0.85
PL	TS_3	VS_3	5.19	3.61	161.72	0.99	81.25	5.20	-0.90

TS_1: GCA1-2009 + GCA2-2010 + GCA3-2011, TS_2: GCA1-2009 + GCA2-2010 + GCA1-2010 + GCA2-2011, TS_3: GCA1-2009 + GCA2-2010 + GCA3-2011 + GCA1-2010 + GCA2-2011 + GCA3-2012 + GCA1-2011 + GCA2-2012 + GCA3-2013, VS_1: GCA1-2012, VS_2: GCA1-2013, VS_3: GCA1-2014

repeated across the other years, whereas in the PL dataset fewer locations are used across years, that is, more locations are repeated across years, i.e., the number of location-year combinations compared to the number of total locations across years are greater in the GER than in the PL datasets (Table C.1).

This situation reflects more confounding for the PL dataset, and as a consequence, the PL dataset does not have as many GL or GYL effects as the GER dataset, so that asymptotic correlations between the variance estimates for GL and GYL effects are slightly higher in absolute value for the PL program than for the GER program (Table 4.2). The confounding is diminished when more years are used in the TS because the number of year-location combinations increases.

4.4.2 Predictive abilities

Predictive abilities were calculated using equation (4.16) (Figures 4.2-4.4). Notice that the year-wise with kinship approach (A1K) is not used for controlTS_1 because the control TS is composed of only one year, thus fitting a GY effect would over-parametrize the model.

There are years or cycles that are easier to predict than others. Predicting the VS_1:GCA1-2012 had, across all datasets, the highest predictive abilities. VS_2:GCA1-2013 had also relatively high ρ_{GP} compared to VS_3:GCA1-2014.

There was a marginal increase in ρ_{GP} along the approaches from TS covering data from two and three selection cycles (TS_2 and TS_3) over TS_1 (one selection cycle). In the GER&PL program, this increase is observed especially for VS_1 and for the 1P-scenario of VS_2. In Germany the difference between TS_2 and TS_3 is small, though there is a general increase of the predictive ability in these two datasets over TS_1. In the PL dataset, ρ_{GP} obtained using TS_3 or TS_2 are not always better than TS_1. They depend on the model and the VS used.

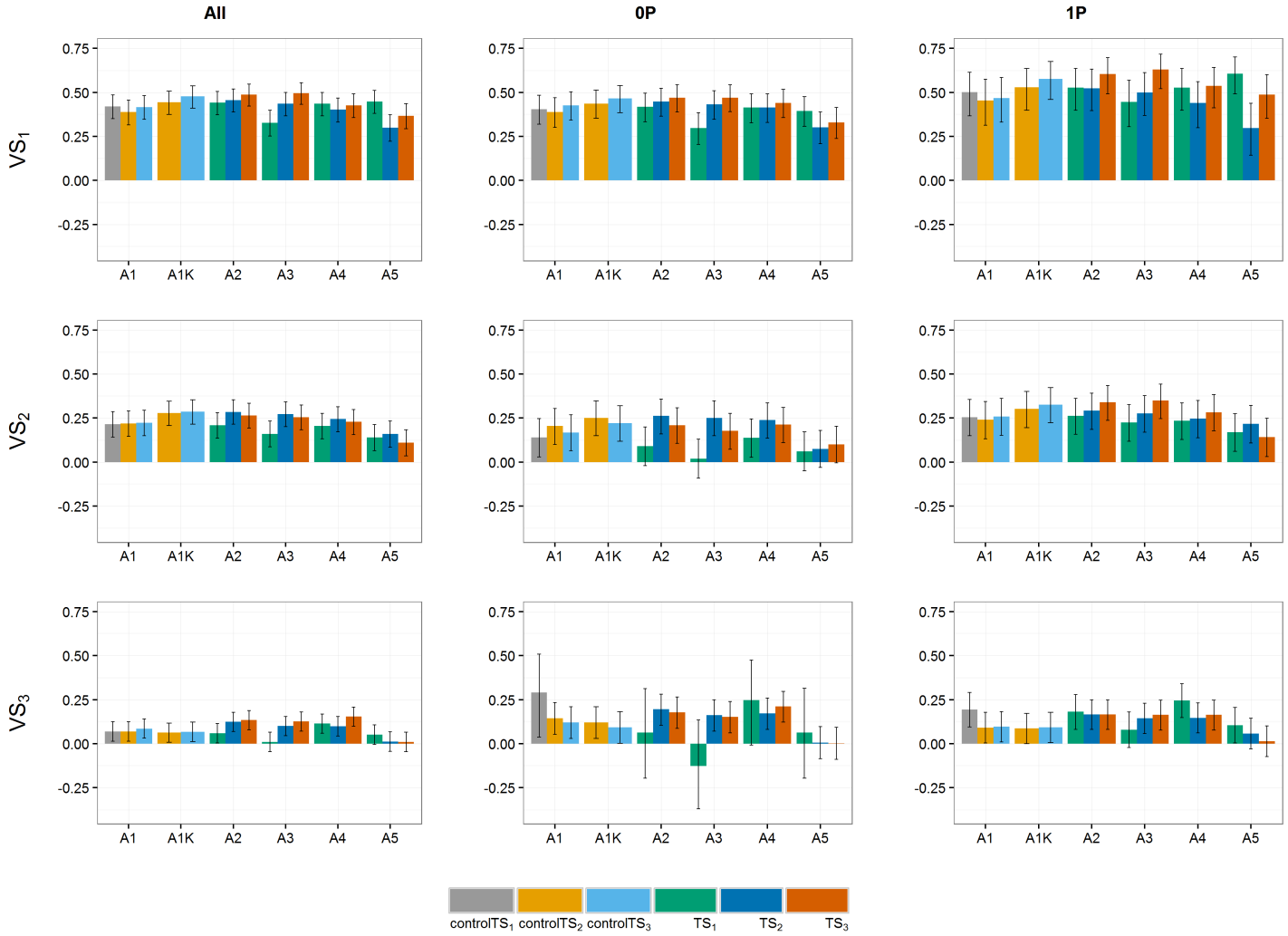


Figure 4.2: **Predictive abilities (y-axis) of the German and Polish dataset for the three scenarios.** TS_1 and controlTS_1, TS_2 and controlTS_2, and TS_3 and controlTS_3 to predict the validation sets VS_1, VS_2 and VS_3 with All, 0P and 1P-scenarios. Black lines for each bar represent the 95% confidence intervals of the predictive ability. Year-wise approach (A1) and year-wise with kinship approach (A1K) were fitted to the control sets, approaches 2-stg-Kin (A2), 2-stg-Kin-het (A3), 3-stg-NoKin (A4) and 3-stg-Kin (A5) to the complete sets. TS_1: GCA1-2009 + GCA2-2010 + GCA3-2011, controlTS_1: GCA1-2009, TS_2: GCA1-2009 + GCA2-2010 + GCA1-2010 + GCA2-2011, controlTS_2: GCA1-2009 + GCA1-2010, TS_3: GCA1-2009 + GCA2-2010 + GCA3-2011 + GCA1-2010 + GCA2-2011 + GCA3-2012 + GCA1-2011 + GCA2-2012 + GCA3-2013, controlTS_3: GCA1-2009 + GCA1-2010 + GCA1-2011, VS_1: GCA1-2012, VS_2: GCA1-2013, VS_3: GCA1-2014.

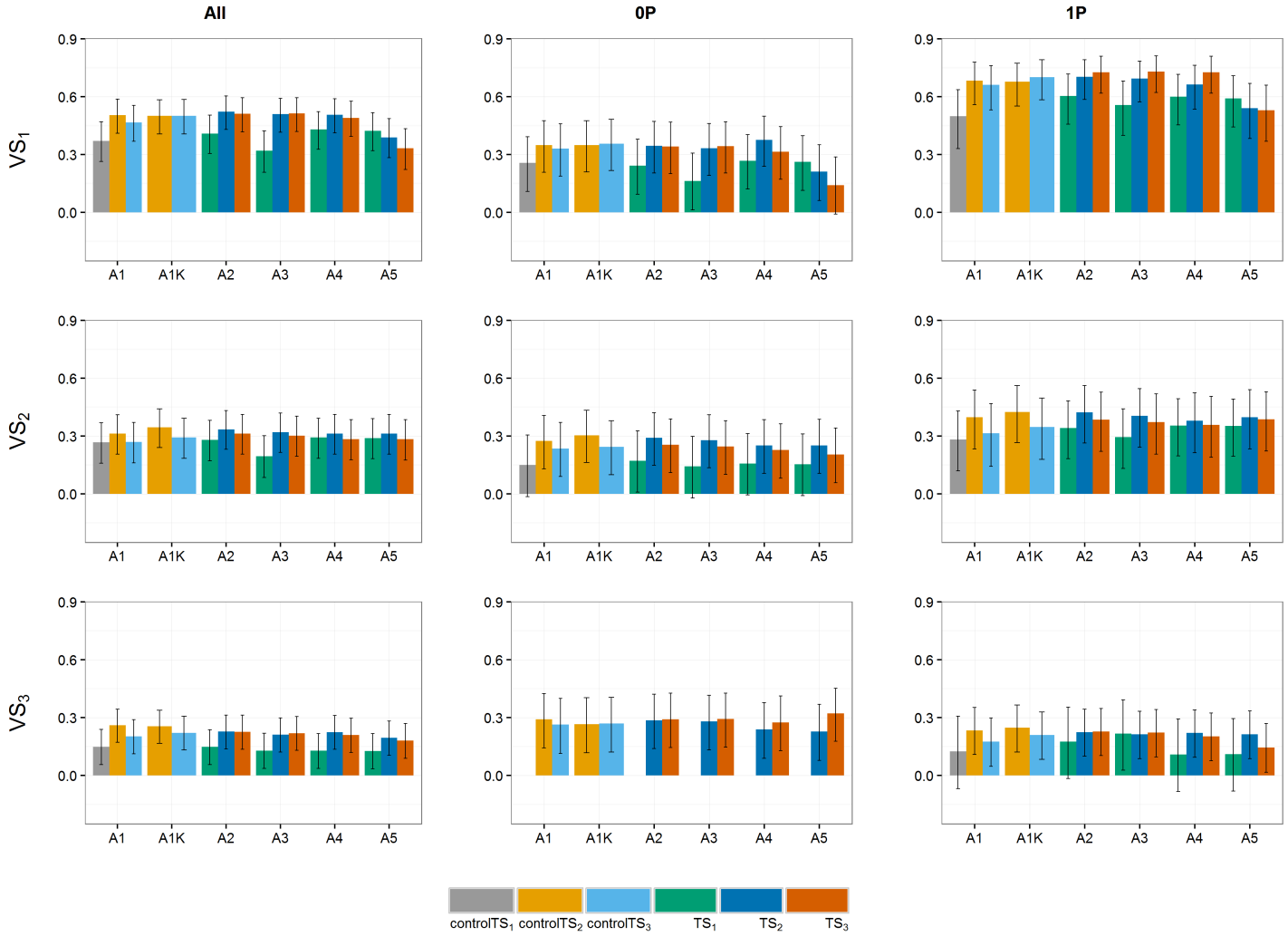


Figure 4.3: **Predictive abilities (y-axis) of the German dataset for the three scenarios.** TS_1 and controlTS_1, TS_2 and controlTS_2, and TS_3 and controlTS_3 to predict the validation sets VS_1, VS_2 and VS_3 with All-, 0P- and 1P-scenarios. Black lines for each bar represent the 95% confidence intervals of the predictive ability. Year-wise approach (A1) and year-wise with kinship approach (A1K) were fitted to the control sets, approaches 2-stg-Kin (A2), 2-stg-Kin-het (A3), 3-stg-NoKin (A4) and 3-stg-Kin (A5) to the complete sets. TS_1: GCA1-2009 + GCA2-2010 + GCA3-2011, controlTS_1: GCA1-2009, TS_2: GCA1-2009 + GCA2-2010 + GCA1-2010 + GCA2-2011, controlTS_2: GCA1-2009 + GCA1-2010, TS_3: GCA1-2009 + GCA2-2010 + GCA3-2011 + GCA1-2010 + GCA2-2011 + GCA3-2012 + GCA1-2011 + GCA2-2012 + GCA3-2013, controlTS_3: GCA1-2009 + GCA1-2010 + GCA1-2011, VS_1: GCA1-2012, VS_2: GCA1-2013, VS_3: GCA1-2014.

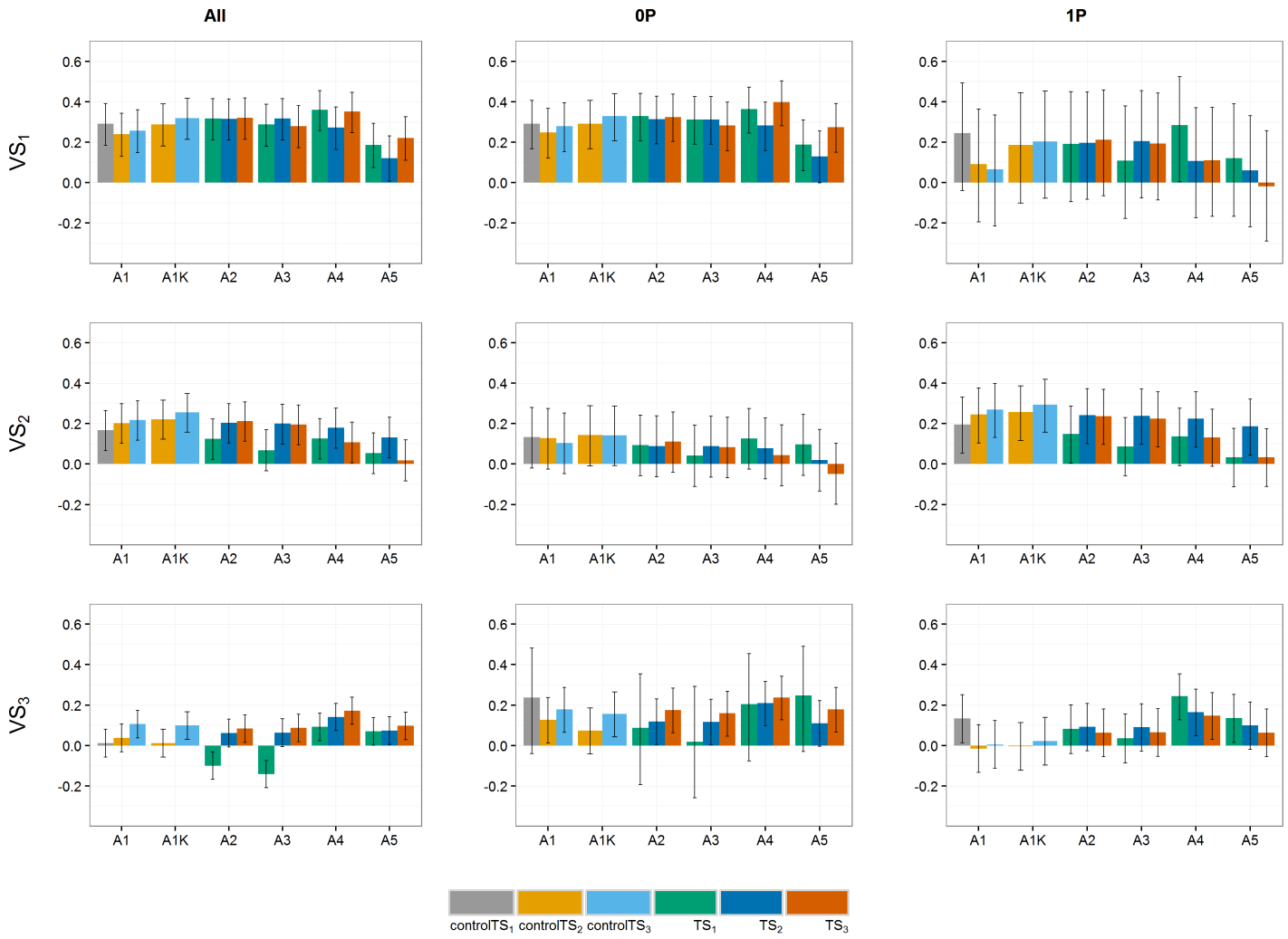


Figure 4.4: **Predictive abilities (y-axis) of the Polish dataset for the three scenarios.** TS_1 and controlTS_1, TS_2 and controlTS_2, and TS_3 and controlTS_3 to predict the validation sets VS_1, VS_2 and VS_3 with All-, 0P- and 1P-scenarios. Black lines for each bar represent the 95% confidence intervals of the predictive ability. Year-wise approach (A1) and year-wise with kinship approach (A1K) were fitted to the control sets, approaches 2-stg-Kin (A2), 2-stg-Kin-het (A3), 3-stg-NoKin (A4) and 3-stg-Kin (A5) to the complete sets. TS_1: GCA1-2009 + GCA2-2010 + GCA3-2011, controlTS_1: GCA1-2009, TS_2: GCA1-2009 + GCA2-2010 + GCA1-2010 + GCA2-2011, controlTS_2: GCA1-2009 + GCA1-2010, TS_3: GCA1-2009 + GCA2-2010 + GCA3-2011 + GCA1-2010 + GCA2-2011 + GCA3-2012 + GCA1-2011 + GCA2-2012 + GCA3-2013, controlTS_3: GCA1-2009 + GCA1-2010 + GCA1-2011, VS_1: GCA1-2012, VS_2: GCA1-2013, VS_3: GCA1-2014.

When relatedness between TS and VS increased, there was a general increase in ρ_{GP}. The increment depends on the dataset, the target VS and the model (particularly for the PL dataset). For example, in the VS_1 of the GER dataset, the increase in ρ_{GP} from the 0P- to the 1P-scenario was from ~ 0.30 to ~ 0.60, and in the pooled dataset (GER&PL) from ~ 0.40 to ~ 0.50, whereas in the PL dataset the 1P-scenario had too wide confidence intervals and varying predictive abilities across models, so that no general trend can be recognized. For VS_3, there was no increase in ρ_{GP} from the 0P- to the 1P-scenario. This is in agreement with the Euclidean distances presented in Table C.11.

Predictive abilities were on average higher for the GER dataset (0.2741) than for the GER&PL program (0.2407) and markedly higher than for the PL dataset (0.1424). When splitting German and Polish genotypes within the GER&PL dataset, ρ_{GP} for only Polish lines was lower than the ρ_{GP} obtained when only considering the PL program, whereas the ρ_{GP} obtained for German lines within the GER&PL dataset was higher than that obtained from the GER dataset alone. The principal component analysis (PCA) of the marker data in Figure 4.5 shows that the genotypes from the PL program form a more compact cloud than those from the GER program and that the Polish lines are well contained within the cloud of the German lines. Although the first two principal components capture little variance ($< 15\%$), the PCA shows that lines in the PL program are more closely related than lines in the GER program, so that some far related German lines could cause a bias in the prediction of the Polish lines within the GER&PL dataset.

For controlTS_2 and controlTS_3, approach A1K (year-wise with kinship) was on average 17% higher in predictive ability than A1 (year-wise without kinship) across programs, relatedness scenarios, TS and VS (17.3 % in the GER dataset, 21.8% in the PL dataset and 13.0% in the GER&PL dataset). Approaches A2 (2-stg-Kin), A3 (2-stg-Kin-het) and A4 (3-stg-NoKin) yielded very similar predictive abilities across datasets, relatedness scenarios and VS for TS_2 and TS_3 (on average 0.2497), and were also very close to predictive abilities obtained by A1K (on average 0.2477). The worst approach was A5 (3-stg-Kin), which led on average to 23% lower ρ_{GP} than the average of A2, A3 and A4 across programs, relatedness scenarios and VS.

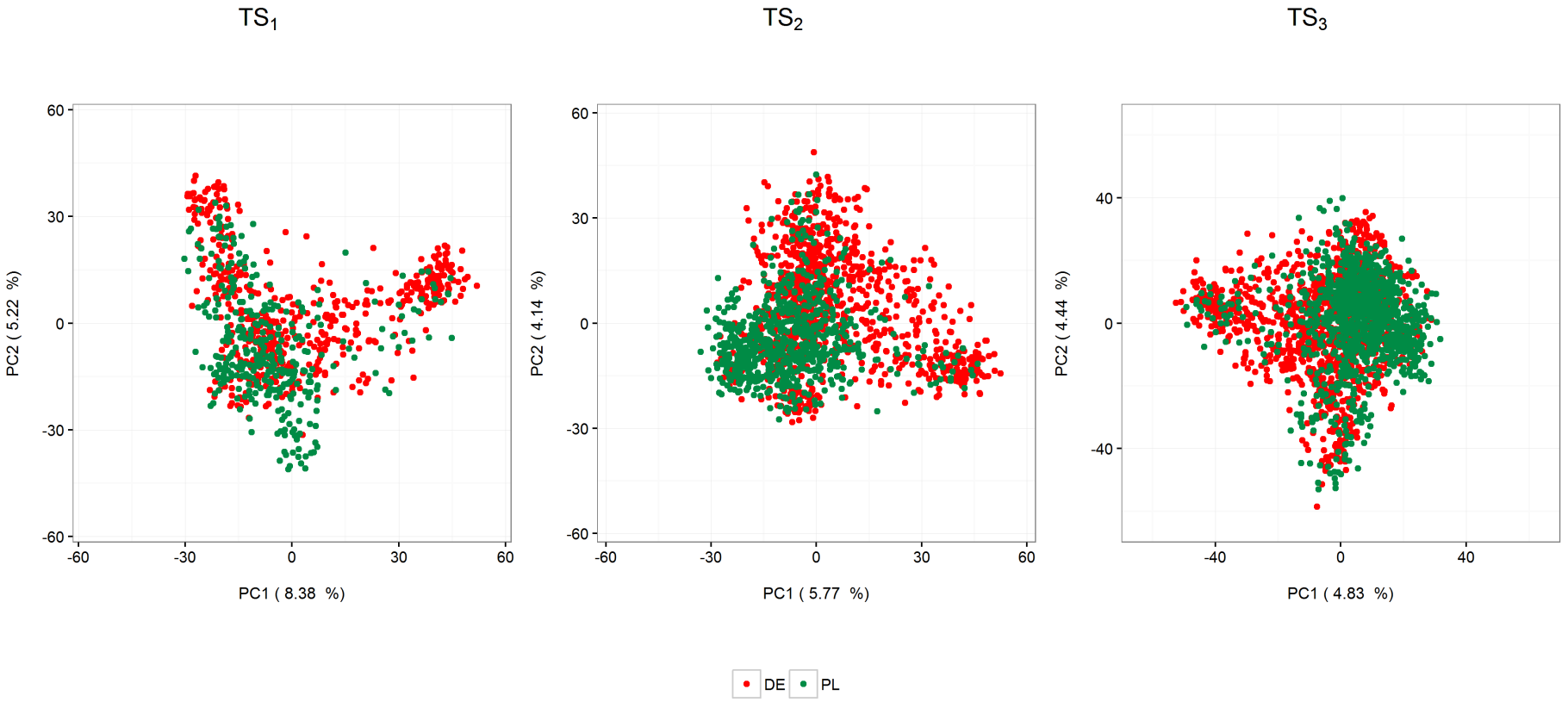


Figure 4.5: **Principal component (PC) plots for the training datasets TS_1, TS_2 and TS_3 of the German (GER) and the Polish (PL) programs.** TS_1: GCA1-2009 + GCA2-2010 + GCA3-2011, TS_2: GCA1-2009 + GCA2-2010 + GCA1-2010 + GCA2-2011, TS_3: GCA1-2009 + GCA2-2010 + GCA3-2011 + GCA1-2010 + GCA2-2011 + GCA3-2012 + GCA1-2011 + GCA2-2012 + GCA3-2013

Predictive abilities in sampling scenarios

To avoid the confounding effect of the VS-size and to objectively compare parent scenarios and models, we defined a VS-size of 100 genotypes to be predicted and iterated the GP-FV 10 times. Results are depicted in Figures C.5 - C.7. The predictive abilities and their 95% confidence intervals are based on the mean of the 10 sample draws.

The predictive abilities obtained for the scenarios with samples of 100 genotypes in the VS confirmed the trends observed for scenarios with complete validation sets (Figures 4.2-4.4). The size of the confidence intervals varied between the sampling scenarios and the scenarios using all available genotypes. For smaller VS-size (sampling 100 genotypes), confidence intervals are wider, suggesting that more and better data would allow better genotype estimates, as expected.

4.4.3 Relatedness scenarios

A PCA for each combination TS-VS-relatedness scenario in all the datasets (GER&PL, GER and PL) showed that PC1 and PC2 captured only little variance ($< 15\%$) (Figures C.8-C.16), but still showed that TS and VS are genetically structured and there is no clear separation for TS and VS using different relatedness degrees, i.e., different parent number in the TS.

Additionally, the mean of the Euclidean distance using the marker matrix for genotypes in TS and all relatedness scenarios of VS (Table C.11), showed no strong variation between relatedness scenarios and between TS-VS combinations. The values were in general slightly higher for the PL dataset than for the GER dataset, showing that the two groups are closely related within themselves but marginally genetically divergent between them. The results are consistent with the PCAs, since there was no clear pattern from the 1P-scenarios that would suggest a closer relatedness between TS and VS than the 0P-scenarios or the All-scenarios.

For the three relatedness scenarios (All, 0P- and 1P-scenarios) across all the datasets (GER, PL and GER&PL), approach A1K (year-wise with kinship) produced, in general, very similar predictive abilities than approaches A2 (2-stg-Kin), A3 (2-stg-Kin-het) and A4 (3-stg-NoKin), and these four approaches were on average 18% better than approaches A1 (year-wise without kinship) and A5 (3-stg-Kin) in terms of ρ_{GP}. In the GER and GER&PL datasets, A1K produced slightly higher predictive abilities than A2, A3 and A4 for All- and 0P-scenarios, whereas for 1P-scenario there was no markedly difference between A1K and A2, A3 and A4. In the PL program, A4 had on average 13% and 8% higher ρ_{GP} than A1K for the 0P- and 1P-scenario, respectively. For the All-scenario, A4 showed no difference with A1K and both approaches yielded on average 14% better ρ_{GP} than A2 and A3.

4.4.4 Top-yield scenarios

In the present study, using a selected fraction of individuals in the TS was useful only in the control TS, i.e., when a given selection cycle (genetic background) was represented by only one year of (GCA1) data (Figures 4.6 - 4.8). In this case, the effects of non-yield QTL are confounded within each genetic background with the GY effects. Consequently, a selected fraction of individuals with higher grain yield performance will reduce variation due to non-yield QTL and, therefore, reduce bias due to confounding effects. In contrast, when two or more years of data are available per genetic background, environmental and non-yield QTL effects can be estimated separately, thus rendering the use of selected fractions in the TS (Top75% or Top50%) non-effective.

For the control TS across all datasets, the Top75% and Top100% scenarios using the year-wise (A1) approach and year-wise with kinship (A1K) approach had a higher ρ_{GP} than the Top50% scenario. For the GER and GER&PL datasets A1K using Top75% was marginally better than A1K using Top100% (on average 4% better) and across all datasets, A1K had 13% higher ρ_{GP} than A1. Additionally, for A2 (2-stg-Kin), A3 (2-stg-Kin-het), A4 (3-stg-NoKin) and A5 (3-stg-Kin) the Top100% scenarios outperformed the Top75% and Top50% scenarios in terms of ρ_{GP}.

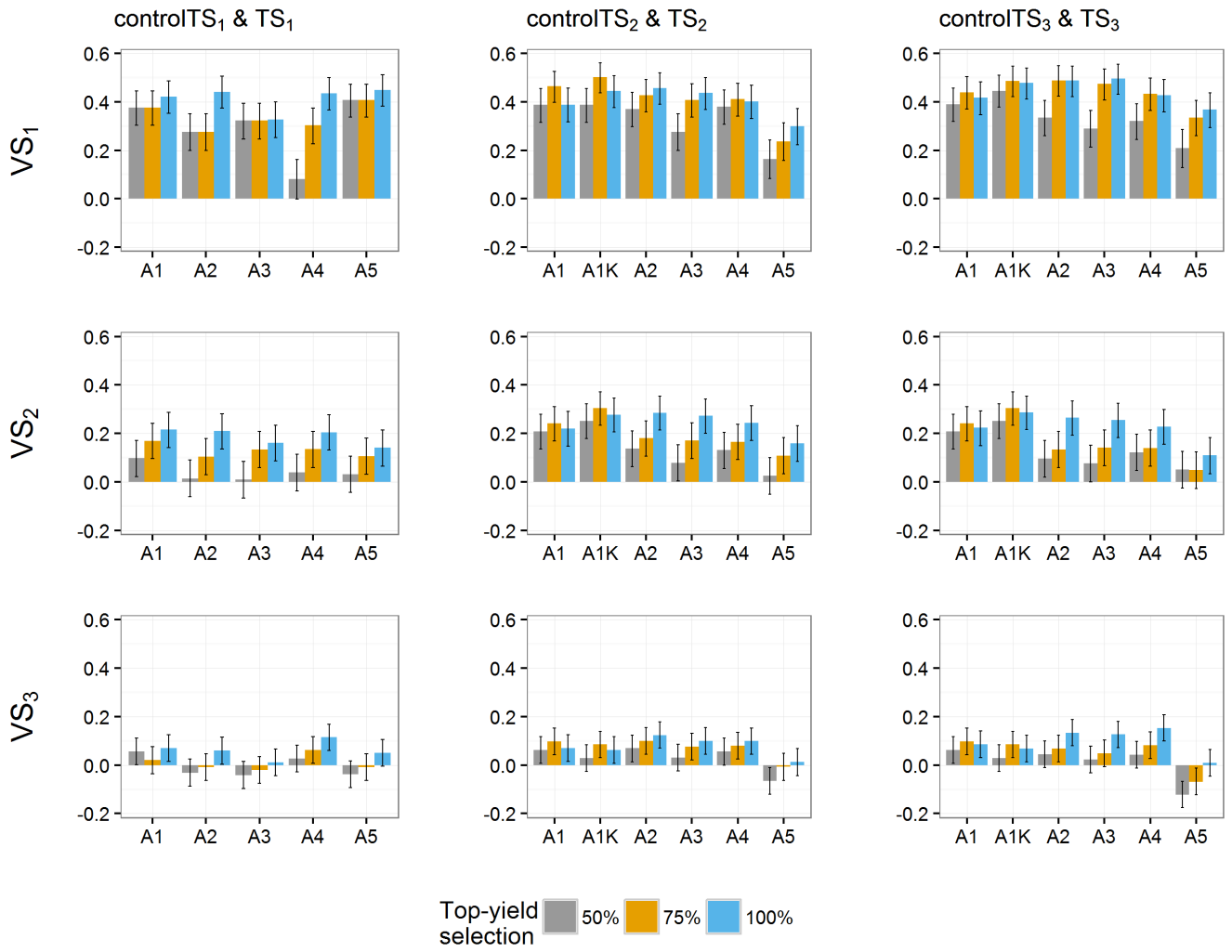


Figure 4.6: **Predictive abilities (y-axis) of the German and Polish dataset for selection scenarios of top-yield performance.** Selection in the training set (TS): 50% of highest yielding genotypes (gray bars), 75% of highest yielding genotypes (yellow bars) and 100% of the genotypes (blue bars), using validation sets VS_1, VS_2 and VS_3. Black lines for each bar represent the 95% confidence intervals of the predictive ability. Year-wise approach (A1) and year-wise with kinship approach (A1K) were fitted to the control TS, approaches 2-stg-Kin (A2), 2-stg-Kin-het (A3), 3-stg-NoKin (A4) and 3-stg-Kin (A5) to the complete TS.

TS_1: GCA1-2009 + GCA2-2010 + GCA3-2011, controlTS_1: GCA1-2009, TS_2: GCA1-2009 + GCA2-2010 + GCA1-2010 + GCA2-2011, controlTS_2: GCA1-2009 + GCA1-2010, TS_3: GCA1-2009 + GCA2-2010 + GCA3-2011 + GCA1-2010 + GCA2-2011 + GCA3-2012 + GCA1-2011 + GCA2-2012 + GCA3-2013, controlTS_3: GCA1-2009 + GCA1-2010 + GCA1-2011, VS_1: GCA1-2012, VS_2: GCA1-2013, VS_3: GCA1-2014.

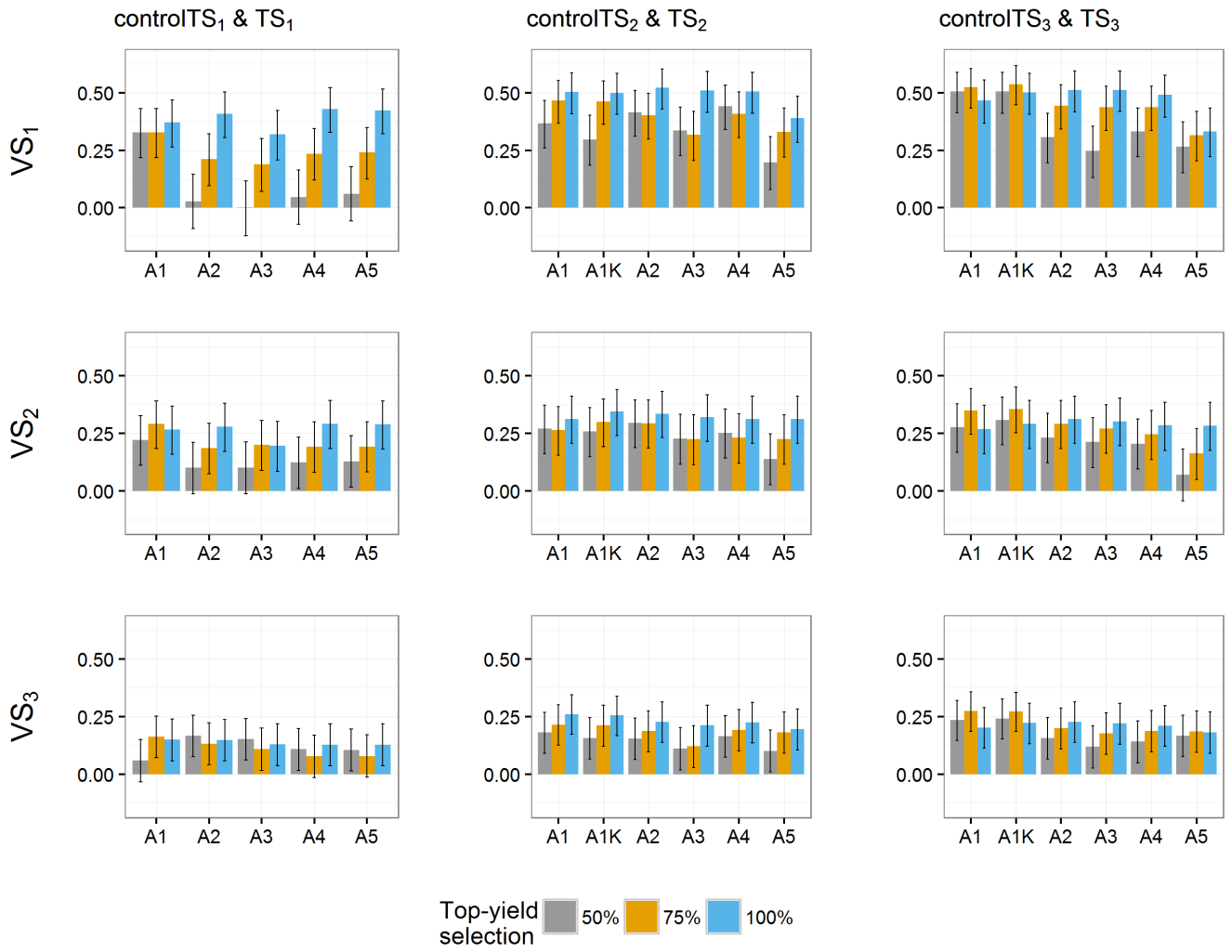


Figure 4.7: **Predictive abilities (y-axis) of the German dataset for selection scenarios of top-yield performance.** Selection in the training set (TS): 50% of highest yielding genotypes (gray bars), 75% of highest yielding genotypes (yellow bars) and 100% of the genotypes (blue bars), using validation sets VS_1, VS_2 and VS_3. Black lines for each bar represent the 95% confidence intervals of the predictive ability. Year-wise approach (A1) and year-wise with kinship approach (A1K) were fitted to the control TS, approaches 2-stg-Kin (A2), 2-stg-Kin-het (A3), 3-stg-NoKin (A4) and 3-stg-Kin (A5) to the complete TS. TS_1: GCA1-2009 + GCA2-2010 + GCA3-2011, $controlTS_1$: GCA1-2009, TS_2: GCA1-2009 + GCA2-2010 + GCA1-2010 + GCA2-2011, $controlTS_2$: GCA1-2009 + GCA1-2010, TS_3: GCA1-2009 + GCA2-2010 + GCA3-2011 + GCA1-2010 + GCA2-2011 + GCA3-2012 + GCA1-2011 + GCA2-2012 + GCA3-2013, $controlTS_3$: GCA1-2009 + GCA1-2010 + GCA1-2011, VS_1: GCA1-2012, VS_2: GCA1-2013, VS_3: GCA1-2014.

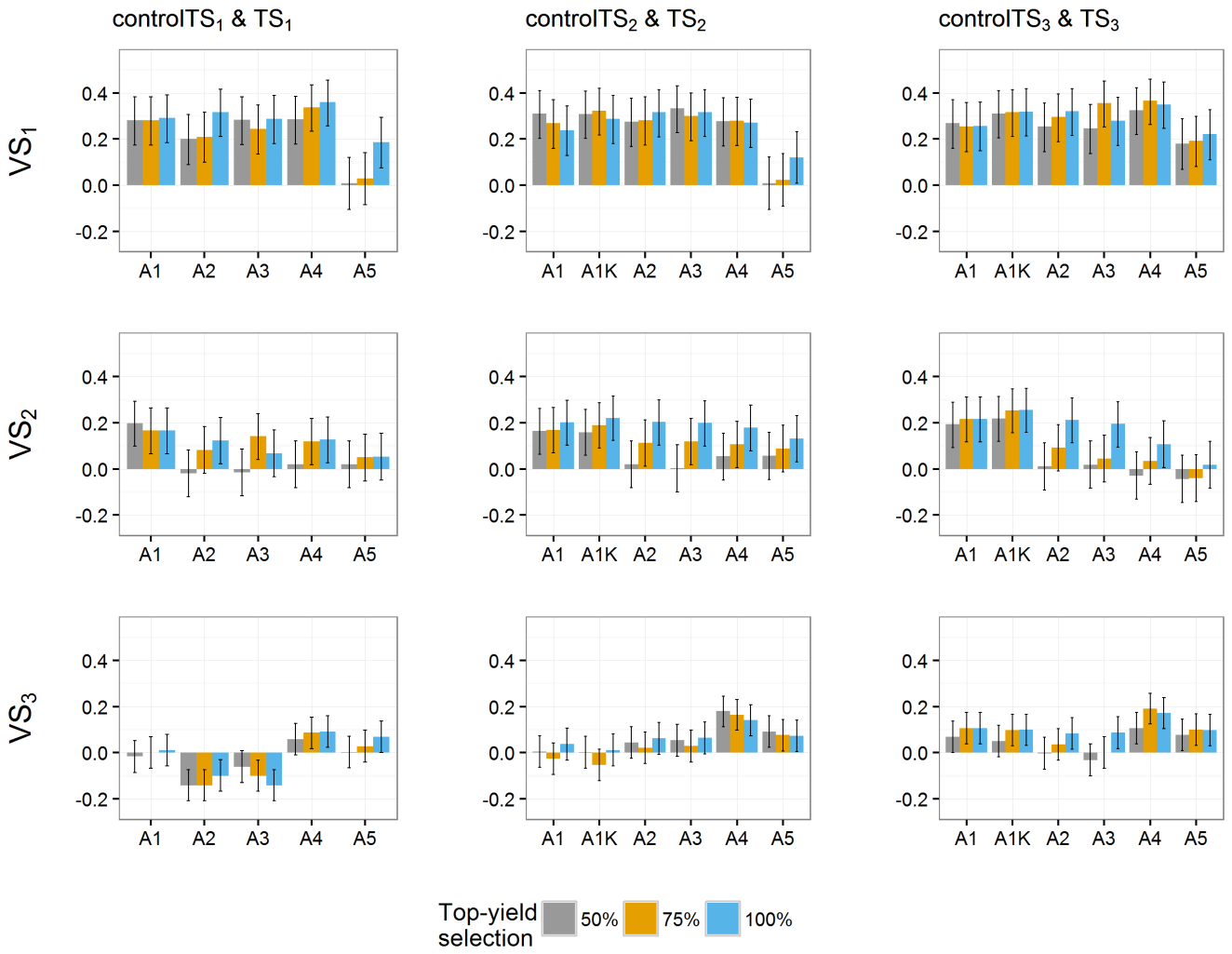


Figure 4.8: **Predictive abilities (y-axis) of the Polish dataset for selection scenarios of top-yield performance.** Selection in the training set (TS): 50% of highest yielding genotypes (gray bars), 75% of highest yielding genotypes (yellow bars) and 100% of the genotypes (blue bars), using validation sets VS_1, VS_2 and VS_3. Black lines for each bar represent the 95% confidence intervals of the predictive ability. Year-wise approach (A1) and year-wise with kinship approach (A1K) were fitted to the control TS, approaches 2-stg-Kin (A2), 2-stg-Kin-het (A3), 3-stg-NoKin (A4) and 3-stg-Kin (A5) to the complete TS. TS_1: GCA1-2009 + GCA2-2010 + GCA3-2011, controlTS_1: GCA1-2009, TS_2: GCA1-2009 + GCA2-2010 + GCA1-2010 + GCA2-2011, controlTS_2: GCA1-2009 + GCA1-2010, TS_3: GCA1-2009 + GCA2-2010 + GCA3-2011 + GCA1-2010 + GCA2-2011 + GCA3-2012 + GCA1-2011 + GCA2-2012 + GCA3-2013, controlTS_3: GCA1-2009 + GCA1-2010 + GCA1-2011, VS_1: GCA1-2012, VS_2: GCA1-2013, VS_3: GCA1-2014.

4.5 Discussion

The key contribution of this paper was an evaluation of the use of kinship to model GY effects in disconnected datasets for a better separation from GEBV. We presented a detailed step-by-step genomic prediction analysis modeling GY with different approaches and extending the use of molecular markers to deal with disconnected trials. We also use a validation set system across years that approximates to the breeders' aim of empirical validation.

In the analyzed datasets, we found that G and GY effects (and other effects that include factor G) were confounded. This was evident from the large negative asymptotic correlations that reflect ill-conditioning (Tables 4.2, C.2-C.4). Using multiple genetic backgrounds as in TS_2 (two selection cycles) and TS_3 (three selection cycles), it is possible in principle to build bridges across years given that GY is specific to the genetic background. Nonetheless, the unbalancedness of the design was still so strong that those effects remained confounded (Tables C.5-C.10). The use of several cycles improved the estimate of the variance of genotype effects because there were more lines repeated across years within cycles (especially in the PL dataset), thus solving the problem of a zero variance estimate with single-cycle data. By contrast, the use of multiple cycles did not solve the ill-conditioning problem.

The main advantage expected from pooling GCA1+GCA2+GCA3 data in the TS is that a better bridge is built between years, leading to more precise adjusted means, thus allowing to dissect GY from GEBV. If most of the interaction is specific to the genetic background (as we assume it to be), multiple genetic backgrounds (selection cycles) are needed for a better separation of main SNP effects, such as in TS_2 and TS_3. Auinger et al. [2016] recently found that aggregating data from several consecutive cycles connected by a sufficient number of common ancestors improves the accuracy of the predictions of candidate genotypes. Our results confirm their conclusion and complement the recommendation towards using additionally a selected fraction of 75% best yielding genotypes in the TS to reduce biasing effects due to non-yield QTL. The most surprising result is that the highest and most stable results are obtained with the controlTS_2 and controlTS_3 with A1K, i.e., using GCA2 and GCA3 data apparently is not only advantageous, but leads to a slight reduction in prediction abilities in comparison to using multiple consecutive GCA1 data, as in A1K. This is probably due to a biased segregation and variation of QTL effects in the selected fractions of GCA2 and GCA3 with respect to the non-selected GCA1 datasets.

The advantage of using a whole cycle with GCA1 to GCA3 is that the genotypes making it to GCA2 and GCA3 have been seen in more than one year, thus models that use a complete TS benefit from the TS structure, allowing reasonable GY estimates with or without kinship. By using only GCA1 experiments (i.e., control TS), a good coverage of the genetic target population is achieved and the use of kinship to model the genetic connection across years (specifically with model A1K) seems to be powerful enough

to estimate GY fairly independent from GEBVs.

The PL dataset produced markedly lower predictive abilities than the GER and the GER&PL datasets. We had stated that the German genotypes profited from the Polish ones but not *vice versa*, perhaps because the GER program is genetically more diverse than the PL program (Figure 4.5), so that there are some SNPs that are monomorphic for the Polish lines but not for the German lines causing a bias in the prediction of the Polish lines within the GER&PL dataset. Probably the main reason why the PL dataset had markedly lower predictive abilities than the GER dataset is that the Polish data have a higher error, i.e, GY, GL and GYL interaction effects are estimated less accurately. The fact that in Poland there are fewer GL and GYL evaluations (Table C.1) could explain why the Polish predictive abilities were lower. Endelman et al. [2014] show that having larger populations spread across more environments produces higher predictive abilities than evaluating the same genotypes in fewer environments. The GER dataset has a higher number of GL and GYL combinations because trials with Tester 1 and Tester 2 are not evaluated in exactly the same locations, whereas in the PL dataset, there is a balanced design of testers across locations within a year.

Predictive abilities were in general 26% higher for the 1P-scenarios than the 0P-scenarios and 15% higher than for the All-scenarios, reinforcing the findings of other genomic prediction studies on the effect of relationships between TS and VS [Albrecht et al., 2011; Brøndum et al., 2011; Daetwyler et al., 2013; Habier et al., 2007; Pszczola et al., 2012]. The use of the kinship to model GY in 0P-scenarios did not consistently compensate the lack of relatedness. Although the three relatedness scenarios (All-, 0P- and 1P-scenarios) showed small differentiation by the mean Euclidean distance (Table C.11) and not so marked divergence in the PCA plots (Figures C.8-C.16), a realized relationship between TS and VS does have a positive impact on the predictive abilities. In the best case, i.e. the GER dataset - VS_1:GCA1-2012, predictive abilities ranged from ~ 0.14 to ~ 0.38 in the 0P-scenario and from ~ 0.50 to ~ 0.73 in the 1P-scenario.

All approaches revealed marked variation in predictive abilities across scenarios. In general, there was a modest increment of the year-wise with kinship approach (A1K) over the year-wise approach (A1), in particular controlTS_2:GCA1-2009 + GCA1-2010 and controlTS_3:GCA1-2009 + GCA1-2010 + GCA1-2011 over controlTS_1:GCA1-2009. The confidence intervals of the predictive abilities of the year-wise approach (A1) overlapped most of the times with predictive abilities of the year-wise with kinship approach (A1K) (black lines of Figures 4.2-4.4), but even so, in challenging programs as the Polish one, the benefit of using the kinship was worth about 22% on the correlation scale. In the GER and GER&PL datasets the approaches A2, A3 and A4 had consistent and very similar predictive abilities. Only A5 was almost always markedly lower in predictive ability than the other models. From these

results we conclude first, that using the kinship to model GY for settings of disconnected years is safer than estimating the year effect as the simple average of the genotypes evaluated in a given year, and second, when the datasets cover multiple genetic backgrounds in the same year (as datasets used for A2, A3, A4 and A5), it is possible to estimate GY effects either by using kinship directly in the GP stage (A2 and A3) or simply using the correct model in the TS to obtain adjusted genotype means across years (A1K) and submit them to GP. Hence, kinship is helpful in the case of disconnected data and no harm is done using it in other cases. Although computational load may increase with the use of kinship to model GY, novel approaches that combine dense and sparse matrix methods alleviate this burden and are starting to become freely available [De Coninck et al., 2016].

It was surprising that the 3-stg-Kin approach (A5) had markedly lower predictive abilities than the approach 3-stg-NoKin (A4) because the difference between both approaches is that in A5, we use kinship to model the GY whereas in A4 we do not, so we would have expected that using kinship in modeling GY improves predictive ability. While this expectation was confirmed in the other approaches that used kinship (A2 and A3), this was not the case here. All methods are designed to approximate the same single-stage model (4.1), so that it was not obvious which one should work better because it uses kinship to model the GY effects, as does model (4.1). While both A5 and A1K seek to approximate the single-stage model (4.1), A1K makes somewhat weaker assumptions because it does not use kinship to model GY in the second stage. So while A5 better approximates the single-stage model, there is no guarantee that the single-stage model is the best model for GP. This may explain why A1K does better in terms of predictive accuracy and also why A4 fared better than A5.

Predictive abilities for VS_1:GCA1-2012 ranged from ~ 0.24 on average in the PL dataset to maximum ~ 0.73 in the GER dataset, and the lowest ρ_{GP} occurred for VS_3:GCA1-2014 ranging from zero (or negative) in the worst case of the PL dataset to ~ 0.33 the best case in the GER dataset. The results that we obtained are in accordance with the predictive abilities reported by Auinger et al. [2016], which ranged between 0.39 and 0.58 (with an average heritability of 0.83) and were based on GP-FV. The validation sets VS_1:GCA1-2012 and VS_2:GCA1-2013 could be predicted more accurately than VS_3:GCA1-2014. Data from the year 2014 has been identified as problematic, since it is not easy to predict within the GP program from KWS-LOCHOW. We acknowledge the fact that the scenarios TS_3-VS_1 and TS_3-VS_2 are less realistic in the sense that data from the same year of prediction is used in TS and VS, but we consider those scenarios because the number of genotypes in GCA3-2012 and GCA3-2013 is low (less than 30 shared genotypes within cycles in all the programs) and there are no genotypes in common between TS and VS, keeping our condition of disconnected TS and VS valid for the presented scenarios. Moreover, removing data from GCA3-2012 or GCA3-2013 from TS_3 led to only a slight variation in the value of the reported predictive abilities, with changes occurring after the

third or fourth decimal place.

Besides focusing in the mean performance across years, another important target in plant breeding is to investigate stability, which refers to the variability from year to year. In the context of genomic prediction, it makes sense to also study the expected consistency of year to year performance aiming to minimize this variability [Malosetti et al., 2016; Mühleisen et al., 2014]. This stability aspect deserves further study.

The results obtained for the top-yield scenarios led us to conclude that using a multiple genetic background in the TS allows capturing the true QTL for yield, whereas when having only one year in the TS (i.e. control sets), the model is not able to do this distinction and hence, a pre-selection of best yielding genotypes may improve the predictive abilities. This explains the ability of the year-wise with kinship approach (A1K) to improve ρ_{GP} using 75% of the best-yielding genotypes even if the TS-size was reduced. Selecting a top fraction of best yielding genotypes for the TS basically allows to reduce the genotype-by-year effects that cannot be accurately estimated due to absence of connectivity across years. In this work, we randomly used 75% top fraction, but other values (e.g 95%, 90%, 85%, 80%) should be further considered. The implementation of the A1K (year-wise with kinship) approach is advantageous from the technical point of view, since the analysis requires lower computing power than using 100% data from complete cycles as for A2, A3 and A4. Given our results, we recommend that for GP in a breeding program with a similar structure to the program described in the present work, the year-wise with kinship (A1K) approach with TS composed of minimum two single years of multiple genetic backgrounds (i.e. controlTS$_2$ and controlTS$_3$) should be used.

4.6 Conclusions

The main conclusions of this study are: (i) Using multi-year datasets is advantageous, (ii) the year-wise with kinship approach (A1K) with two or three years in the TS (controlTS$_2$ or controlTS$_3$) led to slightly better and more consistent ρ_{GP} than any other approach, (iii) the use of kinship to model GY in multi-year datasets is encouraged, especially for datasets covering multiple genetic backgrounds and where disconnected trials across years are evaluated, i.e. year-wise with kinship approach (A1K), (iv) if only data from one selection cycle is available (TS$_1$) there is a loss in ρ_{GP} with no options to improve via kinship or other models, (iv) predictive abilities improved in scenarios where TS and VS were more related (1P-scenario), and (v) pre-selection of top-yielding genotypes is recommended in cases where several single-year data are available within selection cycles and in such cases, the use of the kinship to model GY is also advisable.

Chapter 5

General discussion

The three main results of this work pertain to data pre-processing, phenotypic analysis and genomic prediction as key components of genomic selection programs. The results can be summarized as follows: (i) an outlier detection step before phenotypic and genotypic analyses is justified and can only improve predictive performance in GP analysis, (ii) for the studied field designs, it was enough to model the blocking factors (incomplete block, row and column) to control for the field heterogeneity, whereas spatial modeling provided little improvement in accuracy, and (iii) using the kinship to dissect genotype and GY effects for multi-year data was worthwhile, in particular for TS with disconnected years.

The broader implications of these results are discussed in this chapter and are aligned with current literature on genomic prediction. The three main thematic areas are covered and perspectives for future research are included as closing remarks.

5.1 An evaluation of outlier detection methods

It was shown that the outlier identification methods helped scanning datasets for spurious observations, especially when datasets are large and fast-return-decision is needed (Chapter 2). The decision of what to do after flagging an observation as an outlier is crucial and needs to be made based on sufficient knowledge of the field trials and problems that occurred during their conduct. Identifying outliers may lead to improving the residual plots, hence ensuring that assumptions of normality and homogeneity of error variance (within groups) are fulfilled. These assumptions are essential for LMM analyses, which have been the cornerstone of the models considered in this work. In practice, in later stages of a breeding program, one can afford removing some suspicious observations, whereas in screening trials or other kinds of experiments in different areas of plant science, where data size is usually small and there are no or few field replications, the decision about keeping, replacing or removing an observation may have a

stronger impact; e.g., a breeder may inadvertently discard a promising variety.

In Chapter 2, it was shown that removing outliers in empirical datasets used for GP did not reduce the predictive abilities relative to the use of all available observations. The datasets scrutinized were relatively large (i.e. $N \simeq 25,000$), so one could argue that removing around 100 observations may not have a strong impact on the predictive ability, which was indeed the result of the comparison of different outlier detection methods. The methods varied only by around 0.01 units on the correlation scale. Using less than half of the data (i.e. $N \simeq 7,000$) led to predictive ability differences between outlier detection methods of around 0.05 units on the correlation scale, hence, raising the question regarding the usefulness of flagging, removing or down-weighting outliers when datasets for GP are large (e.g. $N \geq 20,000$), middle-sized (e.g. $2,000 < N < 10,000$) or relatively small (e.g. $N < 2,000$). Estaghvirou et al. [2014] concluded that effects of outliers on the predictive accuracies of genomic selection are greater for small than for large datasets, which confirms the findings in this work. Nonetheless, the methods implemented in Chapter 2 were all performed at the level of the trial, that is, outliers were identified based on local conditions and not global trends, thus detection of outliers was independent of the GP dataset size. One could argue here that other factors such as TS-size are more likely to be responsible of the small or large differences across predictive abilities when comparing outlier detection methods. Independent of the size of the dataset used for GP, identifying outliers at the trial level ensures that adjusted means or genotypes are not severely biased due to outliers increasing variability of records within a group [Yang et al., 2004].

Even though methods of outlier detection seem to be fallible, their use is always suggested as a first step in any analysis procedure [Besag and Kempton, 1986; Schützenmeister et al., 2012; Schützenmeister and Piepho, 2012]. In this work, a variety of methods are proposed to meet this need (Chapter 2). One of the crucial observations on the analysis of the outlier identification methods was that using the robust scale estimate median absolute deviation (MAD) to standardize residuals, which is also implemented in the PlabStat package, showed good properties also when used in a REML framework, reflected in that the predictive abilities obtained later in GP were slightly higher for the methods that considered a robust scale estimate (i.e. M4:BH-MADR , M5:BH-STRO and M1:PlabStat). In general, robust methods in the context of linear models have shown to be valuable tools as they ensure that the analysis is not significantly disturbed by any outlying observation [Lourenço and Pires, 2014].

An additional concern regarding GP using multi-year and/or multi-location data is that this kind of data is *per se* environmentally-driven heterogeneous, which has an impact on prediction of GEBVs [Ou et al., 2015]. This situation has led to other outlier concepts in GP, such as outlier environments, outlier trials or outlier cycles and also to propose GP models that account for heterogeneous residual

variance. Heslot et al. [2013a], for example, used environmental clustering based on Euclidean distances and on predictive accuracies between pairs of environments to identify possible outlier environments. They found that removing environments with a low ability to predict other environments led to higher predictive accuracies. Another example is the identification of outlier breeding cycles. In a GP study across multiple breeding cycles in wheat breeding, Michel et al. [2016] evaluated possibilities to increase the prediction ability. They found that in a between-cycles cross validation scheme, dropping outlier cycles raised predictive abilities for grain yield and protein content of wheat. These two approaches of removing complete environments or breeding cycles from the TS can be very controversial. First, a major conceptual problem can be stressed using a very simple example: if for any dataset on (y, x) paired data a regression is fitted, and then sequentially, residuals are computed and the most outlying (y, x) pair is removed, the fit of the regression will improve till the fit is perfect $R^2 = 1$ for a final dataset comprising two (x, y) pairs. And second, eliminating a complete outlying environment or cycle may improve predictive ability in a given set of environments but may lead to a loss of predictive power with respect to the target population of environment (TPE). The concept of TPE designates the expected mixture of environmental conditions likely to occur across multiple years in a given geographical area [Comstock, 1977]. The breeder creates and selects improved genotypes within this geographical and temporal sets of environments. The authors justify the elimination of outlier environments based on the assumption that the initial dataset is an approximate sample of the TPE, so that environments, which on average have a poor prediction accuracy of other environments, are likely to be less frequent in the TPE. Thus, they see their method as a weighting approach, where some environments receive zero weight. It is crucial to conduct METs for a representative sample of environments from the TPE [Piepho and Möhring, 2005; van Eeuwijk et al., 2016], so if there is a random sample of environments to begin with, dropping environments based on outlier detection may be problematic, because the remaining set of environments may become biased with respect to the TPE. An approach detecting outliers per trial at the level of plots (Chapter 2) is much more conservative, because no entire environment will be removed. One should also keep in mind that the precision of a genotypic value estimate based on MET data primarily depends on the number of environments [Chapman et al., 2000; Endelman et al., 2014]. With efficient weighting methods in place [Damesa et al., 2017; Möhring and Piepho, 2009; Piepho et al., 2012a; Smith et al., 2001], there is no good reason to drop an entire environment.

A different approach was used by Ou et al. [2015], who evaluated simulated data with different degrees of heterogeneous residual variance. They showed that models accounting for heterogeneous residual variance improved predictive accuracies in scenarios with high degree of residual heterogeneity. Thus, since METs are environmentally heterogeneous *per se*, the use of those models in combination with

outlier robustness extensions at the plot level is recommended. This approach ties in well with the stage-wise analyses used in this thesis for genomic prediction procedures. Other studies in the plant breeding context have also highlighted the advantages of stage-wise approaches [Damesa et al., 2017; Piepho et al., 2012a; Schulz-Streeck et al., 2013b]. Foremost, stage-wise analysis allows fitting specificities innate to a given trial. If a trial has low coefficient of variation, the weighting scheme used to pass on from one stage to another will have the ability to account for the precision of the trial, giving little weight to those trials where variation was high. This suggests that breeders should not discard entire trials just because precision is low. It is always worthwhile to include such trials (unless they show extreme atypical ranking of genotypes compared to the TPE), because any trial in a different environment provides valuable information on the genotype-by-environment interaction in the TPE.

5.2 Merit of spatial modeling

Phenotypic analysis has not been a frequent target of attention for GP, but dedicating some effort to selecting an appropriate phenotypic model is justified [Schulz-Streeck et al., 2012, 2013b, Chapter 3, Chapter 4]. This thesis evaluated whether using spatial or non-spatial models has a positive effect on the prediction performance in the GP stage (Chapter 3). Then, it was investigated if the use of genotypic data could also help to model effects that usually are handled only at the phenotypic analysis stage and lead to higher predictive abilities (Chapter 4). In both cases, the choice of an appropriate model for the phenotypic data was crucial. Conversely, when model choice is sub-optimal, confounding effects, over-parametrization or low-quality data can lead to biased estimates and singularity problems [Besag and Kempton, 1986; Pinheiro and Bates, 2000].

Piepho and Williams [2010] proposed extensions of the common experimental designs used to randomize field trials from plant breeding and variety testing (e.g., resolvable block or row-column designs) towards an improvement by post-blocking or addition of spatial model components. They argue that a proper randomization accounting for blocking factors is an efficient strategy, hence spatial modeling should be regarded only as an add-on component of models with a block structure. In Chapter 3, spatial models were used in the phenotypic analysis in comparison to a baseline model (the basic randomization-based model for an α-design) and a baseline model plus post-blocking for rows and columns. As row and column coordinates were available from the trial layout of the rye METs, the addition of these factors during the data analysis to account for other source of variation in the field was straightforward: the initial blocking factors were kept (i.e. replicates and blocks within replicates), and post-blocking factors were added (i.e. rows within replicates and columns within replicates). The extension to fitting a spatial trend can be implemented in the directions of the post-blocking factors.

With the empirical data available for the present study, it turned out that the spatial models did not markedly improve the predictive abilities but the bidirectional post-bocking did yield better prediction performance. This result implies that under the climatic and pedological conditions of the trial locations studied in this thesis attention paid to the blocking settings on the field can be enough to control field heterogeneity, rendering row-column designs as an appealing trial choice to consider. Lado et al. [2013] in the opposite, found that accounting for spatial variation was worthwhile for GP. They used data from three different water regimes in Mediterranean conditions. It seems that in more stress-prone environments, where water may be limiting or heat excessive, a larger gain in precision is found by using spatial models [Leiser et al., 2012] than for example in Germany, where rain and fertilizer levels may not be stress factors.

5.2.1 The model selection controversy

In Chapter 3, it is shown that GP can also be used for selection of phenotypic models. A step-by-step description of the model construction and the way to select a proper model comparing predictive abilities obtained from GP-CV vs. AIC-based selection was depicted. These two approaches selected the same models but showed a different pattern in selection across models, and either any approach selected the model that in overall had the highest predictive ability. An advantage of AIC is that it does not require more calculations after a model has been fitted, but, once a GP-CV pipeline has been established, the GP-CV model selection approach does not need too much computation time. A question here is why sometimes AIC selects a different model than GP-CV. A better understanding of what the model selection aim is and what the validation methods' purposes are would help clarifying the question. The goal of model selection can be estimation or identification; estimation when the goal is to minimize the loss between the "true model" and an approximating model, and identification when it is aimed to identify the "true model". AIC is typically built for estimation, whereas CV has been used for both estimation and identification [Arlot and Celisse, 2010]. AIC and other likelihood-based methods may allow selecting the best model within a set of candidate models hoped to well approximate the true model, but if there are no good models, they cannot be identified by model selection algorithms. Thus, uncertainty plays always a role when selecting a particular model, first, because it is unknown if a good approximating model is in the candidate set, and second, because even if there is a good approximating model in the candidate set, it is unsure that the model selection algorithm identifies the best approximation thereof [Burnham and Anderson, 1998]. Different types of CV display intrinsic properties in terms of bias and variance that may be the cause of the different results between CV and AIC model selection. The TS and VS sizes affect prediction accuracy in GP. Large TS and small VS may lead to high accuracies and higher variances [Erbe et al., 2010]. Thus, the choice of number of folds for TS and VS becomes

important. Other model selection criteria proposed in the literature could be potential alternatives to explore model selection. For example, conditional AIC (cAIC) [Hurvitch and Tsai, 1989], unbiased AIC [McQuarrie et al., 1997] or generalized CV criterion [Ansley and Kohn, 1987]. More recently, Vaida and Blanchard [2005] proposed a cAIC for mixed models considering the degrees of freedom accounting for shrinkage in the random effects. This criterion is more appropriate when the focus is on "clusters" instead of population parameters, e.g. when the interest is to know the effect of the genotype at a specific environment. The cAIC, however, has some limitations as it ignores the uncertainty in the estimation of the covariance matrix of random effects, i.e. it uses a correction term for the bias that assumes known variance. Greven and Kneib [2010] proposed a corrected cAIC that accounts for the estimation of the variance parameters and avoids the high computational cost of other methods also addressing this problem. Likewise, measuring the uncertainty associated with a selected model (e.g., using bootstrap techniques) may add suitable information to the model selection decision [Burnham and Anderson, 1998].

5.2.2 Spatial adjustment in the genomic prediction model

Geostatistical models using marker covariates as "spatial" coordinates also offer a convenient alternative to model the variance-covariance matrix for the markers in the GP stage. Piepho [2009b] and Schulz-Streeck and Piepho [2010] used different spatial variance-covariance structures as function of genomic distances between pairs of genotypes in GP. In both studies, it was found that RR-BLUP, which is equivalent to the quadratic model, was slightly better than the other spatial models (i.e., linear, exponential, spherical, Gaussian, power). Piepho [2009b] demonstrates that accounting for epistatic effects in the GBLUP model is equivalent to fit an additional effect that assumes a Gaussian spatial variance-covariance for the additive $\times$ additive epistatic kinship. Likewise, Jiang and Reif [2015] used a Taylor argument to motivate the equivalence of the extended GBLUP (EG-BLUP) model and Reproducing Kernel Hilbert Space (RKHS) method, both accounting for epistatic effects, and Ober et al. [2011] applied the geostatistical concept of kriging to consider additive, additive-dominance, and epistatic gene-action models.

The use of spatial models is hence not restricted to phenotypic analysis to account for geographical variation, but it is also easily extended to model marker effects in GP. The LMM framework allows this extensions for a high-dimensional space. Consequently, the models described throughout this work are malleable material to further account for additional gene effects, whose covariance structure can be modeled using nonlinear functions of the SNP covariates.

5.3 Importance of modeling genotype-by-year *(GY)* interaction effects

The model choice at the phenotypic analysis stage is decisive to ensure precise genotype estimates and, further, accurate GEBVs at the GP stage. For example, the control datasets used in Chapter 4 were disconnected across years (i.e., there were no common genotypes or check entries across years). A first approach to analyze these data was to perform a phenotypic analysis by year and then submit the adjusted genotype means computed by year to a GP stage, where the simple mean of genotypes per year could be used as the year effect estimate (Chapter 3). In the case of disconnected or weakly connected data across years, this is a promising approach and ensures a more accurate year estimate than using a few checks to estimate year effects (Chapter 3). In this approach, however, it cannot be ensured that GEBVs are not confounded with GY effects. Besides, the annual gain from selection ($\sim 1\% - 2\%$) across years in the program is disregarded. To overcome these limitations, another approach including multi-year and multiple genetic backgrounds was studied (Chapter 4). The result was that using kinship to model GY helps to dissect GEBVs from GY effects. There are several studies that support this finding. On the one hand, doing the analysis by year seems to make sense because year-to-year variation is known to be very strong [Lado et al., 2016; Laidig et al., 2008; Piepho et al., 2014], but on another hand, if environments are understood as location-year combinations, fitting GE effects, at any level in the analysis, would absorb great part of the GY variation [Heslot et al., 2014; Jarquín et al., 2014; Malosetti et al., 2016, 2013].

It is argued that with the available dataset from KWS-Lochow in Chapter 4, splitting the environmental effect into year and location effects is crucial since: (i) year-to-year variation contributes to a large proportion of GE [Heslot et al., 2014; Laidig et al., 2008] and (ii) several breeding cycles are run in the same year, so that genetic effects can be separated from the GY effects. It is shown that either using several genetic backgrounds in the TS or kinship to estimate GY have a positive effect on the predictive ability (Chapter 4). Extensions of this model could be explored, e.g., using a covariate matrix of environmental information (as in Heslot et al. [2014]; Malosetti et al. [2016]) or also testing other variance-covariance structures such as factor analytical [Malosetti et al., 2013; Piepho, 1997, 1998] for GY or genotype-by-location interaction effects.

The advantages of the models used all throughout this work are that the models can be easily implemented using existing statistical software (e.g. SAS, ASReml, R) and can be extended to account for genetic effects not captured by markers, to accommodate location-specific variances or to assume that each marker has its own variance. Although the more extensions are added to the models, the more complex they become and the more computing power is required, novel computing tools are closing the gap between big-data and complex model analysis. Breeding companies producing thousands

of data records across years will be able to benefit from all those developments. Free R packages developed for GP, e.g., synbreed [Wimmer et al., 2012], BGLR [Perez and de los Campos, 2014], RR-BLUP [Endelman, 2011] and Needles [De Coninck et al., 2016]), are becoming popular and easy-to-use functions will start to be spread.

5.4 Genomic prediction: validation and implementation

Genotype adjusted means estimated from stage-wise analyses are the core unit of phenotypic information used for GP and also for CV. As adjusted means usually come from unbalanced designs, they are not independent and thus the basic assumption of independent and identically distributed errors is not fulfilled. In Chapter 4 a forward validation (GP-FV) approach is evaluated, where genotypes of a new breeding cycle or a new year not included in the TS are predicted. In this way, the adjusted means from the TS are independent from the means in the VS. One approach to avoid correlated adjusted means within the TS or the VS is to use spectral decomposition of the means to obtain rotated (orthogonalized) means as proposed by Schulz-Streeck et al. [2013b]. Another simpler but not necessarily less computationally demanding approach is to use a full-efficient stage-wise phenotypic analysis [Damesa et al., 2017], where the complete error variance-covariance matrix is submitted from one stage to the other.

Across the stage-wise analyses used in this work, the method suggested by Smith et al. [2001] is the selected method to pass on the information on precision from one stage to the next. This method consists of using as weights the diagonal values of the inverse variance-covariance matrix from the preceding stage. The best approximation in a least-squares sense of the inverse of the variance-covariance matrix is its diagonal matrix [Smith et al., 2001]. It is reasonable to approximate the inverse variance-covariance because it is this matrix that is part of the mixed model equations. This approximation is useful because it allows to identify the weight of an observation as one single measurement reducing the computational load that a fully-efficient analysis entails [Damesa et al., 2017; Möhring and Piepho, 2009].

The prediction model can be validated through GP-CV or GP-FV, where predictive accuracy, the ratio between predictive ability and square root of heritability, is used to assess the prediction performance of the model. There are, however, better ways to estimate predictive accuracy than this "naive" way. Estaghvirou et al. [2013] demonstrate that combining conflicting assumptions in the estimation of predictive ability and heritability renders predictive accuracy inestimable. For instance, RR-BLUP assumes that genotypes are correlated whereas an *ad hoc* estimate of heritability assumes independent genotypes. One method proposed by Estaghvirou et al. [2013] and another one commonly used in animal breeding [Mrode and Thompson, 2005] consistently produced less bias and more precise estimates of

predictive accuracy.

5.4.1 The impact of the relatedness between TS and VS on predictive accuracy

The relatedness degree within the TS and between the TS and the VS are factors determining predictive abilities of GP [Auinger et al., 2016; Daetwyler et al., 2013; Habier et al., 2007; Pszczola et al., 2012; Riedelsheimer et al., 2013]. It has been shown that the RR-BLUP model is strongly sensitive to genetic relationships among individuals so that its efficiency depends in great part on the relatedness degree between TS and VS. Other methods such as BayesB utilize more efficiently the information of LD between QTL and markers because only a small proportion of the total number of markers is fitted [Habier et al., 2007]. Pszczola et al. [2012] demonstrated through simulations that predictive abilities of a dairy cattle population could be improved if the relationships among animals in the TS was minimized and genetic relatedness degree between TS and VS was maximized. Using empirical data of maize double haploids (DH), Riedelsheimer et al. [2013] showed that including additional crosses in the TS when both parents of the individuals of the VS are already in the TS did not improve predictive abilities.

These works have led to investigation of optimization approaches to construct the TS [Bustos-Korts et al., 2016; Isidro et al., 2015; Rincent et al., 2012]. One approach consists on choosing genotypes in the TS that maximize the generalized coefficient of determination (CD) between TS and VS, i.e. the precision of the contrast between each genotype in the VS and the mean of the TS is maximized [Rincent et al., 2012]. An extension of this method is to use a stratified sampling among subpopulations and apply the coefficient of determination mean criterion (CDmean) [Isidro et al., 2015]. Bustos-Korts et al. [2016] suggest constructing a TS by uniformly covering the genetic space of the population of genotypes. All these methods are based on the use of cross validation within a breeding population, where VS are a subset of the complete population. Thus, it makes full sense to ensure that all families or subpopulations have representation in the TS. In Chapter 3, two GP-CV schemes were presented: within crosses (WC) and across crosses (AC). In the WC approach, a random sample of genotypes from each and all the crosses were included in the TS, whereas in the AC approach complete crosses were either kept in the TS or dropped. As expected, the WC approach ($\rho_{GP} \cong 0.69$) that covered the population genetic space outperformed the AC approach ($\rho_{GP} \cong 0.39$). In Chapter 4, using the GP-FV approach, predictive abilities of ~ 0.65 were reached for a specific scenario of the German program (i.e. VS:GCA1-2012 - 1P-scenario, high relatedness between TS and VS), whereas for the Polish program in the same scenario predictive abilities were ~ 0.17. Studies that compare GP-CV against GP-FV report different results. Albrecht et al. [2014] suggest that GP-FV with an appropriate choice of testers and genetic relationships between TS and VS for grain yield and dry matter content of maize led to only slightly reduced accuracies

compared to GP-CV. Auinger et al. [2016] report substantially lower predictive abilities of across-cycles scenarios (i.e. GP-FV) in comparison to within-cycle scenarios (i.e. GP-CV) for grain dry matter yield, plant height and thousand kernel weight in a rye hybrid breeding program.

The choice of a validation scheme that represents a real breeding scenario is of major importance in the evaluation of GP. Efforts to optimize the construction of TS and VS have led to the conclusions that the more related the validation individuals to the TS are, and the higher the genetic coverage in the TS is, the higher the probability of predicting individuals with more accuracy. The breeders' main objective is to predict the average performance of a genotype across time. Therefore, an interesting scenario would be to predict the GEBV of a genotype that has been evaluated in several years. In the rye dataset, one could predict genotypes of a FACT/GCA3 experiment, which have undergone three (or more) years of evaluations, allowing the estimation of genotype main effects free of GY effects. The limitation is that the number of genotypes in FACT/GCA3 experiments is usually less than 50, which is a small VS-size that may lead to high prediction abilities with large standard errors [Schulz-Streeck et al., 2013a; Chapter 4]. Nevertheless, GP-FV using complete cycles in the VS opens up the way to investigate validation scenarios for stability of the genetic material, i.e., the expected consistency of the performance of genotypes across years.

5.5 Merit of the extensions of the genomic prediction model

Among some reasons for extensions of the GP model towards increasing the prediction accuracy, the following can be listed: First, the presence of genotype-by-environment interaction effects [Griffiths et al., 2000; Laidig et al., 2008; Malosetti et al., 2013], second, phenotypic variation is not only due to additive effects but to non-additive effects [Technow et al., 2012; Viana et al., 2016; Wellmann and Bennewitz, 2012], i.e. dominance and epistatic effects, and third, breeding programs aim to improve several traits at the same time [Falconer and Mackay, 1996; Schulthess et al., 2016].

In a LMM framework, GP allows using REML for the variance estimation and further, allows accounting for other sources of variation by adding fixed and random effects [Piepho, 2009b]. Also, the LMM can be customized to account for other assumptions closer to Bayesian methods. For example, to drop the assumption of homogeneous marker variances assumed in RR-BLUP, one could assume an heterogeneous marker variance so that the model mimic the assumption of BayesB from Meuwissen et al. [2001] [Piepho, 2009b]. Assuming a non-normal distribution for the marker effects instead of a normal distribution could be implemented via h-likelihood in a frequentist setting using the so-called hierarchical generalized linear models (HGLM) [Lee et al., 2006], although the option to select such

non-normal distributions is not yet very broadly used.

5.5.1 Models including non-additive effects

The extension of the GP model towards including dominance effects is relatively straightforward. An additional design matrix identifying the homozygous and heterozygous alleles is used. In general, the alleles can be coded as $\{0,1,0\}$ for $\{$AA, Aa, aa$\}$, respectively [Technow et al., 2012; Xiang et al., 2016]. There are several studies supporting advantages of GP models that account for non-additive effects [Guo et al., 2013; Massman et al., 2013; Toro and Varona, 2010]. Also, in the simulations carried out by Technow et al. [2012] assuming no epistatic effects, it was found that modeling dominance effects in the GP model for simulated traits of maize (grain yield and moisture) produced higher predictive accuracies, specially when BayesB was used. Wellmann and Bennewitz [2012] proposed a Bayesian approach named BayesD (and submodels) particularly suited for traits that show overdominance. They demonstrated that depending on the marker panel, the inclusion of dominance effects increased the accuracy of GEBVs.

There are as well studies which favored the pure GP additive model, e.g., for grain yield, oil yield and oil content in sunflower [Reif et al., 2013] and grain yield in wheat [Zhao et al., 2013]. Xu et al. [2014] carried out an analysis on rice hybrid using three GP methods accounting for additive, additive and dominance, and additive, dominance and epistatic effects. They found no noticeable improvement from non-additive models over additive models, probably due to the small TS-size used causing high correlations among the different types of kinship matrices (additive, dominance and epistatic kinship matrices). Xu and Jia [2007] proposed to model epistatic effects by adding a marker-by-marker interaction effect. They use a model that includes a main effect of a locus l and an epistatic effect between loci l and l'. The main difficulty in the implementation of such models is the computational load due to the large number of marker-by-marker interactions. Jiang and Reif [2015] demonstrated, however, that extended GBLUP and RKHS models are equivalent and are an alternative to explicitly model epistasis with reduced computational load. Recently, Xiang et al. [2016] demonstrated that accounting for inbreeding depression had a great impact on predictive accuracy of a trait whose gene action was mainly additive. Legarra et al. [2008] suggest to be cautious when including additional non-additive effects since including these effects in the model may decrease accuracy when they are not truly present. Additive and non-additive effects can show some degree of collinearity. Thus, associated variance components are difficult to estimate accurately [Xu et al., 2014]. Multicollinearity may occur for example, when covariates for epistasis are computed from covariates for the additive effects.

The increase in prediction abilities due to using a GP model that accounts for non-additive effects

seems to strongly depend on the trait architecture [Guo et al., 2013] and on the ratio between dominance variance and additive variance, where dominance variance must be relatively important [Guo et al., 2013; Reif et al., 2013], but there is no threshold yet known that defines how large the dominance variance should be in relation to the total variance.

A simpler approach to account for polygenic variance (the genetic variance not captured by the markers) includes fitting an independent between-individual effect that is required to be separated from the within-individual error component [Piepho, 2009b; Schulz-Streeck and Piepho, 2010]. In practice, fitting a polygenic effect in a GP model can be achieved by fixing the error variance, using for example the square standard error of adjusted genotype means or the Smith weights [Smith et al., 2001] obtained in a previous stage of a stage-wise analysis, the latter being the preferred method all throughout this study.

Goncharenko et al. [2015] found that in rye the expression of the quantitative traits number of productive stalks per square meter, number of grains per ear, plant height and starch content strongly depended on dominance and epistatic effects. Also, frost tolerance, an important trait in rye (due to being the most tolerant species among small grain cereals), is known to be a complex trait with polygenic inheritance [Li et al., 2011a]. It would make sense to exploit extensions of the GP model towards accounting for non-additive effects, especially for traits known to display dominance or epistasis.

In Chapter 3, a polygenic effect was incorporated in the baseline model (M1) for grain dry matter yield. It was found that about 88% of the total genetic variance was captured by the additive effects in the RR-BLUP model. This result suggests that RR-BLUP or GBLUP may be sufficient for dry grain matter yield in rye, which coincide with studies that favored the pure GP additive model [Reif et al., 2013; Xu et al., 2014; Zhao et al., 2013]. The GP-CV methodology constitutes an alternative to evaluate whether accounting for additional non-additive effects in the baseline GP model is advantageous.

5.5.2 Multi-trait genomic prediction

The trait considered throughout this thesis has been dry grain matter yield, which is one of the most relevant traits towards improving productivity. Nonetheless, many economically important traits are also part of the breeding objectives for rye, such as protein content, single ear weight, frost tolerance or degree of resistance/susceptibility to a certain disease [Laidig et al., 2017]. It is often desired to improve several traits at a time, to improve a given trait without reducing performance of another trait and to improve traits that are expensive or difficult to record [Falconer and Mackay, 1996]. Different strategies can be adopted, e.g., use of selection indices, recording the performance of an indicator trait from a trait of interest and using pedigree or marker information [Cooper et al., 2014; Mrode and Thompson,

2005]. Modeling multiple traits takes into account the fact that traits are most likely associated and share a biological basis [Scutari et al., 2014]. Simulation GP approaches have shown that simultaneously modeling multiple quantitative traits results in higher predictive power than using individual traits [Calus and Veerkamp, 2011; Guo et al., 2014a; Hayashi and Iwata, 2013; Jia and Jannink, 2012]. In theory, multi-trait models are most advantageous when the genetic correlation among the traits analyzed is high [Piepho et al., 2008a], and further, a trait which is difficult to measure with good precision leads to low heritability, so that indirect selection using a genetically correlated trait with higher heritability is beneficial [Falconer and Mackay, 1996].

In an applied GS study using a two-trait GS model for grain yield and protein content, which in rye are negatively correlated traits (the higher the grain yield the lower the protein content), Schulthess et al. [2016] confirmed that the multi-trait GS approach make sense when the aim is to predict a low heritability trait with scarce phenotypic records which is supported by a genetically correlated indicator trait highly heritable and extensively phenotyped. Also, Wang et al. [2016] fitted a multi-trait GP model including additive and dominance effects for hybrid rice. They concluded that the prediction accuracy of this model was superior over single-trait GP models, in particular when the trait of interest with low heritability was supported by highly correlated auxiliary traits, whereas for a high-heritability trait, single-trait GP was sufficient. In general, if there is sufficient correlation between a covariate and a variable of interest, using a covariate may lead to considerable gain in accuracy [Piepho and McCulloch, 2002].

As multi-trait GP may become computationally demanding, in particular when the number of traits and phenotypic records increases, a selection index method may be an option to potentially close the gap between GP and multi-trait improvement. It has been advocated that selection index methods are optimal only for balanced phenotypic information, thus they cannot yet benefit from the advantages of multi-trait analysis such as borrowing information from other traits with more phenotypic observations [Schulthess et al., 2016]. Selection index methods are, however, closely related to BLUP-based analysis. Selection indices are essentially a linear combination of genotypic effects of traits with coefficients corresponding to economic weights. In case of unbalanced data, the genotypic effects can be estimated via BLUPs and subsequently be plugged into the linear combination allowing the calculation of the index, and therefore the application of GP for selection indices is feasible.

5.5.3 Non-normally distributed traits

In Chapter 2 the assumptions of the LMM are scrutinized and outlier detection methods are presented as consideration to improve, in particular, the homogeneous error variance assumption. An alternative approach is the use of generalized linear mixed models (GLMM) [Lee et al., 2006; Stroup, 2012], which

allow to relax the distributional assumptions while focusing in searching for a suitable transformation (link function in GLMMs) to achieve linearity or additivity. Extensions of the GP models using GLMM allowing to fit dichotomous, ordinal, or count traits have been studied [Montesinos-López et al., 2015a,b; Technow and Melchinger, 2013]. In the analysis of those traits, one could disregard the fact that the underlying distribution is not normal and use a GBLUP model that assumes that phenotypes follow a normal distribution and that the variance is constant and not a function of the mean value, given the premise that for large sample size and, in the case of ordinal data, a large number of categories, the data may follow an approximate normal distribution [Atkinson, 1988]. Likewise, in plant sciences it is very common to use transformations to stabilize variance, which could be an alternative. It has been shown that transformations can be ineffective and may fail to address the problem of skewness [Stroup, 2015], but also that a correct transformation can be advantageous allowing the use of LMM and facilitating the interpretation [Piepho, 2009a].

In general, the studies on GP using non-normally distributed traits have led to the conclusion that extension of the GP towards GLMM is a viable alternative to deal with non-quantitative data, but have not totally ruled out the use of GBLUP with transformed or untransformed traits [Montesinos-López et al., 2015b; Technow and Melchinger, 2013]. The merit of all these extensions, even if they do not always lead to significantly higher predictive abilities, is that they allow to relax the normality assumption as well as the assumption of homogeneity of variance. This is advantageous but there is still a price to pay, which is that interpretation may be seemingly more difficult for breeders, as they are used to the mean and variance output from a LMM framework. In practice, this translates into curtailment of the take-off of GP based on GLMM models.

5.6 Future perspectives

The results in this thesis demonstrated that genomic selection in hybrid rye breeding is a worthwhile enterprise and goes hand in hand with a holistic analysis process, including pre-processing data, phenotypic analysis and genomic prediction. A refined start for data management towards implementation of genomic selection is provided. The models and procedures are flexible and allow further modifications that may come in the future.

Parallel to the development of GP methods, optimization to the breeding methodology is of great importance since it determines effective resource allocation in routine hybrid breeding schemes [Marulanda et al., 2016]. Certainty on where in the program GS gives the most profit may lead to a more efficient breeding scheme. Research to develop such hybrid schemes (phenotypic and genomic selection) is well

justified.

Preparation for GS using whole-genome sequence (WGS) data should also be a matter of further study. The main advantage of WGS over dense marker data is that the former allows tagging precisely genetic variations (i.e., causal mutations). The computational load for GP methodology represents a remaining challenge. Some simulations of GP using WGS data have already been conducted and show that genotyping by sequencing produces no advantage compared to other marker data [Ober et al., 2012; Pérez-Enciso et al., 2015; van Binsbergen et al., 2015] since there are still the same limitations, such as excessive number of linked and uninformative markers [Gianola, 2013; Pérez-Enciso et al., 2015]. Nonetheless, cases were TS-size is small, WGS integrated with the use of multiple populations and variable selection methods may be advantageous [Iheshiulor et al., 2016]. Since multicollinearity among predictors may lead to incorrect inference of marker effects, recent proposals of whole genome prediction approaches are recently incorporating the correlation among chromosome segments by modeling covariances between SNP effects, e.g., using a first-order antedependence correlation structure [Yang and Tempelman, 2012], or as an autoregressive prior derived from the haplotype frequencies of the population from which an individual parent was derived [Wittenburg et al., 2016].

Optimization of GP methods may generate some advance in prediction accuracies, however, it seems that the main gain lies in the incorporation of meaningful biological information into the prediction model [MacLeod et al., 2016; Pérez-Enciso et al., 2015; van Binsbergen et al., 2015]. Revolutionizing genotyping is tightly linked with a revolution in phenotyping. Information from crop-growth models constitutes a promising source of information to improve GP models and moving from a purely statistical view to an ecophysiological understanding [Malosetti et al., 2016; van Eeuwijk et al., 2005]. A gain in knowledge about genetic factors underlying a trait via GP is only possible if GP is used together with additional genetic analyses, e.g. QTL mapping, genome-wide association analysis (GWAS). This issue has recently been addressed in the frequentist [Bernardo, 2014; Spindel et al., 2016] and the Bayesian contexts [Bennewitz et al., 2017]. Moreover, functional information such as presence or absence of transcription factors and empirical information from population genetic studies (e.g., selective sweeps evidence) constitutes potential biological information to feed into GP methods [Pérez-Enciso et al., 2015]. Exploitation of a better understanding of the selection traits' architecture is another route to extend GP models.

Chapter 6

Conclusions

This thesis developed models at phenotypic level towards incorporating high-dimensional marker data for genomic prediction in a rye hybrid breeding program. Here, the most important conclusions from this work are summarized:

- The linear mixed model framework constitutes a flexible means for data analysis towards genomic selection. Verification of the assumptions in routine analysis can be achieved by implementation of robust outlier detection methods. The routine use of such procedures pinpoints spurious data, controls family-wise error rate and can improve prediction accuracy in further analyses.

- In the German and Polish multi-environment trials analyzed, there was no advantage of spatial modeling over baseline model plus row and column post-blocking factors. Row-column designs may be a promising experimental design alternative in order to account for field heterogeneity and further, allowing a framework to extend to spatial modeling when disease or stress pressure may emerge.

- A viable option for plant breeders is to use historical data to increase training set size and thereby improve predictive accuracy. Bulked multi-year datasets are usually disconnected, making it challenging to precisely dissect year effects from genetic variation. The use of the genetic kinship in modeling genotype-by-year effects allows the separation of genotype-by-year from GEBVs.

Chapter 7

Summary

Technical progress in the genomic field is accelerating developments in plant and animal breeding programs. The access to high-dimensional molecular data has facilitated acquisition of knowledge of genome sequences in many economically important species, which can be used routinely to predict genetic merit. Genomic prediction (GP) has emerged as an approach that allows predicting the genomic estimated breeding value (GEBV) of an unphenotyped individual based on its marker profile. The approach can considerably increase the genetic gain per unit time, as not all individuals need to be phenotyped. Accuracy of the predictions are influenced by several factors and require proper statistical models able to overcome the problem of having more predictor variables than observations.

Plant breeding programs run for several years and genotypes are evaluated in multi-environment trials. Selection decisions are based on the mean performance of genotypes across locations and later on, across years. Under this conditions, linear mixed models offer a suitable and flexible framework to undertake the phenotypic and genomic prediction analyses using a stage-wise approach, allowing refinement of each particular stage. In this work, an evaluation and comparison of outlier detection methods, phenotypic analyses and GP models were considered. In particular, it was studied whether at the plot level, identification and removal of possible outlying observations has an impact on the predictive ability. Further, if an enhancement of phenotypic models by spatial trends leads to improvement of GP accuracy, and finally, whether the use of the kinship matrix can enhance the dissection of GEBVs from genotype-by-year (GY) interaction effects. Here, the methods related to the mentioned objectives are compared using experimental datasets from a rye hybrid breeding program.

Outlier detection methods widely used in many German plant breeding companies were assessed in terms of control of the family-wise error rate and their merits evaluated in a GP framework (Chapter 2). The benefit of implementation of the methods based on a robust scale estimate was that in routine

analysis, such procedures reliably identified spurious data. This outlier detection approach per trial at the plot level is conservative and ensures that adjusted genotype means are not severely biased due to outlying observations. Whenever it is possible, breeders should manually flag suspicious observations based on subject-matter knowledge. Further, removing the flagged outliers identified by the recommended methods did not reduce predictive abilities estimated by cross validation (GP-CV) using data of a complete breeding cycle.

A crucial step towards an accurate calibration of the genomic prediction procedure is the identification of phenotypic models capable of producing accurate adjusted genotype mean estimates across locations and years. Using a two-year dataset connected through a single check, a three-stage GP approach was implemented (Chapter 3). In the first stage, spatial and non-spatial models were fitted per locations and years to obtain adjusted genotype-tester means. In the second stage, adjusted genotype means were obtained per year, and in the third stage, GP models were evaluated. Akaike information criterion (AIC) and predictive abilities estimated from GP-CV were used as model selection criteria in the first and in the third stage. These criteria were used in the first stage, because a choice had to be made between the spatial and non-spatial models and in the third stage, because the predictive abilities allow a comparison of the results of the complete analysis obtained by the alternative stage-wise approaches presented in this thesis. The second stage was a transitional stage where no model selection was needed for a given method of stage-wise analysis. The predictive abilities displayed a different ranking pattern for the models than the AIC, but both approaches pointed to the same best models. The highest predictive abilities obtained for the GP-CV at the last stage did not coincide with the models that AIC and predictive ability of GP-CV selected in the first stage. Nonetheless, GP-CV can be used to further support model selection decisions that are usually based only upon AIC. There was a trend of models accounting for row and column variation to have better accuracies than the counterpart model without row and column effects, thus suggesting that row-column designs may be a potential option to set up breeding trials.

While bulking multi-year data allows increasing the training set size and covering a wider genetic background, it remains a challenge to separate GEBVs from GY effects, when there are no common genotypes across years, i.e., years are poorly connected or totally disconnected. First, an approach considering the two-year dataset connected through a single check, adjusted genotype means were computed per year and submitted to the GP stage (Chapter 3). The year adjustment was done in the GP model by assuming that the mean across genotypes in a given year is a good estimate of the year effect. This assumption is valid because the genotypes evaluated in a year are a sample of the population. Results indicated that this approach is more realistic than relying on the adjustment of a single check.

A further approach entailed the use of kinship to dissect GY effects from GEBVs (Chapter 4). It

was not obvious which method best models the GY effect, thus several approaches were compared and evaluated in terms of predictive abilities in forward validation (GP-FV) scenarios. It was found that for training sets formed by several disconnected years' data, the use of kinship to model GY effects was crucial. In training sets where two or three complete cycles were available (i.e. there were some common genotypes across years within a cycle), using kinship or not yielded similar predictive abilities. It was further shown that predictive abilities are higher for scenarios with high relatedness degree between training and validation sets, and that predicting a selection of top-yielding genotypes was more accurate than predicting the complete validation set when kinship was used to model GY effects.

In conclusion, stage-wise analysis is recommended and it is stressed that the careful choice of phenotypic and genomic prediction models should be made case by case based on subject-matter knowledge and specificities of the data. The analyses presented in this thesis provide general guidelines for breeders to develop phenotypic models integrated with GP. The methods and models described are flexible and allow extensions that can be easily implemented in routine applications.

Chapter 8

Zusammenfassung

Der technische Fortschritt auf dem Gebiet der Genomik ermöglicht eine schnellere Entwicklung in Pflanzen- und Tierzuchtprogrammen. Die Verfügbarkeit von hochdimensionalen, molekularen Daten in vielen ökonomisch wichtigen Tier- und Pflanzenarten erlaubt dessen routinemäßigen Einsatz zur Schätzung und Vorhersage von genetischen Werten. Die genomische Vorhersage (genomic prediction = GP) ermöglicht die Schätzung des genomischen Zuchtwertes eines nicht phänotypisierten Individuums allein auf Grund des Markerprofils. Da nicht alle Individuen phänotypisiert werden müssen, erreicht man mit dieser Herangehensweise einen höheren Selektionsgewinn pro Zeiteinheit. Die Vorhersagegenauigkeit wird durch verschiedene Faktoren beeinflusst und bedarf geeigneter statistischer Modelle. Diese müssen in der Lage sein, Lösungen für ein Gleichungssystem zu finden, obwohl es mehr erklärende Variablen als Beobachtungen gibt. Pflanzenzuchtprogramme erstrecken sich über mehrere Jahre in denen Genotypen an mehreren Versuchsorten wiederholt geprüft werden. Die Selektionsentscheidungen basieren auf der durchschnittlichen Leistung der Genotypen standortübergreifend und später über Jahre hinweg. Für diese Daten stellen gemischte lineare Modelle ein geeignetes und flexibles Werkzeug dar, um die Zuchtwerte der Individuen anhand von phänotypischen oder genetischen Daten vorherzusagen. Die Anwendung dieser Modelle zur Zuchtwertvorhersage kann in zwei Stufen erfolgen, wobei in den beiden Stufen verschiedene Aspekte berücksichtigt werden müssen, um eine valide Zuchtwertschätzung zu erhalten. In dieser Arbeit wurden verschiedene Verfahren zur Bestimmung von Ausreißern, phänotypische Analysen und genomische Vorhersage-Modelle betrachtet. Insbesondere wurde untersucht, ob anhand der Beobachtungsdaten die Identifizierung und Entfernung von möglichen Ausreißern einen Einfluss auf die Vorhersagefähigkeit der verwendeten Modelle hat. Ferner wurde analysiert, ob geostatistische Modelle zu einer Verbesserung der genomischen Vorhersagegenauigkeit führen. Ein weiteres Ziel dieser Arbeit bestand darin, herauszufinden, ob die Verwandtschaftsmatrix eine Trennung des Zuchtwertes von

der Genotyp-Jahr-Interaktion ermöglicht. Die genannten Ziele wurden mit Hilfe eines Hybridroggendatensatzes aus einem Züchtungsprogramm untersucht.

In deutschen Züchtungsunternehmen weitverbreitete Ausreißeridentifizierungsmethoden wurden im Hinblick auf Kontrolle der versuchsbezogenen Irrtumswahrscheinlichkeit und in Bezug auf Verbesserung der genomischen Selektion untersucht (Kapitel 2). Dabei stellte sich heraus, dass diese Verfahren die Ausreißer zuverlässig identifizieren. Dieser Ansatz zur Ausreißererkennung auf Grund von Beobachtungswerten ist konservativ und gewährleistet, dass adjustierte genotypische Mittelwerte nicht aufgrund von Ausreißern verzerrt werden. Züchter sollten verdächtige Beobachtungen basierend auf ihrer Fachkenntnis markieren. Ferner hat ein Entfernen der so identifizierten Ausreißer die Vorhersagefähigkeit nicht reduziert. Die Vorhersagefähigkeit wurde über eine Kreuzvalidierung (cross validation = CV) bestimmt.

Ein entscheidender Schritt zu einer genauen Kalibrierung des genomischen Vorhersageverfahrens ist die Identifizierung von phänotypischen Modellen, die fähig sind, genaue adjustierte genotypische Mittelwerte über Standorte und Jahre hinweg zu liefern. In der vorliegenden Arbeit wurde eine dreistufige GP-Auswertung für einen zweijährigen Datensatz implementiert. Die Daten beider Jahre sind über eine einzige Standardsorte verbunden (Kapitel 3). In der ersten Stufe wurden räumliche und nicht räumliche Modelle an die Daten jedes Standorts und jedes Jahrs angepasst, um die adjustierten Genotyp-Testermittelwerte zu erhalten. In der zweiten Stufe wurden adjustierte genotypische Mittelwerte pro Jahr ermittelt und in der dritten Stufe wurde die Vorhersagegüte der Modelle bewertet. Hierfür wurde sowohl das Akaike Informationkriterium (Akaike information criteria = AIC) als auch die Vorhersagefähigkeit der GP-CV in der ersten und dritten Stufe als Modellauswahlkriterium eingesetzt. In der ersten Stufe wurden diese Kriterien verwendet, weil eine Entscheidung über räumliches und nicht räumliches Modell getroffen werden musste. In der dritten Stufe wurden diese Kriterien verwendet, weil die Vorhersagefähigkeit einen Vergleich der verschiedenen Analysemethoden, die in dieser Arbeit verwendet wurden, ermöglicht. Die zweite Stufe war eine Übergangsstufe, in der keine Modellauswahl benötigt wurde. Die Vorhersagefähigkeit der Modelle zeigt unterschiedliche Rangfolgen, aber beide Modellauswahlkriterien präferieren dasselbe Modell. Das mit GP-CV in der letzten Stufe bestimmte Modell mit der besten Vorhersagefähigkeit, stimmte nicht mit den mittels AIC und mittels GP-CV in der ersten Stufe präferierten Modellen überein. Nichtsdestotrotz kann GP-CV anstelle des AIC zur Modellselektion verwendet werden. Es gab eine Tendenz, dass Modelle, die Zeilen- und Spaltenvariabilität erfassen, eine bessere Vorhersagegenauigkeit aufweisen als Modelle ohne Zeilen- und Spalteneffekte. Dies suggeriert, dass Zeilen-Spalten-Designs eine mögliche Option darstellen, Zuchtversuche anzulegen.

Während kombinierte, mehrjährige Daten größere Trainingsdatensätze erlauben und eine größere

genetischen Variabilität abdecken, bleibt es eine Herausforderung, die Zuchtwerte von der Genotyp-Jahr-Interaktion zu trennen, wenn es kaum oder keine gemeinsame Genotypen über Jahre hinweg gibt. In dem Fall sind Jahre nur schwach verbunden oder komplett unabhängig. Zunächst wurde der zweijährige Datensatz ausgewertet, wobei die Jahre nur über eine einzelne Standardsorte verbunden sind. Es wurden adjustierte genotypische Mittelwerte pro Jahr berechnet, und anschließend in der GP-Stufe verwendet (Kapitel 3). Die Jahreseffekte wurden im GP-Modell als Mittelwert der Genotypen in den verschiedenen Jahren geschätzt. Diese Annahme ist gültig, weil die Genotypen eines Jahres eine Stichprobe der Grundgesamtheit sind. Die Ergebnisse weisen darauf hin, dass dieser Ansatz realistischer ist, als das Abschätzen der Jahreseffekte durch die Standardsorte. Ein weiterer Ansatz bestand darin, Genotyp-Jahr-Interaktionen vom Zuchtwert durch die Nutzung der Verwandtschaftsmatrix zu separieren (Kapitel 4). Hierbei war jedoch nicht offensichtlich, welche Methode die Genotyp-Jahr-Interaktion am besten abbildet. Daher wurden verschiedene Ansätze hinsichtlich der Vorhersagefähigkeit in einer Vorwärts-Validierung verglichen. Dabei stellte sich heraus, dass die Nutzung der Verwandtschaftsmatrix insbesondere dann, wenn es keine gemeinsame Standardsorte gibt, zu einer Verbesserung der Vorhersagefähigkeit führt.

Wenn jedoch ausreichend Genotypen, die in mehreren Jahren getestet wurden, benutzt werden, um Jahreseffekte im GP Modell anzupassen, hat die Nutzung der Verwandtschaftsmatrix weniger Einfluss auf die Vorhersagefähigkeit. Außerdem wurde in den Analysen deutlich, dass bei zunehmendem Verwandtschaftsgrad der Genotypen in den Trainingsdatensätzen die Vorhersagefähigkeit verbessert werden kann und dass die Vorhersagefähigkeit von Genotypen mit den höchsten Zuchtwerten größer ist als die Vorhersagefähigkeit für die restlichen Genotypen. Zusammenfassend kann eine stufenweise Analyse empfohlen werden. Es sei darauf hingewiesen, dass die Modellauswahl für die genomische Selektion von Fall zu Fall, also in Abhängigkeit der Daten und anhand von fachspezifischen Entscheidungen, getroffen werden sollte. Die hier vorgestellten Analysen und Methoden stellen generelle Richtlinien zur Modellselektion in der genomischen Selektion dar, die von Züchtern angewandt werden können.

References

Albrecht, T., Auinger, H.-J., Wimmer, V., Ogutu, J. O., Knaak, C., Ouzunova, M., Piepho, H.-P., and Schön, C.-C. (2014). Genome-based prediction of maize hybrid performance across genetic groups, testers, locations, and years. *Theor Appl Genet*, 127:1375–1386.

Albrecht, T., Wimmer, V., Auinger, H. J., Erbe, M., Knaak, C., Ouzunova, M., Simianer, H., and Schön, C.-C. (2011). Genome-based prediction of testcross values in maize. *Theor Appl Genet*, 123:339–350.

Anscombe, F. J. (1960). Rejection of outliers. *Technometrics*, 2:123–147.

Anscombe, F. J. and Tukey, J. W. (1963). The examination and analysis of residuals. *Technometrics*, 5:141–160.

Ansley, C. F. and Kohn, R. (1987). Efficient generalized cross-validation for state space models. *Biometrika*, 74:139–148.

Arlot, S. and Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Stat Surv*, 4:40–79.

Atkinson, L. (1988). The measurement-statistics controversy: factor analysis and subinterval data. *Bull Psychon Soc*, 26:361–364.

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

Babadi, B., Rasekh, A., Rasekhi, A. A., Zare, K., and Zadkarami, M. R. (2014). A variance shift model for detection of outliers in the linear measurement error model. *Abstr Appl Anal*, 2014:396875.

Barnett, V. and Lewis, T. (2000). *Outliers in Statistical Data*. John Wiley & Sons, New York.

Battenfield, S. D., Guzmán, C., Gaynor, R. C., Singh, R. P., Peña, R. J., Dreisigacker, S., Fritz, A. K., and Poland, J. A. (2016). Genomic selection for processing and end-use quality traits in the CIMMYT spring bread wheat breeding program. *Plant Genome*, 9:2.

Bennewitz, J., Edel, C., Fries, R., Meuwissen, T. H. E., and Wellmann, R. (2017). Application of a Bayesian dominance model improves power in quantitative trait genomewide association analysis. *Genet Sel Evol*, 49:7.

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

Bernal-Vasquez, A.-M., Utz, H. F., and Piepho, H.-P. (2016). Outlier detection methods for generalized lattices: a case study on the transition from ANOVA to REML. *Theor Appl Genet*, 129:787–804.

Bernardo, R. (2008). Molecular markers and selection for complex traits in plants: Learning from the last 20 years. *Crop Sci*, 48:1649–1664.

Bernardo, R. (2014). Genomewide selection when major genes are known. *Crop Sci*, 54:68–75.

Besag, J. and Kempton, R. (1986). Statistical analysis of field experiments using neighbouring plots. *Biometrics*, 42:231–251.

Bradu, D. and Hawkins, D. M. (1982). Location of multiple outliers in two-way tables, using tetrads. *Technometics*, 24:103–108.

Brien, C. J. (1983). Analysis of variance tables based on experimental structure. *Biometrics*, 39:53–59.

Brøndum, R. F., Rius-Vilarrasa, E., Strandén, I., Su, G., Guldbrandtsen, B., Fikse, W. F., and Lund, M. S. (2011). Reliabilities of genomic prediction using combined reference data of the Nordic Red dairy cattle populations. *J Dairy Sci*, 94:4700–4707.

Burgueño, J., Crossa, J., Cotes, J. M., San Vicente, F., and Das, B. (2011). Prediction assessment of linear mixed models for multienvironment trials. *Crop Sci*, 51:944–954.

Burgueño, J., de los Campos, G., Weigel, K., and Crossa, J. (2012). Genomic prediction of breeding values when modeling genotype $\times$ environment interaction using pedigree and dense molecular markers. *Crop Sci*, 52:707–719.

Burnham, K. P. and Anderson, D. R. (1998). *Model Selection and Inference: A Practical Information-Theoretic Approach*. Springer-Verlag, New York.

Bustos-Korts, D., Malosetti, M., Chapman, S., Biddulph, B., and van Eeuwijk, F. A. (2016). Improvement of predictive ability by uniform coverage of the target genetic space. *G3 (Bethesda)*, 6:3733–3747.

Calus, M. P. L. and Veerkamp, R. F. (2011). Accuracy of multi-trait genomic selection using different methods. *Genet Sel Evol*, 43:26.

Cerioli, A., Farcomeni, A., and Riani, M. (2013). Robust distances for outlier-free goodness-of-fit testing. *Comput Stat Data An*, 65:29–45.

Chapman, S. C., Cooper, M., Butler, D. G., and Henzell, R. G. (2000). Genotype by environment interactions affecting grain sorghum. I. Characteristics that confound interpretation of hybrid yield. *Aust J Agric Res*, 51:197–207.

Cochran, W. G. and Cox, G. M. (1957). *Experimental Designs*. John Wiley & Sons, New York, second edition.

Comstock, R. (1977). Quantitative genetics and the design of breeding programs. In Pollack, O., Kempthorne, T., and Jr, B., editors, *Proceedings of the International Conference on Quantitative Genetics, 1621 August 1976*. Iowa State University Press.

Cook, R. D. and Weisberg, S. (1982). *Residuals and Influence in Regression*. Chapman and Hall, London.

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

Crossa, J., Burgueño, J., Cornelius, P. L., McLaren, G., Trethowan, R., and Krishnamachari, A. (2006). Modelling genotype x environment interaction using additive genetic covariances of relatives for predicting breeding values of wheat genotypes. *Crop Sci*, 46:1722–1733.

Crossa, J., Pérez, P., Hickey, J., Burgueño, J., Ornella, L., Ceró N-Rojas, J., Zhang, X., Dreisigacker, S., Babu, R., Li, Y., Bonnett, D., and Mathews, K. (2014). Genomic prediction in CIMMYT maize and wheat breeding programs. *Heredity*, 112:48–60.

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

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

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

Daetwyler, H. D., Calus, M. P. L., Pong-Wong, R., de los Campos, G., and Hickey, J. M. (2013). Genomic prediction in animals and plants: Simulation of data, validation, reporting, and benchmarking. *Genetics*, 193:347–365.

Damesa, T., Möhring, J., Worku, M., and Piepho, H.-P. (2017). One step at a time: stage-wise analysis of series of experiments. *Agron J*, 109:1–13.

De Coninck, A., De Baets, B., Kourounis, D., Verbosio, F., Schenk, O., Maenhout, S., and Fostier, J. (2016). Needles: toward large-scale genomic prediction with marker-by-environment interaction. *Genetics*, 203:543–555.

Dekkers, J. C. M. (2007). Prediction of response to marker-assisted and genomic selection using selection index theory. *J Anim Breed Genet*, 124:331–341.

Duarte, J. B. and Vencovsky, R. (2005). Spatial statistical analysis and selection of genotypes in plant breeding. *Pesqui Agropecu Bras*, 40:107–114.

Endelman, J. B. (2011). Ridge regression and other kernels for genomic selection with R package rrBLUP. *Plant Genome*, 4:250–255.

Endelman, J. B., Atlin, G. N., Beyene, Y., Semagn, K., Zhang, X., Sorrells, M. E., and Jannink, J.-L. (2014). Optimal design of preliminary yield trials with genome-wide markers. *Crop Sci*, 54:48–59.

Erbe, M., Pimentel, E., Sharifi, A., and Simianer, H. (2010). Assessment of cross-validation strategies for genomic prediction in cattle. In *Book of Abstracts of the 9th World Congress of Genetics Applied to Livestock Production*, page S 129, Leipzig, Germany.

Estaghvirou, S. B. O., Ogutu, J. O., and Piepho, H.-P. (2014). Influence of outliers on accuracy estimation in genomic prediction in plant breeding. *G3 (Bethesda)*, 4:2317–2328.

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

Falconer, D. S. and Mackay, T. F. C. (1996). *Introduction to quantitative genetics*. Pearson Prentice Hall, Harlow, fourth edition.

Geiger, H. H. and Miedaner, T. (2009). Rye Breeding. In Carena, M., editor, *Cereals*, pages 157–181. Springer.

Gianola, D. (2013). Priors in whole-genome regression: the Bayesian alphabet returns. *Genetics*, 194:573–596.

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

Goddard, M. (2009). Genomic selection: prediction of accuracy and maximisation of long term response. *Genetica*, 136:245–257.

Gomez, K. A. and Gomez, A. A. (1984). *Statistical Procedures for Agricultural Research*. John Wiley & Sons, New York.

Goncharenko, A., Krahmalev, S., Makarov, V., and Yermakov, S. (2015). Genetic research of quantitative traits of inbred lines of winter rye *(Secale cereale L.)* in diallel crossings. *Agricultural Biology*, 50:75–84.

Greven, S. and Kneib, T. (2010). On the behaviour of marginal and conditional AIC in linear mixed models. *Biometrika*, 97:773–789.

Griffiths, A. J., Miller, J. H., , Suzuki, D. T., Lewontin, R. C., and Gelbart, W. M. (2000). *An Introduction to Genetic Analysis*. Freeman and Company, New York, seventh edition.

Gumedze, F. N. and Chatora, T. D. (2014). Detection of outliers in longitudinal count data via overdispersion. *Comput Stat Data An*, 79:192–202.

Gumedze, F. N. and Jackson, D. (2011). A random effects variance shift model for detecting and accommodating outliers in meta-analysis. *BMC Med Res Methodol*, 11:19.

Gumedze, F. N., Welham, S. J., Gogel, B. J., and Thompson, R. (2010). A variance shift model for detection of outliers in the linear mixed model. *Comput Stat Data An*, 54:2128–2144.

Guo, G., Zhao, F., Wang, Y., Zhang, Y., Du, L., and Su, G. (2014a). Comparison of single-trait and multiple-trait genomic prediction models. *BMC Genet*, 15:30.

Guo, T., Li, H., Yan, J., Tang, J., Li, J., Zhang, Z., Zhang, L., and Wang, J. (2013). Performance prediction of F1 hybrids between recombinant inbred lines derived from two elite maize inbred lines. *Theor Appl Genet*, 126:189–201.

Guo, Z., Tucker, D. M., Basten, C. J., Gandhi, H., Ersoz, E., Guo, B., Xu, Z., Wang, D., and Gay, G. (2014b). The impact of population structure on genomic prediction in stratified populations. *Theor Appl Genet.*, 127:749–762.

Habier, D., Fernando, R. L., and Dekkers, J. C. M. (2007). The impact of genetic relationship information on genome-assisted breeding values. *Genetics*, 177:2389–2397.

Habier, D., Fernando, R. L., and Garrick, D. J. (2013). Genomic BLUP decoded: a look into the black box of genomic prediction. *Genetics*, 194:597–607.

Hampel, F. R. (1985). The breakdown points of the mean combined with some rejection rules. *Technometics*, 27:95–107.

Hayashi, T. and Iwata, H. (2013). A Bayesian method and its variational approximation for prediction of genomic breeding values in multiple traits. *BMC Bioinformatics*, 14:34.

Heffner, E. L., Sorrells, M. E., and Jannink, J.-L. (2009). Genomic selection for crop improvement. *Crop Sci*, 49:1–12.

Heslot, N., Akdemir, D., Sorrells, M. E., and Jannink, J.-L. (2014). Integrating environmental covariates and crop modeling into the genomic selection framework to predict genotype by environment interactions. *Theor Appl Genet*, 127:463–480.

Heslot, N., Jannink, J.-L., and Sorrells, M. E. (2013a). Using genomic prediction to characterize environments and optimize prediction accuracy in applied breeding data. *Crop Sci*, 53:921–933.

Heslot, N., Rutkoski, J., Poland, J., Jannink, J.-L., and Sorrells, M. E. (2013b). Impact of marker ascertainment bias on genomic selection accuracy and estimates of genetic diversity. *PLoS ONE*, 8:e74612.

Heslot, N., Yang, H.-P., Sorrells, M. E., and Jannink, J.-L. (2012). Genomic selection in plant breeding: a comparison of models. *Crop Sci*, 52:146–160.

Hochberg, Y. and Tamhane, A. C. (1987). *Multiple Comparison Procedures*. John Wiley & Sons, New York.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scand J Stat*, 6:65–70.

Hurvitch, C. M. and Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, 76:297–307.

Iglewicz, B. (2000). Robust scale estimators and confidence intervals for location. In Hoaglin, D., Mosteller, F., and Tukey, J. W., editors, *Understanding Robust and Exploratory Data Analysis*. John Wiley & Sons, New York.

Iheshiulor, O. O. M., Woolliams, J. A., Yu, X., Wellmann, R., and Meuwissen, T. H. E. (2016). Within- and across-breed genomic prediction using whole-genome sequence and single nucleotide polymorphism panels. *Genet Sel Evol*, 48:15.

Isidro, J., Jannink, J.-L., Akdemir, D., Poland, J., Heslot, N., and Sorrells, M. E. (2015). Training set optimization under population structure in genomic selection. *Theor Appl Genet*, 128:145–158.

Jannink, J.-L., Lorenz, A. J., and Iwata, H. (2010). Genomic selection in plant breeding: from theory to practice. *Brief Funct Genomics*, 9:166–177.

Jarquín, D., Crossa, J., Lacaze, X., Du Cheyron, P., Daucourt, J., Lorgeou, J., Piraux, F., Guerreiro, L., Pérez, P., Calus, M., Burgueño, J., and de los Campos, G. (2014). A reaction norm model for genomic selection using high-dimensional genomic and environmental data. *Theor Appl Genet*, 127:595–607.

Jia, Y. and Jannink, J.-L. (2012). Multiple-trait genomic selection methods increase genetic value prediction accuracy. *Genetics*, 192:1513–1522.

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

John, J. A. and Williams, E. R. (1995). *Cyclic and Computer Generated Designs*. Chapman and Hall, London, second edition.

Kang, H. M., Zaitlen, N. A., Wade, C. M., Kirby, A., and Heckerman, D. (2008). Efficient control of population structure in model organism association mapping. *Genetics*, 178:1709–1723.

Kemper, K. E., Reich, C. M., Bowman, P. J., Vander Jagt, C. J., Chamberlain, A. J., Mason, B. A., Hayes, B. J., and Goddard, M. E. (2015). Improved precision of QTL mapping using a nonlinear Bayesian method in a multi-breed population leads to greater accuracy of across-breed genomic predictions. *Genet Sel Evol*, 47:29.

Lado, B., González-Barrios, P., Quinke, M., Silva, P., and Gutiérrez, L. (2016). Modeling genotype $\times$ environment interaction for genomic selection with unbalanced data from a wheat breeding program. *Crop Sci*, 56:1–15.

Lado, B., Matus, I., Rodríguez, A., Inostroza, L., Poland, J., Belzile, F., del Pozo, A., Quincke, M., Castro, M., and von Zitzewitz, J. (2013). Increased genomic prediction accuracy in wheat breeding through spatial adjustment of field trial data. *G3 (Bethesda)*, 3:2105–2114.

Laidig, F., Drobek, T., and Meyer, U. (2008). Genotypic and environmental variability of yield for cultivars from 30 different crops in German official variety trials. *Plant Breed*, 127:541–547.

Laidig, F., Piepho, H.-P., Rentel, D., Drobek, T., Meyer, U., and Huesken, A. (2017). Breeding progress, variation and correlation of grain and quality traits in winter rye hybrid and population varieties and national on-farm progress over 26 years. *Theor Appl Genet*, in Press:1–18.

Le Roy, P., Filangi, O., Demeure, O., and Elsen, J.-M. (2012). Comparison of analyses of the XVth QTLMAS common dataset III: Genomic Estimations of Breeding Values. *BMC Proc*, 6 Suppl 2:S3.

Lee, H. and Ghosh, S. K. (2009). Performance of information criteria for spatial models. *JSCS*, 79:93–106.

Lee, Y., Nelder, J. A., and Pawitan, Y. (2006). *Generalized Linear Models with Random Effects: Unified Analysis via H-Likelihood*. Chapman & Hall/CRC, Boca Raton, FL.

Legarra, A., Robert-Granié, C., Manfredi, E., and Elsen, J.-M. (2008). Performance of genomic selection in mice. *Genetics*, 180:611–618.

Leiser, W. L., Rattunde, H. F., Piepho, H.-P., and Parzies, H. K. (2012). Getting the most out of sorghum low-input field trials in West Africa using spatial adjustment. *J Agron Crop Sci*, 198:349–359.

Li, Y., Böck, A., Haseneyer, G., Korzun, V., Wilde, P., Schön, C.-C., Ankerst, D. P., and Bauer, E. (2011a). Association analysis of frost tolerance in rye using candidate genes and phenotypic data from controlled, semi-controlled, and field phenotyping platforms. *BMC Plant Biol*, 11:146.

Li, Y., Haseneyer, G., Schön, C.-C., Ankerst, D., Korzun, V., Wilde, P., and Bauer, E. (2011b). High levels of nucleotide diversity and fast decline of linkage disequilibrium in rye (*Secale cereale L.*) genes involved in frost response. *BMC Plant Biol*, 11:6.

Littell, R. C. (2002). Analysis of unbalanced mixed model data: A case study comparison of ANOVA versus REML/GLS. *J Agric Biol Envir S*, 7:472–490.

Littell, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D., and Schabenberger, O. (2006). *SAS for mixed models*. SAS Institute Inc., NC, USA, second edition.

Lopez-Cruz, M., Crossa, J., Bonnett, D., Dreisigacker, S., Poland, J., Jannink, J.-L., Singh, R. P., Autrique, E., and de los Campos, G. (2015). Increased prediction accuracy in wheat breeding trials using a marker $\times$ environment interaction genomic selection model. *G3 (Bethesda)*, 5:569–582.

Lourenço, V. M. and Pires, A. M. (2014). M-regression, false discovery rates and outlier detection with application to genetic association studies. *Comput Stat Data An*, 78:33–42.

MacLeod, I. M., Bowman, P. J., Vander Jagt, C. J., Haile-Mariam, M., Kemper, K. E., Chamberlain, A. J., Schrooten, C., Hayes, B. J., and Goddard, M. E. (2016). Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits. *BMC Genomics*, 17:144.

Malosetti, M., Bustos-Korts, D., Boer, M., and van Eeuwijk, F. (2016). Predicting responses in multiple environments: Issues in relation to genotype $\times$ environment interactions. *Crop Sci*, 13:2210–2222.

Malosetti, M., Ribaut, J.-M., and van Eeuwijk, F. A. (2013). The statistical analysis of multi-environment data: modeling genotype-by-environment interaction and its genetic basis. *Front Physiol*, 4:44.

Marubini, E. and Orenti, A. (2014). Detecting outliers and/or leverage points: a robust two-stage procedure with bootstrap cut-off points. *Epidemiology Biostatistics and Public Health*, 11:1–17.

Marulanda, J. J., Mi, X., Melchinger, A. E., Xu, J.-L., Würschum, T., and Longin, C. F. H. (2016). Optimum breeding strategies using genomic selection for hybrid breeding in wheat, maize, rye, barley, rice and triticale. *Theor Appl Genet*, 120:1901–1913.

Massman, J. M., Gordillo, A., Lorenzana, R. E., and Bernardo, R. (2013). Genomewide predictions from maize single-cross data. *Theor Appl Genet*, 126:13–22.

McQuarrie, A., Shumway, R., and Tsai, C.-L. (1997). The model selection criterion AIC_u. *Stat Probabil Lett*, 34:285–292.

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

Meyer, K. (2009). Factor-analytic models for genotype $\times$ environment type problems and structured covariance matrices. *Genet Select Evol*, 41:21.

Michel, S., Ametz, C., Gungor, H., Epure, D., Grausgruber, H., Löschenberger, F., and Buerstmayr, H. (2016). Genomic selection across multiple breeding cycles in applied bread wheat breeding. *Theor Appl Genet*, 129:1–11.

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

Möhring, J., Williams, E. R., and Piepho, H.-P. (2014). Efficiency of augmented p-rep designs in multi-environmental trials. *Theor Appl Genet*, 127:1049–1060.

Montesinos-López, O. A., Montesinos-López, A., Pérez-Rodríguez, P., de los Campos, G., Eskridge, K., and Crossa, J. (2015a). Threshold models for genome-enabled prediction of ordinal categorical traits in plant breeding. *G3 (Bethesda)*, 5:291–300.

Montesinos-López, O. A., Montesinos-López, A., Pérez-Rodríguez, P., Eskridge, K., He, X., Juliana, P., Singh, P., and Crossa, J. (2015b). Genomic prediction models for count data. *J Agric Biol Environ Stat*, 20:533–554.

Mrode, R. A. and Thompson, R. (2005). *Linear Models for the Prediction of Animal Breeding Values*. CABI Publishing, Wallingford, UK, second edition.

Mühleisen, J., Piepho, H.-P., Maurer, H. P., Longin, C. F. H., and Reif, J. C. (2014). Yield stability of hybrids versus lines in wheat, barley, and triticale. *Theor Appl Genet*, 127:309–316.

Nakagawa, S. and Schielzeth, H. (2010). Repeatability for Gaussian and non-Gaussian data: a practical guide for biologists. *Biol Rev*, 85:935–956.

Nobre, J. S. and Singer, J. M. (2007). Residual analysis for linear mixed models. *Biom J*, 49:863–875.

Nobre, J. S. and Singer, J. M. (2011). Leverage analysis for linear mixed models. *J Appl Stat*, 38:1063–1072.

Ober, U., Ayroles, J. F., Stone, E. A., Richards, S., Zhu, D., Gibbs, R. A., Stricker, C., Gianola, D., Schlather, M., Mackay, T. F. C., and Simianer, H. (2012). Using whole-genome sequence data to predict quantitative trait phenotypes in *Drosophila melanogaster*. *PLoS Genet*, 8:e1002685.

Ober, U., Erbe, M., Long, N., Porcu, E., Schlather, M., and Simianer, H. (2011). Predicting genetic values: A kernel-based best linear unbiased prediction with genomic data. *Genetics*, 188:695–708.

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

Ou, Z., Tempelman, R. J., Steibel, J. P., Ernst, C. W., Bates, R. O., and Bello, N. M. (2015). Genomic prediction accounting for residual heteroskedasticity. *G3 (Bethesda)*, 6:1–13.

Patterson, H. D. and Hunter, E. A. (1983). The efficiency of incomplete block designs in national list and recommended list cereal variety trials. *J Agric Sci*, 101:427–433.

Perez, P. and de los Campos, G. (2014). BGLR : A statistical package for whole genome regression and prediction. *Genetics*, 198:483–495.

Pérez-Enciso, M., Rincón, J. C., and Legarra, A. (2015). Sequence- vs. chip-assisted genomic selection: accurate biological information is advised. *Genet Sel Evol*, 47:43.

Piepho, H.-P. (1997). Analyzing genotype-environment data by mixed models with multiplicative effects. *Biometrics*, 53:761–766.

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

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

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

Piepho, H.-P., Büchse, A., and Emrich, K. (2003). A hitchhiker's guide to mixed models for randomized experiments. *J Agron Crop Sci*, 189:310–322.

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

Piepho, H.-P., Laidig, F., Drobek, T., and Meyer, U. (2014). Dissecting genetic and non-genetic sources of long-term yield trend in German official variety trials. *Theor Appl Genet*, 127:1009–1018.

Piepho, H.-P. and McCulloch, C. (2002). Can the sample variance estimator be improved by using a covariate?. *J Agr Biol Envir St*, 7:157–175.

Piepho, H.-P. and Möhring, J. (2005). Best linear unbiased prediction of cultivar effects for subdivided target regions. *Crop Sci*, 45:1151–1159.

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

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

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

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

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

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

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

Pinheiro, J. C. and Bates, D. M. (2000). *Mixed-Efects Models in S and S-PLUS*. Statistics and Computing. Springler, New York.

Pinho, L. G. B., Nobre, J. S., and Singer, J. M. (2015). Cook's distance for generalized linear mixed models. *Comput Stat Data An*, 82:126–136.

Plieschke, L., Edel, C., Pimentel, E. C., Emmerling, R., Bennewitz, J., and Götz, K.-U. (2015). A simple method to separate base population and segregation effects in genomic relationship matrices. *Genet Sel Evol*, 47:53.

Pryce, J. E., Gredler, B., Bolormaa, S., Bowman, P. J., Egger-Danner, C., Fuerst, C., Emmerling, R., Sölkner, J., Goddard, M. E., and Hayes, B. J. (2011). Genomic selection using a multi-breed, across-country reference population. *J Dairy Sci*, 94:2625–2630.

Pszczola, M., Strabel, T., Mulder, H. a., and Calus, M. P. L. (2012). Reliability of direct genomic values for animals with different relationships within and to the reference population. *J Dairy Sci*, 95:389–400.

Reif, J. C., Zhao, Y., Würschum, T., Gowda, M., and Hahn, V. (2013). Genomic prediction of sunflower hybrid performance. *Plant Breed*, 132:107–114.

Riedelsheimer, C., Czedik-Eysenberg, A., Grieder, C., Lisec, J., Technow, F., Sulpice, R., Altmann, T., Stitt, M., Willmitzer, L., and Melchinger, A. E. (2012). Genomic and metabolic prediction of complex heterotic traits in hybrid maize. *Nature Genet*, 44:217–220.

Riedelsheimer, C., Endelman, J. B., Stange, M., Sorrells, M. E., Jannink, J.-L., and Melchinger, A. E. (2013). Genomic predictability of interconnected biparental maize populations. *Genetics*, 194:493–503.

Rincent, R., Laloe, D., Nicolas, S., Altmann, T., Brunel, D., Revilla, P., Rodriguez, V. M., Moreno-Gonzalez, J., Melchinger, A. E., Bauer, E., Schön, C.-C., Meyer, N., Giauffret, C., Bauland, C., Jamin, P., Laborde, J., Monod, H., Flament, P., Charcosset, A., and Moreau, L. (2012). Maximizing the reliability of genomic selection by optimizing the calibration set of reference individuals: Comparison of methods in two diverse groups of maize inbreds (*Zea mays L.*). *Genetics*, 192:715–728.

Rocke, D. M. and Woodruff, D. L. (1996). Identification of outliers in multivariate data. *J Am Stat Assoc*, 91:1047–1061.

Ruppert, D. (2011). *Statistics and Data Analysis for Financial Engineering*. Springer, New York.

Rutkoski, J., Singh, R. P., Huerta-Espino, J., Bhavani, S., Poland, J., Jannink, J.-L., and Sorrells, M. E. (2015). Efficient use of historical data for genomic selection: A case study of stem rust resistance in wheat. *Plant Genome*, 8:1.

Schlegel, R. (2016). Hybrid breeding boosted molecular genetics in rye. *Russ J Genet*, 6:569–583.

Schmidt, M., Kollers, S., Maasberg-Prelle, A., Großer, J., Schinkel, B., Tomerius, A., Graner, A., and Korzun, V. (2016). Prediction of malting quality traits in barley based on genome-wide marker data to assess the potential of genomic selection. *Theor Appl Genet*, 129:203–213.

Schulthess, A. W., Wang, Y., Miedaner, T., Wilde, P., Reif, J. C., and Zhao, Y. (2016). Multiple-trait- and selection indices-genomic predictions for grain yield and protein content in rye for feeding purposes. *Theor Appl Genet*, 129:273–287.

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

Schulz-Streeck, T., Ogutu, J. O., Karaman, Z., Knaak, C., and Piepho, H.-P. (2012). Genomic selection using multiple populations. *Crop Sci*, 52:2453–2461.

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

Schulz-Streeck, T. and Piepho, H.-P. (2010). Genome-wide selection by mixed model ridge regression and extensions based on geostatistical models. *BMC Proc*, 4 Suppl 1:S8.

Schützenmeister, A., Jensen, U., and Piepho, H.-P. (2012). Checking normality and homoscedasticity in the general linear model using diagnostic plots. *Commun Stat Simul C*, 41:141–154.

Schützenmeister, A. and Piepho, H.-P. (2012). Residual analysis of linear mixed models using a simulation approach. *Comput Stat Data An*, 56:1405–1416.

Scutari, M., Howell, P., Balding, D. J., and Mackay, I. (2014). Multiple quantitative trait analysis using Bayesian networks. *Genetics*, 198:129–137.

Searle, S. R. (1987). *Linear Models for Unbalanced Data*. John Wiley & Sons, New York.

Searle, S. R., Casella, G., and McCulloch, C. E. (1992). *Variance Components*. John Wiley & Sons, New York.

Smith, A., Cullis, B., and Gilmour, A. (2001). The analysis of crop variety evaluation data in Australia. *Aust NZ J Stat*, 43:129–145.

Snedecor, G. W. and Cochran, W. G. (1980). *Statistical Methods*. The Iowa State University Press, Iowa, seventh edition.

Spilke, J., Richter, C., and Piepho, H.-P. (2010). Model selection and its consequences for different split-plot designs with spatial covariance and trend. *Plant Breed*, 129:590–598.

Spindel, J. E., Begum, H., Akdemir, D., Collard, B., Redoña, E., Jannink, J.-L., and McCouch, S. (2016). Genome-wide prediction models that incorporate *de novo* GWAS are a powerful new tool for tropical rice improvement. *Heredity*, 116:395–408.

Stroup, W. W. (2012). *Generalized Linear Mixed Models: Modern Concepts, Methods and Applications*. CRC Press, Boca Raton, FL.

Stroup, W. W. (2015). Rethinking the analysis of non-normal data in plant and soil science. *Agron J*, 107:811–827.

Swallow, W. and Kianifard, F. (1996). Using robust scale estimates in detecting multiple outliers in linear regression. *Biometrics*, 52:545–556.

Technow, F. and Melchinger, A. E. (2013). Genomic prediction of dichotomous traits with Bayesian logistic models. *Theor Appl Genet*, 126:1133–1143.

Technow, F., Riedelsheimer, C., Schrag, T. A., and Melchinger, A. E. (2012). Genomic prediction of hybrid performance in maize with models incorporating dominance and population specific marker effects. *Theor Appl Genet*, 125:1181–1194.

Technow, F., Schrag, T. A., Schipprack, W., Bauer, E., Simianer, H., and Melchinger, A. E. (2014). Genome properties and prospects of genomic prediction of hybrid performance in a breeding program of maize. *Genetics*, 197:1343–1355.

Thompson, W. A. (1962). The problem of negative estimates of variance components. *Ann Math Stat*, 33:273–289.

Toro, M. A. and Varona, L. (2010). A note on mate allocation for dominance handling in genomic selection. *Genet Sel Evol*, 42:33.

Utz, H. F. (2003). *PLABSTAT Manual*. http://www.uni-hohenheim.de/ipsp/soft.html. Version 3A of 2010-07-19.

Vaida, F. and Blanchard, S. (2005). Conditional Akaike information for mixed-effects models. *Biometrika*, 92:351–370.

van Binsbergen, R., Calus, M. P. L., Bink, M. C. A. M., van Eeuwijk, F. A., Schrooten, C., and Veerkamp, R. F. (2015). Genomic prediction using imputed whole-genome sequence data in Holstein Friesian cattle. *Genet Sel Evol*, 47:71.

Van den Berg, S., Calus, M. P. L., Meuwissen, T. H. E., and Wientjes, Y. C. J. (2015). Across population genomic prediction scenarios in which Bayesian variable selection outperforms GBLUP. *BMC Genet*, 16:146.

van Eeuwijk, F., Malosetti, M., Yin, X., Struik, P., and Stam, P. (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., Bustos-Korts, D. V., and Malosetti, M. (2016). What should students in plant breeding know about the statistical aspects of genotype $\times$ environment interactions?. *Crop Sci*, 56:2119.

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *J Dairy Sci*, 91:4414–4423.

Verbyla, A. P., Cullis, B. R., Kenward, M. G., and Welham, S. J. (1999). The analysis of designed experiments and longitudinal data by using smoothing splines. *J R Stat Soc Ser C Appl Stat*, 48:269–300.

Viana, J. M. S., Piepho, H.-p., and Fonseca, F. (2016). Quantitative genetics theory for genomic selection and efficiency of breeding value prediction in open-pollinated populations. *Sci Agric (Piracicaba, Braz.)*, 73:243–251.

Wang, C., Prakapenka, D., Wang, S., Pulugurta, S., Runesha, H. B., and Da, Y. (2014). GVCBLUP: a computer package for genomic prediction and variance component estimation of additive and dominance effects. *BMC Bioinformatics*, 15:270.

Wang, C.-L., Ma, P.-P., Zhang, Z., Ding, X.-D., Liu, J.-F., Fu, W.-X., Weng, Z.-Q., and Zhang, Q. (2012). Comparison of five methods for genomic breeding value estimation for the common dataset of the 15th QTL-MAS Workshop. *BMC Proc*, 6 Suppl 2:S13.

Wang, X., Li, L., Yang, Z., Zheng, X., Yu, S., Xu, C., and Hu, Z. (2016). Predicting rice hybrid performance using univariate and multivariate GBLUP models based on North Carolina mating design II. *Heredity*, 118:302–310.

Wellmann, R. and Bennewitz, J. (2011). The contribution of dominance to the understanding of quantitative genetic variation. *Genet Res (Camb)*, 93:139–154.

Wellmann, R. and Bennewitz, J. (2012). Bayesian models with dominance effects for genomic evaluation of quantitative traits. *Genet Res (Camb)*, 94:21–37.

Wensch, J., Wensch-Dorendorf, M., and Swalve, H. H. (2013). The evaluation of variance component estimation software: Generating benchmark problems by exact and approximate methods. *Computation Stat*, 28:1725–1748.

Wientjes, Y. C. J., Veerkamp, R. F., and Calus, M. P. L. (2013). The effect of linkage disequilibrium and family relationships on the reliability of genomic prediction. *Genetics*, 193:621–631.

Wilde, P., Bajgain, P., Dopierala, P., Gordillo, A., Korzun, V., Menzel, J., Schmiedchen, B., and Steffan, P. (2015). Genetic gain from hybrid rye breeding: achivements and challenges. In *Book of Abstracts of the the International Conference on Rye Breeding and Genetics*, pages 20–21.

Wilkinson, G. N., Eckert, S. R., Hancock, T. W., and Mayo, O. (1983). Nearest neighbour (nn) analysis of field experiments. *J R Stat Soc Ser B Stat Methodol*, 45:151–211.

Williams, E. R. (1977). Iterative analysis of generalized lattice designs. *Aust J Stat*, 19:39–42.

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

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

Williams, E. R. and Luckett, D. J. (1988). The use of uniformity data in the design and analysis of cotton and barley variety trials. *Aust J Agric Res*, 39:339–350.

Wimmer, V., Albrecht, T., Auinger, H. J., and Schön, C.-C. (2012). synbreed: a framework for the analysis of genomic prediction data using R. *Bioinformatics*, 28:1–29.

Wimmer, V., Lehermeier, C., Albrecht, T., Auinger, H. J., Wang, Y., and Schön, C.-C. (2013). Genome-wide prediction of traits with different genetic architecture through efficient variable selection. *Genetics*, 195:573–587.

Windhausen, V. S., Atlin, G. N., Hickey, J. M., Crossa, J., Jannink, J.-L., Sorrells, M. E., Raman, B., Cairns, J. E., Tarekegne, A., Semagn, K., Beyene, Y., Grudloyma, P., Technow, F., Riedelsheimer, C., and Melchinger, A. E. (2012). Effectiveness of genomic prediction of maize hybrid performance in different breeding populations and environments. *G3 (Bethesda)*, 2:1427–1436.

Wittenburg, D., Teuscher, F., Klosa, J., and Reinsch, N. (2016). Covariance between genotypic effects and its use for genomic inference in half-sib families. *G3 (Bethesda)*, 6:2761–2772.

Wolfinger, R. D. (1996). Heterogeneous variance-covariance structures for repeated measures. *J Agric Biol Environ Stat*, 1:205–230.

Wulff, S. S. (2008). The equality of REML and ANOVA estimators of variance components in unbalanced normal classification models. *Stat Probabil Lett*, 78:405–411.

Xiang, T., Christensen, O. F., Vitezica, Z. G., and Legarra, A. (2016). Genomic evaluation by including dominance effects and inbreeding depression for purebred and crossbred performance with an application in pigs. *Genet Sel Evol*, 48:92.

Xu, S. and Jia, Z. (2007). Genomewide analysis of epistatic effects for quantitative traits in barley. *Genetics*, 175(4):1955–1963.

Xu, S., Zhu, D., and Zhang, Q. (2014). Predicting hybrid performance in rice using genomic best linear unbiased prediction. *Proc Natl Acad Sci U.S.A.*, 111:12456–12461.

Yang, R., Schaeffer, L. R., and Jamrozik, J. (2004). Robust estimation of breeding values in a random regression test-day model. *J Anim Breed Genet*, 121:221–228.

Yang, W. and Tempelman, R. J. (2012). A Bayesian antedependence model for whole genome prediction. *Genetics*, 190:1491–1501.

Zhao, Y., Zeng, J., Fernando, R., and Reif, J. C. (2013). Genomic prediction of hybrid wheat performance. *Crop Sci*, 53:802–810.

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

Zweig, M. H. and Campbell, G. (1993). Receiver-operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine. *Clin Chem*, 39:561–577.

Appendix A

Supplementary material of Chapter 2

A.1 Published datasets: original, with 3 missing observation and with 3 outlying observations

Table A.1: **Alpha design Dataset from John and Williams, 1995, p. 146.** Yield of 24 oats genotypes (gen) were laid out as a -design using 3 replicates (rep) each consisting on six blocks (bl). Three observations were randomly removed, noted as missing values (w3miss), and also modified as representing outliers (w3outl).

Obs	rep	bl	gen	yield	w3miss	w3outl	Obs	rep	bl	gen	yield	w3miss	w3outl
1	1	5	1	5.1202	5.1202	5.1202	51	3	17	24	3.5655	3.5655	3.5655
2	2	10	1	5.7161	.	25.7161	52	1	4	3	3.342	3.342	3.342
3	3	13	1	4.6512	4.6512	4.6512	53	2	8	3	3.7999	3.7999	3.7999
4	1	2	10	4.1736	4.1736	4.1736	54	3	18	3	2.8873	2.8873	2.8873
5	2	10	10	4.9057	4.9057	4.9057	55	1	1	4	4.4461	4.4461	4.4461
6	3	16	10	4.0875	4.0875	4.0875	56	2	7	4	4.3599	4.3599	4.3599
7	1	1	11	4.1172	.	-0.8828	57	3	15	4	4.396	4.396	4.396
8	2	9	11	5.1163	5.1163	5.1163	58	1	1	5	5.8757	5.8757	5.8757
9	3	13	11	3.9205	3.9205	3.9205	59	2	10	5	5.1202	5.1202	5.1202
10	1	6	12	5.256	5.256	5.256	60	3	18	5	4.1972	4.1972	4.1972
11	2	9	12	5.3127	5.3127	5.3127	61	1	6	6	4.7085	.	-5.2915
12	3	16	12	4.1746	4.1746	4.1746	62	2	12	6	5.1751	5.1751	5.1751
13	1	4	13	4.253	4.253	4.253	63	3	15	6	4.2474	4.2474	4.2474
14	2	11	13	5.484	5.484	5.484	64	1	5	7	4.1505	4.1505	4.1505
15	3	16	13	4.7512	4.7512	4.7512	65	2	12	7	4.6297	4.6297	4.6297
16	1	3	14	4.7572	4.7572	4.7572	66	3	18	7	3.6096	3.6096	3.6096
17	2	7	14	4.5294	4.5294	4.5294	67	1	4	8	4.9989	4.9989	4.9989
18	3	13	14	4.3887	4.3887	4.3887	68	2	7	8	3.9926	3.9926	3.9926
19	1	5	15	5.0902	5.0902	5.0902	69	3	14	8	3.9821	3.9821	3.9821
20	2	8	15	4.9114	4.9114	4.9114	70	1	6	9	3.3986	3.3986	3.3986
21	3	14	15	4.6783	4.6783	4.6783	71	2	10	9	4.2955	4.2955	4.2955
22	1	3	16	4.4906	4.4906	4.4906	72	3	14	9	3.1407	3.1407	3.1407

Continue Table A.1

Obs	rep	bl	gen	yield	w3miss	w3outl	Obs	rep	bl	gen	yield	w3miss	w3outl
23	2	12	16	5.3024	5.3024	5.3024	38	2	7	20	3.6056	3.6056	3.6056
24	3	17	16	4.3852	4.3852	4.3852	39	3	18	20	3.7349	3.7349	3.7349
25	1	5	17	4.7876	4.7876	4.7876	40	1	2	21	4.654	4.654	4.654
26	2	9	17	5.0744	5.0744	5.0744	41	2	9	21	5.3802	5.3802	5.3802
27	3	15	17	4.3234	4.3234	4.3234	42	3	17	21	4.413	4.413	4.413
28	1	3	18	3.9737	3.9737	3.9737	43	1	1	22	4.5784	4.5784	4.5784
29	2	11	18	5.0988	5.0988	5.0988	44	2	11	22	5.0969	5.0969	5.0969
30	3	15	18	4.2486	4.2486	4.2486	45	3	17	22	4.2397	4.2397	4.2397
31	1	4	19	4.7269	4.7269	4.7269	46	1	3	23	4.2323	4.2323	4.2323
32	2	12	19	5.3148	5.3148	5.3148	47	2	8	23	4.3042	4.3042	4.3042
33	3	13	19	4.5552	4.5552	4.5552	48	3	16	23	3.8721	3.8721	3.8721
34	1	2	2	4.335	4.335	4.335	49	1	6	24	4.9577	4.9577	4.9577
35	2	11	2	5.1566	5.1566	5.1566	50	2	8	24	3.9039	3.9039	3.9039
36	3	14	2	4.051	4.051	4.051							
37	1	2	20	4.0141	4.0141	4.0141							

Table A.2: **Triple lattice design Dataset from Gomez and Gomez, 1984, p.55-56.** Triple lattice design - dataset. Grain yield data (yield) in ton/ha from a trial of 81 upland rice varieties was conducted in a 9 x 9 triple lattice design [9 replicates (rep) and 9 blocks (block)].). Three observations were randomly removed, noted as missing values (w3miss), and also modified as representing outliers (w3outl).

Obs	rep	block	gen	yield	w3miss	w3out	Obs	rep	block	gen	yield	w3miss	w3out
1	1	1	1	2.7	2.7	2.7	26	1	3	26	5.87	5.87	5.87
2	1	1	2	1.6	1.6	1.6	27	1	3	27	4.2	4.2	4.2
3	1	1	3	4.45	4.45	4.45	28	1	4	28	3.74	3.74	3.74
4	1	1	4	2.91	2.91	2.91	29	1	4	29	3.05	3.05	3.05
5	1	1	5	2.78	2.78	2.78	30	1	4	30	5.16	5.16	5.16
6	1	1	6	3.32	3.32	3.32	31	1	4	31	4.76	4.76	4.76
7	1	1	7	1.7	1.7	1.7	32	1	4	32	3.75	3.75	3.75
8	1	1	8	4.72	.	14.72	33	1	4	33	3.66	3.66	3.66
9	1	1	9	4.79	4.79	4.79	34	1	4	34	4.52	4.52	4.52
10	1	2	10	4.2	4.2	4.2	35	1	4	35	4.64	4.64	4.64
11	1	2	11	5.22	5.22	5.22	36	1	4	36	5.36	5.36	5.36
12	1	2	12	3.96	3.96	3.96	37	1	5	37	4.76	4.76	4.76
13	1	2	13	1.51	1.51	1.51	38	1	5	38	4.43	4.43	4.43
14	1	2	14	3.48	3.48	3.48	39	1	5	39	5.36	5.36	5.36
15	1	2	15	4.69	4.69	4.69	40	1	5	40	4.73	4.73	4.73
16	1	2	16	1.57	1.57	1.57	41	1	5	41	5.3	5.3	5.3
17	1	2	17	2.61	2.61	2.61	42	1	5	42	3.93	3.93	3.93
18	1	2	18	3.16	3.16	3.16	43	1	5	43	3.37	3.37	3.37
19	1	3	19	4.63	4.63	4.63	44	1	5	44	3.74	3.74	3.74
20	1	3	20	3.33	3.33	3.33	45	1	5	45	4.06	4.06	4.06
21	1	3	21	6.31	6.31	6.31	46	1	6	46	3.45	3.45	3.45
22	1	3	22	6.08	6.08	6.08	47	1	6	47	2.56	2.56	2.56
23	1	3	23	1.86	1.86	1.86	48	1	6	48	2.39	2.39	2.39
24	1	3	24	4.1	4.1	4.1	49	1	6	49	2.3	2.3	2.3
25	1	3	25	5.72	5.72	5.72	50	1	6	50	3.54	3.54	3.54

Continue Table A.2

Obs	rep	block	gen	yield	w3miss	w3out	Obs	rep	block	gen	yield	w3miss	w3out
51	1	6	51	3.66	3.66	3.66	106	2	7	25	5.55	5.55	5.55
52	1	6	52	1.2	1.2	1.2	107	2	8	26	5.14	5.14	5.14
53	1	6	53	3.34	3.34	3.34	108	2	9	27	3.94	3.94	3.94
54	1	6	54	4.04	4.04	4.04	109	2	1	28	3.75	3.75	3.75
55	1	7	55	3.99	3.99	3.99	110	2	2	29	4.06	.	24.06
56	1	7	56	4.48	4.48	4.48	111	2	3	30	4.99	4.99	4.99
57	1	7	57	2.69	2.69	2.69	112	2	4	31	3.71	3.71	3.71
58	1	7	58	3.95	3.95	3.95	113	2	5	32	4.34	4.34	4.34
59	1	7	59	2.59	2.59	2.59	114	2	6	33	3.84	3.84	3.84
60	1	7	60	3.99	3.99	3.99	115	2	7	34	3.52	3.52	3.52
61	1	7	61	4.37	4.37	4.37	116	2	8	35	4.32	4.32	4.32
62	1	7	62	4.24	4.24	4.24	117	2	9	36	4.51	4.51	4.51
63	1	7	63	3.7	3.7	3.7	118	2	1	37	4.08	4.08	4.08
64	1	8	64	5.29	5.29	5.29	119	2	2	38	3.89	3.89	3.89
65	1	8	65	3.58	3.58	3.58	120	2	3	39	4.58	4.58	4.58
66	1	8	66	2.14	2.14	2.14	121	2	4	40	4.85	4.85	4.85
67	1	8	67	5.54	5.54	5.54	122	2	5	41	4.36	4.36	4.36
68	1	8	68	5.14	5.14	5.14	123	2	6	42	4.25	4.25	4.25
69	1	8	69	5.73	5.73	5.73	124	2	7	43	4.03	4.03	4.03
70	1	8	70	3.38	3.38	3.38	125	2	8	44	3.47	3.47	3.47
71	1	8	71	3.63	3.63	3.63	126	2	9	45	3.1	3.1	3.1
72	1	8	72	5.08	5.08	5.08	127	2	1	46	3.88	3.88	3.88
73	1	9	73	3.76	3.76	3.76	128	2	2	47	2.6	2.6	2.6
74	1	9	74	6.45	6.45	6.45	129	2	3	48	3.17	3.17	3.17
75	1	9	75	3.96	3.96	3.96	130	2	4	49	2.87	2.87	2.87
76	1	9	76	3.64	3.64	3.64	131	2	5	50	3.24	3.24	3.24
77	1	9	77	4.42	4.42	4.42	132	2	6	51	3.9	3.9	3.9
78	1	9	78	6.57	6.57	6.57	133	2	7	52	1.2	1.2	1.2
79	1	9	79	6.39	6.39	6.39	134	2	8	53	3.41	3.41	3.41
80	1	9	80	3.39	3.39	3.39	135	2	9	54	3.59	3.59	3.59
81	1	9	81	4.89	4.89	4.89	136	2	1	55	2.14	2.14	2.14
82	2	1	1	3.06	3.06	3.06	137	2	2	56	4.19	4.19	4.19
83	2	2	2	1.61	1.61	1.61	138	2	3	57	2.69	2.69	2.69
84	2	3	3	4.19	4.19	4.19	139	2	4	58	3.79	3.79	3.79
85	2	4	4	2.99	2.99	2.99	140	2	5	59	3.62	3.62	3.62
86	2	5	5	3.81	3.81	3.81	141	2	6	60	3.64	3.64	3.64
87	2	6	6	3.34	3.34	3.34	142	2	7	61	4.36	4.36	4.36
88	2	7	7	2.98	2.98	2.98	143	2	8	62	3.74	3.74	3.74
89	2	8	8	4.2	4.2	4.2	144	2	9	63	2.7	2.7	2.7
90	2	9	9	4.75	4.75	4.75	145	2	1	64	3.68	3.68	3.68
91	2	1	10	2.08	2.08	2.08	146	2	2	65	3.14	3.14	3.14
92	2	2	11	5.3	5.3	5.3	147	2	3	66	2.57	2.57	2.57
93	2	3	12	3.33	3.33	3.33	148	2	4	67	5.28	5.28	5.28
94	2	4	13	2.5	2.5	2.5	149	2	5	68	4.49	4.49	4.49
95	2	5	14	3.48	3.48	3.48	150	2	6	69	5.09	5.09	5.09
96	2	6	15	3.3	3.3	3.3	151	2	7	70	3.18	3.18	3.18
97	2	7	16	2.69	2.69	2.69	152	2	8	71	3.67	3.67	3.67
98	2	8	17	2.69	2.69	2.69	153	2	9	72	4.4	4.4	4.4
99	2	9	18	2.59	2.59	2.59	154	2	1	73	2.85	2.85	2.85
100	2	1	19	2.95	2.95	2.95	155	2	2	74	4.82	4.82	4.82
101	2	2	20	2.75	2.75	2.75	156	2	3	75	3.82	3.82	3.82
102	2	3	21	4.67	4.67	4.67	157	2	4	76	3.32	3.32	3.32
103	2	4	22	4.87	4.87	4.87	158	2	5	77	3.62	3.62	3.62
104	2	5	23	1.87	1.87	1.87	159	2	6	78	6.1	6.1	6.1
105	2	6	24	3.68	3.68	3.68	160	2	7	79	6.77	6.77	6.77

Continue Table A.2

Obs	rep	block	gen	yield	w3miss	w3out	Obs	rep	block	gen	yield	w3miss	w3out
161	2	8	80	2.27	2.27	2.27	211	3	4	47	2.58	2.58	2.58
162	2	9	81	4.86	4.86	4.86	212	3	5	48	1.89	1.89	1.89
163	3	1	1	3.52	3.52	3.52	213	3	6	46	4.18	4.18	4.18
164	3	2	2	0.79	0.79	0.79	214	3	7	50	2.87	2.87	2.87
165	3	3	3	4.69	4.69	4.69	215	3	8	51	3.35	3.35	3.35
166	3	4	4	3.06	3.06	3.06	216	3	9	49	3.05	3.05	3.05
167	3	5	5	3.79	3.79	3.79	217	3	1	58	3.75	3.75	3.75
168	3	6	6	3.34	3.34	3.34	218	3	2	59	3.59	3.59	3.59
169	3	7	7	2.35	2.35	2.35	219	3	3	60	4.66	4.66	4.66
170	3	8	8	4.51	4.51	4.51	220	3	4	61	4.27	4.27	4.27
171	3	9	9	4.21	4.21	4.21	221	3	5	62	3.73	3.73	3.73
172	3	1	12	2.18	2.18	2.18	222	3	6	63	2.7	2.7	2.7
173	3	2	10	3.58	3.58	3.58	223	3	7	55	2.99	2.99	2.99
174	3	3	11	5.33	5.33	5.33	224	3	8	56	3.61	3.61	3.61
175	3	4	15	4.3	4.3	4.3	225	3	9	57	3.19	3.19	3.19
176	3	5	13	0.88	0.88	0.88	226	3	1	69	4.45	4.45	4.45
177	3	6	14	3.94	3.94	3.94	227	3	2	67	5.06	5.06	5.06
178	3	7	18	2.87	2.87	2.87	228	3	3	68	4.5	4.5	4.5
179	3	8	16	1.26	1.26	1.26	229	3	4	72	4.84	4.84	4.84
180	3	9	17	3.17	3.17	3.17	230	3	5	70	3.51	3.51	3.51
181	3	1	20	3.5	3.5	3.5	231	3	6	71	3.96	3.96	3.96
182	3	2	21	4.83	4.83	4.83	232	3	7	66	1.62	1.62	1.62
183	3	3	19	4.43	4.43	4.43	233	3	8	64	4.52	4.52	4.52
184	3	4	23	2.02	2.02	2.02	234	3	9	65	2.63	2.63	2.63
185	3	5	24	3.4	3.4	3.4	235	3	1	77	4.14	4.14	4.14
186	3	6	22	5.72	5.72	5.72	236	3	2	78	6.51	.	0.51
187	3	7	26	5.5	5.5	5.5	237	3	3	76	4.5	4.5	4.5
188	3	8	27	4.2	4.2	4.2	238	3	4	80	2.74	2.74	2.74
189	3	9	25	5.03	5.03	5.03	239	3	5	81	3.5	3.5	3.5
190	3	1	34	3.3	3.3	3.3	240	3	6	79	3.48	3.48	3.48
191	3	2	35	3.63	3.63	3.63	241	3	7	74	5.33	5.33	5.33
192	3	3	36	5.31	5.31	5.31	242	3	8	75	3.38	3.38	3.38
193	3	4	28	3.57	3.57	3.57	243	3	9	73	4.06	4.06	4.06
194	3	5	29	4.92	4.92	4.92							
195	3	6	30	5.34	5.34	5.34							
196	3	7	31	2.72	2.72	2.72							
197	3	8	32	3.19	3.19	3.19							
198	3	9	33	3.34	3.34	3.34							
199	3	1	45	3.88	3.88	3.88							
200	3	2	43	3.02	3.02	3.02							
201	3	3	44	4.13	4.13	4.13							
202	3	4	39	5.8	5.8	5.8							
203	3	5	37	2.12	2.12	2.12							
204	3	6	38	4.47	4.47	4.47							
205	3	7	42	4.2	4.2	4.2							
206	3	8	40	4.76	4.76	4.76							
207	3	9	41	5.31	5.31	5.31							
208	3	1	53	2.45	2.45	2.45							
209	3	2	54	4.2	4.2	4.2							
210	3	3	52	1.98	1.98	1.98							

Table A.3: **Square lattice Dataset from Cochran and Cox, 1987, p 406.** Yield (obs) of 25 soybean varieties (t) laid out in a 5 x 5 simple lattice [5 replicates (rep) and 5 blocks (bl)]. Three observations were randomly removed, noted as missing values (obs_3m), and also modified as representing outliers (3outl).

Obs	rep	bl	t	obs	obs_3m	3outl
1	1	1	1	6	.	6
2	2	6	1	24	24	44
3	1	1	2	7	7	7
4	2	7	2	21	21	21
5	1	1	3	5	5	5
6	2	8	3	16	16	16
7	1	1	4	8	8	8
8	2	9	4	17	17	17
9	1	1	5	6	6	6
10	2	10	5	15	15	15
11	1	2	6	16	16	6
12	2	6	6	13	13	13
13	1	2	7	12	12	12
14	2	7	7	11	11	11
15	1	2	8	12	12	12
16	2	8	8	4	4	4
17	1	2	9	13	13	13
18	2	9	9	10	10	10
19	1	2	10	8	8	8
20	2	10	10	15	15	15
21	1	3	11	17	17	12
22	2	6	11	24	24	24
23	1	3	12	7	.	7
24	2	7	12	14	14	14
25	1	3	13	7	7	7
26	2	8	13	12	12	12
27	1	3	14	9	9	9
28	2	9	14	30	.	30
29	1	3	15	14	14	14
30	2	10	15	22	22	22
31	1	4	16	18	18	18
32	2	6	16	11	11	11
33	1	4	17	16	16	16
34	2	7	17	11	11	11
35	1	4	18	13	13	13
36	2	8	18	12	12	12
37	1	4	19	13	13	13
38	2	9	19	9	9	9
39	1	4	20	14	14	14
40	2	10	20	16	16	16
41	1	5	21	14	14	14
42	2	6	21	8	8	8
43	1	5	22	15	15	15
44	2	7	22	23	23	23
45	1	5	23	11	11	11
46	2	8	23	12	12	12
47	1	5	24	14	14	14
48	2	9	24	23	23	23
49	1	5	25	14	14	14
50	2	10	25	19	19	19

Table A.4: **Rectangular lattice Dataset from Cochran and Cox, 1987, p 418.** Artificial data (obs) for 12 treatments (t) laid out in a 3 x 4 rectangular lattice [3 replicates (rep) and 4 blocks (bl)]. Three observations were randomly removed, noted as missing values (obs3m), and also modified as representing outliers (obs3outl).

Obs	rep	bl	t	obs	obs3m	obs3outl
1	1	1	1	16	16	16
2	2	6	1	17	.	37
3	3	12	1	22	22	22
4	1	1	2	9	9	9
5	2	7	2	10	10	10
6	3	10	2	15	15	15
7	1	1	3	4	4	4
8	2	8	3	11	11	11
9	3	11	3	3	3	3
10	1	2	4	0	0	0
11	2	5	4	5	5	5
12	3	11	4	1	1	1
13	1	2	5	3	3	3
14	2	7	5	6	6	6
15	3	12	5	11	11	11
16	1	2	6	11	11	11
17	2	8	6	20	.	10
18	3	9	6	15	15	15
19	1	3	7	16	16	16
20	2	5	7	14	14	14
21	3	12	7	17	17	17
22	1	3	8	23	23	23
23	2	6	8	19	19	19
24	3	9	8	20	20	20
25	1	3	9	15	15	15
26	2	8	9	17	17	17
27	3	10	9	16	16	16
28	1	4	10	7	7	7
29	2	5	10	6	6	6
30	3	10	10	9	9	9
31	1	4	11	11	.	6
32	2	6	11	8	8	8
33	3	11	11	6	6	6
34	1	4	12	12	12	12
35	2	7	12	9	9	9
36	3	9	12	10	10	10

A.2 Codes - SAS and R

Codes for generalized lattice model

```
************************  CODES IN SAS ***********************************;

***  Code to resemble PlabStat (ANOVA) procedure for analysis
     of generalized lattices ------------------------------------;

proc mixed data=dataset_name method=type1;
* method=type1 switch to ANOVA approach;
class rep        block        genotype;
model yield = rep genotype  ;
random rep*block  ;
run;


***  Code for REML-based analysis
     of generalized lattices ------------------------------------;

proc mixed data=dataset_name method=reml; *nobound;
* method=reml is the default method in SAS. The user can also omit this option;
* nobound allows computation of negative variance components;
class rep        block        genotype;
model yield = rep genotype;
random rep*block  ;
run;


##############################  CODES IN R ###################################

###  Code to resemble REML procedure for analysis
###  of generalized lattices

# Set working directory
Setwd("E:/Folder")
# Libraries
library(lme4)
library(lsmeans)
library(pbkrtest)

# Read file
dataset_name <- read.delim("E:/Folder/dataset_name.txt")

# Set as factors
dataset_name $rep=as.factor(dataset_name $rep)
dataset _name $block=as.factor(dataset_name $block)
dataset _name $gen=as.factor(dataset_name $gen)

# Analysis for original data
lmer.data=lmer(yield ~  gen + rep + (1|rep:block), data=  dataset_name )

################################## Using ASReml-R ########################
library(asreml)

attach(dataset_name)
```

```
asreml.data = asreml(fixed= yield ~  gen + rep, random=~rep:block,
                           data=dataset_name,
                           na.method.Y="omit",
                           na.method.X="omit", maxiter=100,
                           workspace=1e9)
summary(asreml.data)
# Values are slightly different.
# We only use ASReml to produce the student residuals.
```

Codes for outlier methods

```
*************************  CODES IN SAS ***********************************;


*****************************************************************************;
**** Method 1: PlabStat - REML (PS-REML) ********************************;
*****************************************************************************;


*** Analysis of generalized lattice and generation of residuals *************;

*** M1r: PlabStat using incomplete blocks as random effects *****************;
proc mixed data=dataset_name method=reml ;
class rep        block        genotype;
model yield = rep genotype  /s  residual outp=file_with_residuals;
* file_with_residuals is the file containing residuals and
* error degrees of freedom;
random rep*block /s;
lsmeans genotype /diff;
* Produce pair-wise comparisons of treatment or genotype means;
ods output diffs= diffs;
* Save comparisons and standard errors in a file called diffs;
run;


*** M1f: PlabStat using incomplete blocks as fixed effects ******************;
proc mixed data=dataset_name method=reml ;
class rep        block        genotype;
model yield = rep genotype rep*block /s  residual outp=file_with_residuals;
lsmeans genotype /diff;
ods output diffs= diffs;
run;


**********************;
*** Flag the outliers (procedure is the same for M1r and M1f);

** Produce a variable with the variance of each comparison;
data diffs;
set diffs;
vd = stderr**2;
run;

** Produce a file (mvd) containing the mean of the variance of the
   differences between treatment or genotype means ;
```

```
proc means data=diffs noprint;
VAR vd;
output out=mvd mean=mean;
run;

** Store the mean of the variance of the difference between treatment means
   in a variable called mvd;
data _null_;
set mvd;
call symput('mvd',mean); /*call symput('macro_variable',value) */
run;

** Calculate the standard factor. Rep_num is the number of replicates
   and is given manually;
data MSE_Eff;
MSE_Eff = sqrt( &mvd * Rep_num/2);
run;

** Store the standard factor in a variable called MSE_Eff;
data _null_;
set MSE_Eff;
call symput('MSE_Eff',MSE_Eff); /*call symput('macro_variable',value) */
run;

** Compute the MAD of the residuals and store it in a file named Mad;
proc univariate data=file_with_residuals noprint;
output out=Mad MAD=MAD n=n;
var resid;
run;

** Compute the re-scaled MAD in a variable called re_MAD;
data Mad;
set Mad;
re_MAD = MAD*1.4826;
run;

** Store the re-scaled MAD in a variable called s_rob;
data _null_;
set Mad;
call symput('re_MAD',re_MAD); /*call symput('macro_variable',value) */
call symput('n',n);
run;

/* PlabStat-Reml main procedure */
data PS_file;
set file_with_residuals;

MSE_Eff = &MSE_Eff;
res_PS = resid/MSE_Eff;

z = 1 - (DF*0.005/&n); /* z = dfe*premium/n_obs, premium = 0.005, DF=errorDF */
Gau_z = probit(z); /* inverse of the standard normal cdf of z */
K = 1.40 + 0.85*Gau_z;
C = K*(1 - ((K**2 - 2) / (4 * DF)))* sqrt(DF/&n);
if C < 1.5 then C = 1.5;

s_thresh = &re_MAD * C * 1.15;
```

```
if MSE_Eff < s_thresh then res_tresh = s_thresh;
else res_tresh = MSE_Eff;

if abs(resid) > res_tresh then PS_out=1; else PS_out = 0;
* Outliers are the observations flagged with 1;
run;


*************************************************************************;
**** Method 2: Bonferroni-Holm using studentized residual (BH-ST) ***********;
*************************************************************************;

*** Analysis of generalized lattice and generation of residuals *************;

*** M2r: Bonferroni-Holm using studentized residuals and with
             incomplete blocks as random effects;
proc mixed data=dataset_name method=reml;
class rep        block        genotype        ;
model w3outl = rep genotype  /s residual  outp=file_with_residuals;
random rep*block/s;
run;

*** M2f: Bonferroni-Holm using studentized residuals and with
             incomplete blocks as fixed effects;
proc mixed data=dataset_name method=reml;
class rep        block        genotype        ;
model w3outl = rep genotype rep*block /s residual  outp=file_with_residuals;
run;

*** Generate two variables: one with the absolute values of the studentized
    residuals, the other with the corresponding p-value.
    The name of this second variable must be
    raw_p to let the multtest procedure recognize the p-values
    -------------------------;
data file_with_residuals;
set file_with_residuals;
stud_res = abs(StudentResid);
raw_p = 2 * (1 - probnorm(stud_res));
run;

*** Sort by p-value;
proc sort data =file_with_residuals;
by raw_p ;
run;

*** Bonferroni-Holm test ;
proc multtest inpvalues=file_with_residuals holm out=bholm_file;
run;

*** Flag observations with significant test;
data bholm_file ;
set bholm_file ;
if stpbon_p < 0.05 then out_bon_res = 1; else out_bon_res = 0;
* Outliers are the observations flagged with 1;
run;


*************************************************************************;
```

```
**** Method 3: Studentized residual razor (SRR) ****************************;
*************************************************************************;

*** Analysis of generalized lattice and generation of residuals *************;

*** M3r: Studentized residual razor with blocks as random effects;
proc mixed data=dataset_name method=reml;
class rep        block        genotype        ;
model w3outl = rep genotype  /s residual  outp=file_with_residuals;
random rep*block/s;
run;

*** M3f: Studentized residual razor with blocks as fixed effects;
proc mixed data=dataset_name method=reml;
class rep        block        genotype        ;
model w3outl = rep genotype rep*block /s residual  outp=file_with_residuals;
run;
*** Generate a variable with the absolute values of the studentized residuals
    and flag the observations with absolute value of studentized residual
    greater than 2.8;
data srr_file;
set file_with_residuals;
stud_res = abs(StudentResid);
if stud_res > 2.8 then out_stud = 1; else out_stud = 0;
* Outliers are the observations flagged with 1;
run;


*************************************************************************;
**** Method 4: Bonferroni-Holm  using re-scaled MAD for standardizing *******;
************************** residuals (BH-MADR) ****************************;
*************************************************************************;


*** Analysis of generalized lattice and generation of residuals *************;

*** M4r: Bonferroni-Holm using re-scaled MAD and incomplete blocks as random;
proc mixed data=dataset_name method=reml;
class rep        block        genotype        ;
model w3outl = rep genotype  /s residual  outp=file_with_residuals;
random rep*block/s;
run;

*** M4f: Bonferroni-Holm using re-scaled MAD and incomplete blocks as fixed;
proc mixed data=dataset_name method=reml;
class rep        block        genotype        ;
model w3outl = rep genotype rep*block /s residual  outp=file_with_residuals;
run;

** Compute the MAD of the residuals and store in a file called Mad;
proc univariate data=preds noprint;
output out=Mad MAD=MAD n=n;
var resid;
run;

** Compute the re-scaled MAD in a variable called re_MAD;
data Mad;
set Mad;
```

```
re_MAD = MAD*1.4826;
run;

*** Store the re-scaled MAD in a variable called re_MAD;
data _null_;
set Mad;
call symput('re_MAD',re_MAD); /*call symput('macro_variable',value) */
call symput('n',n);
run;

*** Generate two variables: one with the absolute values of the residuals
    standadized using the re-scaled MAD (re_MAD), the other with the
    corresponding p-value. The name of this second variable must be
    raw_p to let the multtest procedure recognize the values ------------;
data BHmad_file;
set file_with_residuals;
studMAD_res = abs(resid/&re_MAD);
raw_p = 2 * (1 - probnorm(studMAD_res));
run;

*** Sort by p-value;
proc sort data=BHmad_file;
by raw_p;
run;

*** Bonferroni-Holm test ;
proc multtest inpvalues=BHmad_file holm out=holm_BHmad_file noprint;
run;

*** Flag observations with significant test;
data holm_BHmad_file;
set holm_BHmad_file;
if stpbon_p < 0.05 then bhmad_res = 1; else bhmad_res = 0;
* Outliers are the observations flagged with 1;
if raw_p = . then bhmad_res = 0;
* This step needs to be added to avoid flagging missing
  observations as outliers;
run;


*************************************************************************;
**** Method 5: Bonferroni-Holm using studentized residuals by a *************;
*************** robust scale estimate (BH-STRO) ***************************;
*************************************************************************;

*** Analysis of generalized lattice and generation of residuals *************;

*** M5r: Bonferroni-Holm using a robust scale estimate for
         studentization of residuals and incomplete blocks as random;
proc mixed data=dataset_name method=reml;
class rep        block        genotype        ;
model w3outl = rep genotype  /s residual  outp=file_with_residuals;
random rep*block/s;
ods output CovParms=cp_file;
* save variance components in a file called cp_file;
run;

*** M5f: Bonferroni-Holm using a robust scale estimate for
```

```
         studentization of residuals and incomplete blocks as fixed;
proc mixed data=dataset_name method=reml;
class rep        block        genotype        ;
model w3outl = rep genotype rep*block /s residual  outp=file_with_residuals;
ods output CovParms=cp_file;
* save error variance in a file called cp_file;
run;


*** Store the square root of the error mean square in a variable called sqrt_mse;
data _null_;
set cp_file;
if CovParm = 'Residual';
s_estimate = sqrt(estimate);
call symput('sqrt_mse',sqrt_mse);
run;

** Compute the MAD of the residuals and store in a file called Mad;
proc univariate data=preds noprint;
output out=Mad MAD=MAD n=n;
var resid;
run;

** Compute the re-scaled MAD in a variable called re_MAD;
data Mad;
set Mad;
re_MAD = MAD*1.4826;
run;

*** Store the re-scaled MAD in a variable called s_rob;
data _null_;
set Mad;
call symput('re_MAD',re_MAD);
call symput('n',n);
run; *--------------------------------;

*** Generate two variables: one with the absolute values of the residuals
    studentized using the re-scaled MAD (stud_rob_restest), the other with the
    corresponding p-value. The name of this second variable must be
    raw_p to let the multtest procedure recognize the values -------------;

data BHSTRO_file;
set file_with_residuals;
stud_rob_res = StudentResid * &sqrt_mse / &re_MAD;
stud_rob_restest = abs(stud_rob_res);
raw_p = 2 * (1 - probnorm(stud_rob_restest));
run;

*** Sort by p-value;
proc sort data =BHSTRO_file;
by raw_p ;
run;

*** Bonferroni-Holm test ;
proc multtest inpvalues=BHSTRO_file holm out=holm_BHSTRO_file ;
run;

*** Flag observations with significant test;
```

```
data holm_BHSTRO_file;
set holm_BHSTRO_file;
if stpbon_p < 0.05 then out_stud_rob = 1; else out_stud_rob = 0;
* Outliers are the observations flagged with 1;
if raw_p = . then out_stud_rob = 0;
* This step needs to be added to avoid flagging missing
  observations as outliers;
run;
```

```
##############################  CODES IN R ###################################

#############################################################################
### Basic model for methods using random incomplete blocks effects
# Analysis for original data

lmer.data=lmer(yield ~  gen + rep + (1|rep:block), data=  dataset_name )

#############################################################################
######## METHOD 1: PlabStat  ################################################
#############################################################################

# Print predictions
pred=cbind(predict(lmer.data))
summary(lmer.data)

# m.v.d. and MSE_Eff
diffs=lsmeans(lmer.data ,pairwise ~ gen, lsm.options(disable.pbkrtest=TRUE))
sum.diffs=summary(diffs)
vd=(cbind(sum.diffs[[2]][3]))**2
mvd=(mean(vd, na.omit=TRUE))*(length(unique(dataset_name$rep)))/2
#"(length(unique(dataset_name$rep)))" is the number of replicates
MSE_Eff=mvd**.5
# end

# Re-scaled MAD
resi=cbind(residuals(lmer.data,type = "response" ))
median=median(resi)
MAD=median((abs(resi - median)))
re_MAD=MAD*1.4826
# end

# Plabstat standardized residuals
res_PS = resi /MSE_Eff
# end


#  get DF
aov=summary(aov(yield ~  gen + rep + rep:block, data=  dataset_name))
DF=aov[[1]]$Df[4]
n=length(resi)
# end


premium=0.005
z=1- (premium*DF)/n
N=qnorm(z)
```

```
K=1.4+.85*N


C=K*(1-(K**2-2)/(4*DF))*(DF/n)**.5
C=ifelse(C<1.5,1.5,C)

#threshold
s_thresh = re_MAD*C*1.15


res_thresh=ifelse(MSE_Eff< s_thresh, res_thresh<- s_thresh,
                  res_thresh<-MSE_Eff)

absresi=cbind(abs(resi))

test=ifelse(absresi>res_thresh, "OUTLIER ", ".")

#Reduce file: Drop "NA's" to get same length as resi and test
dataset_name.1= subset(dataset_name, dataset_name$yield!="NA")

all=cbind(dataset_name.1, resi, test)
# all contains the data-labels, the residuals and the "outlier" labels.


# Take a look at the outliers
outliers_PS <- all[which(all$test!="."),]
#all[which(all$test=="OUTLIER "),]
##############################################################################

##############################################################################
### METHOD 2: Bonferroni-Holm using studentized residuals (BH-ST) ########
##############################################################################
#### This method uses the ASReml-R package
#### For code using lme4, please contact the authors
library(asreml)

attach(dataset_name)
asreml.data = asreml(fixed = yield ~  gen + rep, random = ~ rep:block,
                     data=dataset_name,
                     na.method.Y="omit", na.method.X="omit", maxiter=100,
                     workspace=1e9)
summary(asreml.data)
# Warning: SE are slightly different than the ones computed with lmer

# Produce externally studentized residuals
studresid.data <- asreml.data$resid/sd(asreml.data$resid, na.rm=TRUE)

# Install package "multtest" from the bioconductor
source("http://bioconductor.org/biocLite.R")
biocLite("multtest")
library(multtest)

# Calculate adjusted p-values
rawp.BHStud = 2 * (1 - pnorm(abs(studresid.data)))


#Combine the dataset, the residuals and the adjusted p-values
rawp.BHStud.all <- cbind(dataset_name, studresid.data, rawp.BHStud)
```

```
#Produce a Bonferroni-Holm tests for the adjusted p-values
#The output is a list
test.BHStud<-mt.rawp2adjp(rawp.BHStud,proc=c("Holm"))


#Create vectors/matrices out of the list of the BH tests
adjp = cbind(test.BHStud[[1]][,1])
bholm = cbind(test.BHStud[[1]][,2])
index = cbind(test.BHStud[[2]])

# Condition to flag outliers according to the BH test
out_flag = ifelse(bholm<0.05, "OUTLIER ", ".")

#Create a matrix with all the output of the BH test
BHStud_test = cbind(adjp,bholm,index,out_flag)

#Order the file by index
BHStud_test2 = BHStud_test[order(index),]

#Label colums
names = c("rawp","bholm","index","out_flag")
colnames(BHStud_test2) <- names

#Create a final file, with the data and the test and the labels
#for the outliers
total.m2_data <- cbind(rawp.BHStud.all, BHStud_test2)

# Take a look at the outliers
outliers_BH <- total.m2_data[which(total.m2_data$out_flag!="."),]
##########################################################################

##########################################################################
######## METHOD 3: Studentized residual razor (SRR) ######################
##########################################################################

#### This method uses the ASReml-R package
#### For code using lme4, please contact the authors
library(asreml)

attach(dataset_name)
asreml.data = asreml(fixed = yield  ~ gen + rep, random = ~ rep:block,
                     data=dataset_name,
                     na.method.Y="omit", na.method.X="omit", maxiter=100,
                     workspace=1e9)
summary(asreml.data)
# Warning: SE are different than the ones computed with lmer

# Produce externally studentized residuals
studresid.data <- asreml.data$resid/sd(asreml.data$resid, na.rm=TRUE)

# number of outliers detected

# Rule for flagging residuals
test.SRR=ifelse(abs(studresid.data) > 2.8, "OUTLIER ", ".")

# File containing all the dataset info + student residuals + test output
```

```
all.data.SRR=cbind dataset_name, studresid.data, test.SRR)

# Take a look at the outliers
outliers_SRR <- all.data[which(test.SRR=="OUTLIER "),]
##############################################################################

##############################################################################
### METHOD 4: Bonferroni-Holm using re-scaled MAD for ####################
################## standardizing residuals (BH-MADR) #####################
##############################################################################

## Basic model
# Analysis for original data
lmer.data=lmer(yield ~  gen + rep + (1|rep:block), data=dataset_name)

# re-scaled MAD
resi=cbind(residuals(lmer.data, type = "response" ))
median=median(resi)
MAD=median((abs(resi - median)))
re_MAD=MAD*1.4826
# end

# MAD standardized residuals
res_MAD = resi /re_MAD
# end

# Install package "multtest" from the bioconductor
source("http://bioconductor.org/biocLite.R")
biocLite("multtest")
library(multtest)

# Calculate adjusted p-values
rawp = 2 * (1 - pnorm(abs(res_MAD)))


#Reduce the dataset in case of missing values.
#These are not carried over to the rawp matrix
dataset_name.1= subset(dataset_name, dataset_name$yield!="NA")
#Warning: Change variable that is under analysis, in this case yield


#Combine the dataset, the residuals and the adjusted p-values
rawp2 <- cbind(dataset_name.1, resi,  res_MAD, rawp)


#Produce a Bonferroni-Holm tests for the adjusted p-values
#The output is a list
res2<-mt.rawp2adjp(rawp,proc=c("Holm"))


#Create vectors/matrices out of the list of the BH tests
adjp = cbind(res2[[1]][,1])
bholm = cbind(res2[[1]][,2])
index = cbind(res2[[2]])

# Condition to flag outliers according to the BH test
out_flag = ifelse(bholm<0.05, "OUTLIER ", ".")
```

```
#Create a matrix with all the output of the BH test
bholm_test = cbind(adjp,bholm,index,out_flag)

#Order the file by index
bholm_test2 = bholm_test[order(index),]

#Label colums
names = c("rawp","bholm","index","out_flag")
colnames(bholm_test2) <- names

#Create a final file, with the data and the test and the labels for the outliers
total.m4_data <- cbind(rawp2,bholm_test2)

# Take a look at the outliers
outliers_BHMAD <- total.m4_data[which(total.m4_data$out_flag!="."),]
##########################################################################
```

A.3 Additional information - Online resource 3

A.3.1 Comparison of methods: *Premium* vs. α_B vs. t_{SRR}

The PlabStat threshold includes the robust standard deviation estimate s^r. Raw residuals can be standardized by this robust estimate, so that the threshold can be re-expressed as only CP, with C and P two constants defined in Materials and Methods Section (See Method M1). The same standardized residuals can be used for the classical Bonferroni test.

To obtain the α/n of the classical Bonferroni test that corresponds to a given *premium*, we may equate the threshold CP, corresponding to the PlabStat $premium$, to the $(1-\alpha_B/2)$-quantile of the standard normal distribution, corresponding to the positive bound of the Bonferroni threshold of a two-sided test, where $\alpha_B = \alpha/n$ is the significance level for an individual test:

$$\Phi^{-1}\left(1-\frac{\alpha}{2n}\right) = CP \quad \text{or} \quad \Phi^{-1}\left(\frac{\alpha}{2n}\right) = -CP \tag{A.1}$$

where Φ^{-1} stands for the inverted cumulative distribution of the standard normal.

Let

$$\alpha_B = \frac{\alpha}{n},$$

then we solve for α_B the equation

$$\Phi^{-1}\left(1-\frac{\alpha_B}{2}\right) = CP,$$

where C is a function of the premium.

By Eq. (2.5) we have

$$\Phi^{-1}\left(1-\frac{\alpha_B}{2}\right)=\left[K\left\{1-\frac{K^2-2}{4df_e}\right\}\sqrt{\frac{df_e}{n}}\right]P$$

By Eq. (2.6)

$$\Phi^{-1}\left(1-\frac{\alpha_B}{2}\right)=\left[(1.4+0.85N)\left\{1-\frac{(1.4+0.85N)^2-2}{4df_e}\right\}\sqrt{\frac{df_e}{n}}\right]P$$

where, by Eq. (2.7)

$$\begin{aligned}\Phi(-N)&=\frac{premium*df_e}{n}\\ \Phi^{-1}\left[\Phi(-N)\right]&=\Phi^{-1}\left[premium*\frac{df_e}{n}\right]\\ N&=-\Phi^{-1}\left[premium*\frac{df_e}{n}\right].\end{aligned}$$

Then, from

$$\Phi^{-1}\left(1-\frac{\alpha_B}{2}\right)=$$
$$\left[\left\{1.4-0.85\cdot\Phi^{-1}\left(premium*\frac{df_e}{n}\right)\right\}\left\{1-\frac{\left[1.4-0.85\cdot\Phi^{-1}\left(premium*\frac{df_e}{n}\right)\right]^2-2}{4df_e}\right\}\sqrt{\frac{df_e}{n}}\right]P \tag{A.2}$$

it is possible to see the complexity of the relationship between α_B and $premium$.

If n is large and df_e close to n, then $df_e/n\sim 1$ and the expression above simplifies to

$$\Phi^{-1}\left(1-\frac{\alpha_B}{2}\right)=\left[\left\{1.4-0.85\cdot\Phi^{-1}\left(premium\right)\right\}\left\{1-\frac{\left[1.4-0.85\cdot\Phi^{-1}\left(premium\right)\right]^2-2}{4df_e}\right\}\right]P$$

Further, if df_e is large $\sim\infty$, then

$$\Phi^{-1}\left(1-\frac{\alpha_B}{2}\right) = \left[1.4 - 0.85 \cdot \Phi^{-1}\left(premium\right)\right] P$$
$$\Phi\left[\Phi^{-1}\left(1-\frac{\alpha_B}{2}\right)\right] = \Phi\left\{\left[1.4 - 0.85 \cdot \Phi^{-1}\left(premium\right)\right] P\right\}$$
$$1-\frac{\alpha_B}{2} = \Phi\left\{\left[1.4 - 0.85 \cdot \Phi^{-1}\left(premium\right)\right] P\right\}$$

$$\alpha_B = 2\left(1 - \Phi\left\{\left[1.4 - 0.85 \cdot \Phi^{-1}\left(premium\right)\right] P\right\}\right) \tag{A.3}$$

and thus a small *premium* lead to very small α_B, though somewhat higher than a fixed $\alpha_B = 0.05/n$, with n large (Fig. A.1a).

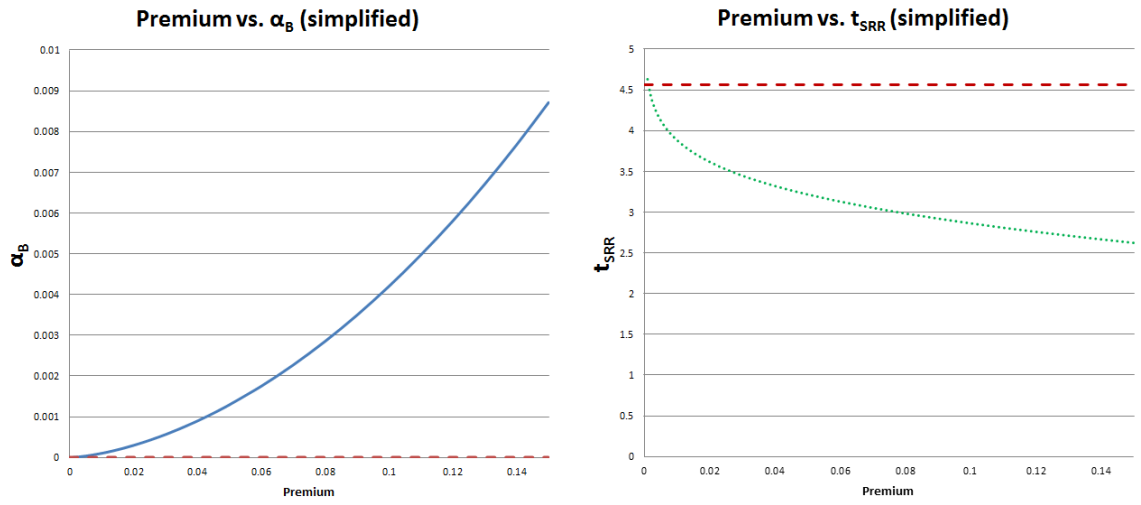
The fundamental outcome of this comparison ($premium$ vs. α_B) is that the PlabStat method, at a fixed $premium$ of 0.005, may flag more outlying observations than the classical Bonferroni test with $\alpha_B = 0.05/n$. As an example, let a trial have $n = 100$ and $df_e = 60$, keeping $premium$ fixed at 0.005, we would need $\alpha_B \cong 0.00157$, that is, an $\alpha \cong 0.157$, to get the same thresholds (using equation Eq. A.2). For trials with large n (as is the case of the rye example), using $premium = 0.005$ flags more outliers than the classical Bonferroni test with $\alpha = 0.05$. This effect is clear because more observations can fall within the area between the two thresholds, increasing the number of flagged outliers for PlabStat procedure. A big difference is not anticipated in case of small trials, where fewer outliers are expected (otherwise it would be better to resort to transformations or a different type of analysis), and when residuals clearly stand out from the main data cloud, having low p-values that are prone to fall in the rejection area of both methods (Figs. 1, S2, S3 and S4).

We can identify the α_B that corresponds to the SRR threshold t_{SRR} from:

$$\Phi^{-1}\left(1-\frac{\alpha_B}{2}\right) = t_{SRR} \quad \Leftrightarrow \quad \alpha_B = 2[1 - \Phi(t_{SRR})]$$

and plot the SRR threshold varying across the *premium* grid (Fig. A.1b).

Under simplification in Fig. A.1, with $df_e/n \sim 1$ and $df_e \sim \infty$, if *premium*$= 0.005$ (as in PlabStat), α_B is very low and t_{SRR} is very high. An exemplary SRR threshold, say $t_{SRR} = 2.8$, would be reachead by using a *premium* of 0.112. What we grasp from Fig. A.1b is that the classical Bonferroni threshold in comparison to SRR would flag a lot less outliers and as long as the *premium* increases, the corresponding SRR threshold reduces. Thus, for a *premium* $= 0.005$ a high SRR threshold is needed to select the same outliers. In practice, however, this behaviour of SRR flagging more outliers than PlabStat is not observed. What we see is that PlabStat with *premium* $= 0.005$ declares more (or the same) outliers than SRR (which may be false positives), but this response cannot be depicted as it is in Fig. A.1 since we are using a simplification and we are not considering strictly the difference attributed to the standardization of the residuals.



(a) (b)

Figure A.1: Correspondence of (a) $\alpha_B = \frac{\alpha}{n}$ values of the classical Bonferroni test and (b) threshold values t_{SRR} of the Studentized residual razor (SRR) for a grid of *premiums* of the PlabStat procedure assuming $df_e/n \sim 1$ and $df_e \sim \infty$. The solid blue line represents the PlabStat threshold varying according to the *premium*, the dashed red line represents the classical Bonferroni threshold at (a) $\alpha_B = \frac{0.05}{n}$ with n large, and (b) $\Phi^{-1}\left(1 - \frac{\alpha_B}{2}\right)$, and the dotted green line shows the Studentized residual razor (SRR) threshold t_{SRR} varying according to the *premium*.

A.3.2 Threshold and re-scaled MAD comparison

Table A.5: Comparison of re-scaled MAD (s^r) and thresholds computed in the datasets with missing observations using PlabStat (PS), SAS-REML (REML) and SAS-ANOVA (ANOVA). Threshold is $s^r CP$, where s^r is the re-scaled MAD, C is a constant depending on degrees of freedom of the error df_e and total number of observations n, and $P = 1.15$.

		Threshold			s^r		
Example	# outliers	PS	REML	ANOVA	PS	REML	ANOVA
1.2	5	0.161	0.155	0.154	0.395	0.379	0.376
2.2	2	1.193	1.192	1.184	0.378	0.381	0.381
3.2	0	3.243	3.503	3.515	1.681	1.593	1.608
4.2	0	0.016	0.000	–§	0.000	0.000	–§

§ Not computed

A.3.3 Residual plots for the methods and the examples

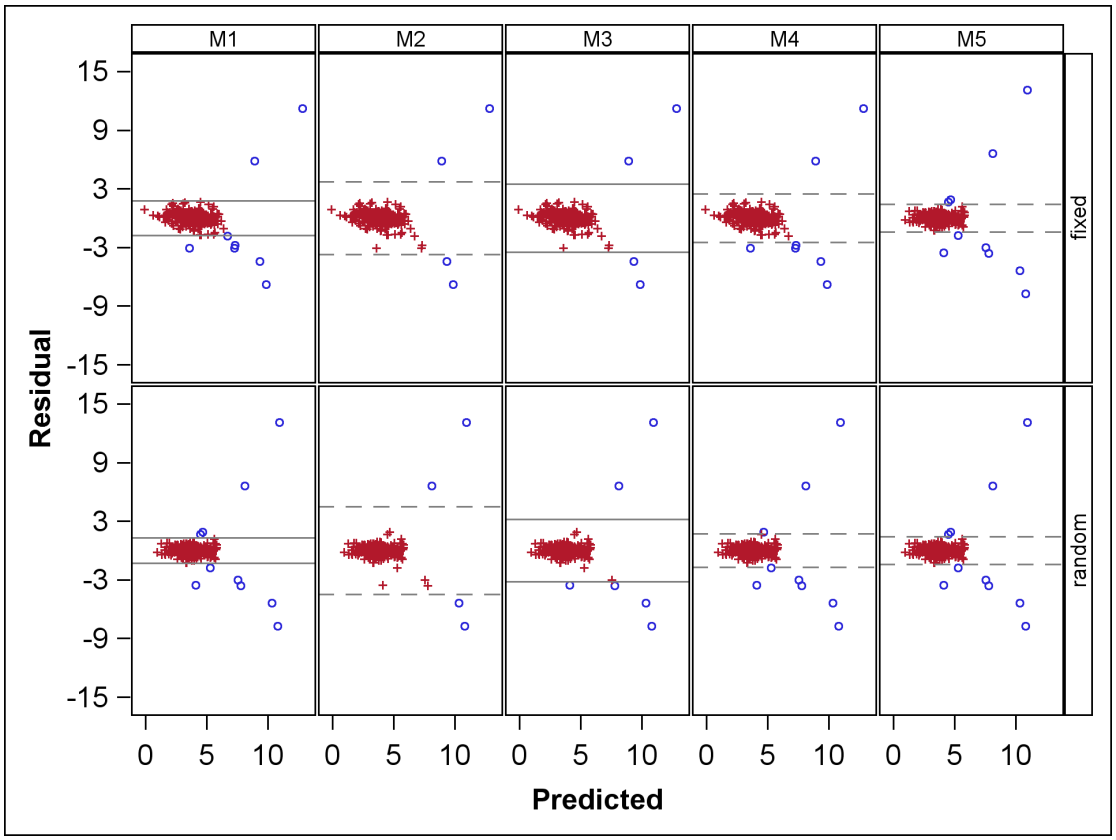


Figure A.2: Scatter plots of raw residuals vs. predictions for the triple lattice design with 3 outlying observations (Example 2.3) using PlabStat outlier detection method (M1), Bonferroni-Holm test using studentized residuals (M2), Studentized residual razor (M3), Bonferroni-Holm test using re-scaled MAD to standardize residuals (M4), and Bonferroni-Holm test using the robust studentized residuals (M5). In the first row, methods that used fixed incomplete block effects and in the second row methods that used random incomplete block effects. Solid reference lines are used for methods with fixed thresholds and dashed reference lines for methods with varying thresholds representing the threshold calculated for the largest residual. Flagged outliers are indicated with an empty circle and non-suspicious observations with a cross.

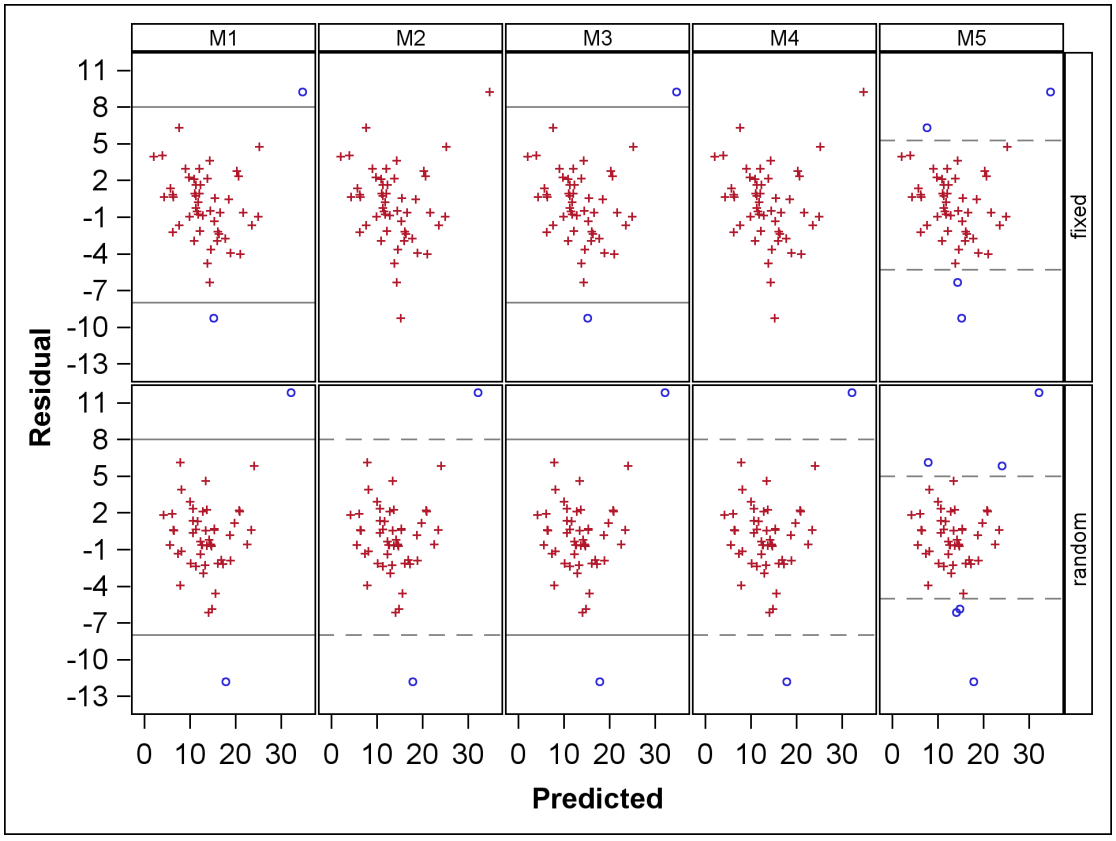


Figure A.3: Scatter plots of raw residuals vs. predictions for the square lattice design with 3 outlying observations (Example 3.3) using PlabStat outlier detection method (M1), Bonferroni-Holm test using studentized residuals (M2), Studentized residual razor (M3), Bonferroni-Holm test using re-scaled MAD to standardize residuals (M4), and Bonferroni-Holm test using the robust studentized residuals (M5). In the first row, methods that used fixed incomplete block effects and in the second row methods that used random incomplete block effects. Solid reference lines are used for methods with fixed thresholds and dashed reference lines for methods with varying thresholds representing the threshold calculated for the largest residual. Flagged outliers are indicated with an empty circle and non-suspicious observations with a cross.

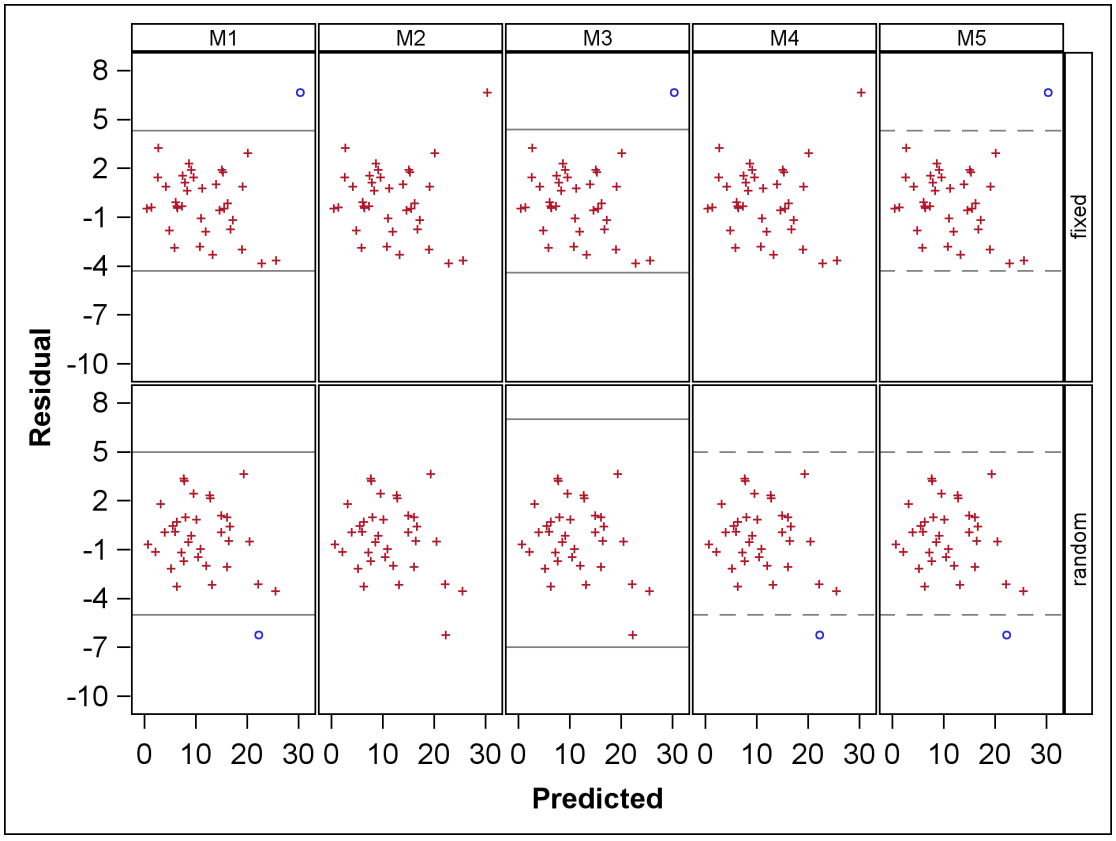


Figure A.4: Scatter plots of raw residuals vs. predictions for the rectangular lattice design with 3 outlying observations (Example 4.3) using PlabStat outlier detection method (M1), Bonferroni-Holm test using studentized residuals (M2), Studentized residual razor (M3), Bonferroni-Holm test using re-scaled MAD to standardize residuals (M4), and Bonferroni-Holm test using the robust studentized residuals (M5). In the first row, methods that used fixed incomplete block effects and in the second row methods that used random incomplete block effects. Solid reference lines are used for methods with fixed thresholds and dashed reference lines for methods with varying thresholds representing the threshold calculated for the largest residual. Flagged outliers are indicated with an empty circle and non-suspicious observations with a cross.

A.4 ROC curves (additionals)

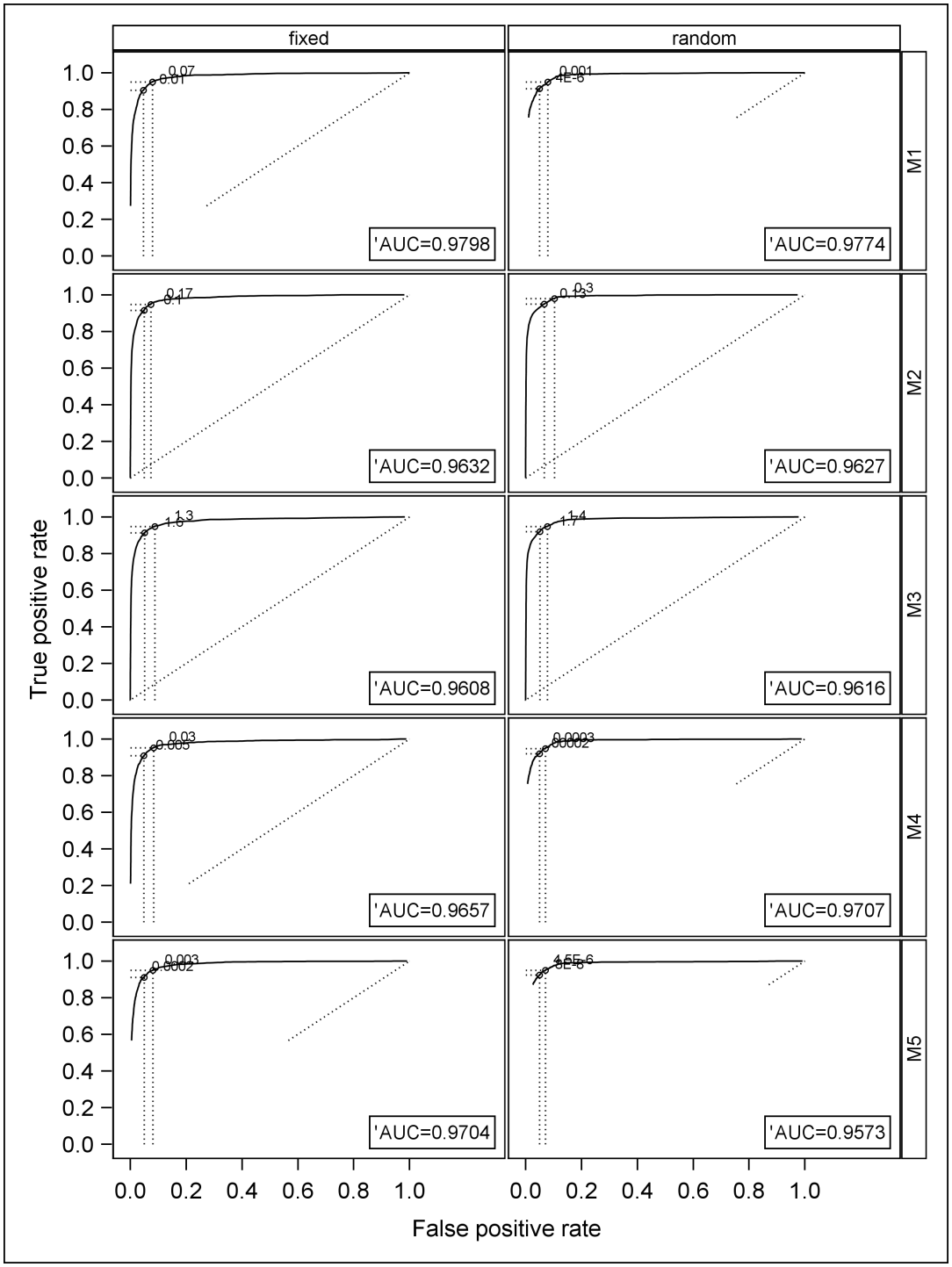


Figure A.5: ROC curves of all methods using fixed (first column) and random (second column) incomplete block effects under a scenario with 5% contamination and 7 deviation units from the mean (Scenario 2). Methods used were PlabStat (M1) outlier detection method, Bonferroni-Holm test using studentized residuals (M2), Studentized residual razor (M3), Bonferroni-Holm test using re-scaled MAD to standardize residuals (M4) and Bonferroni-Holm test using robust studentized residuals (M5).

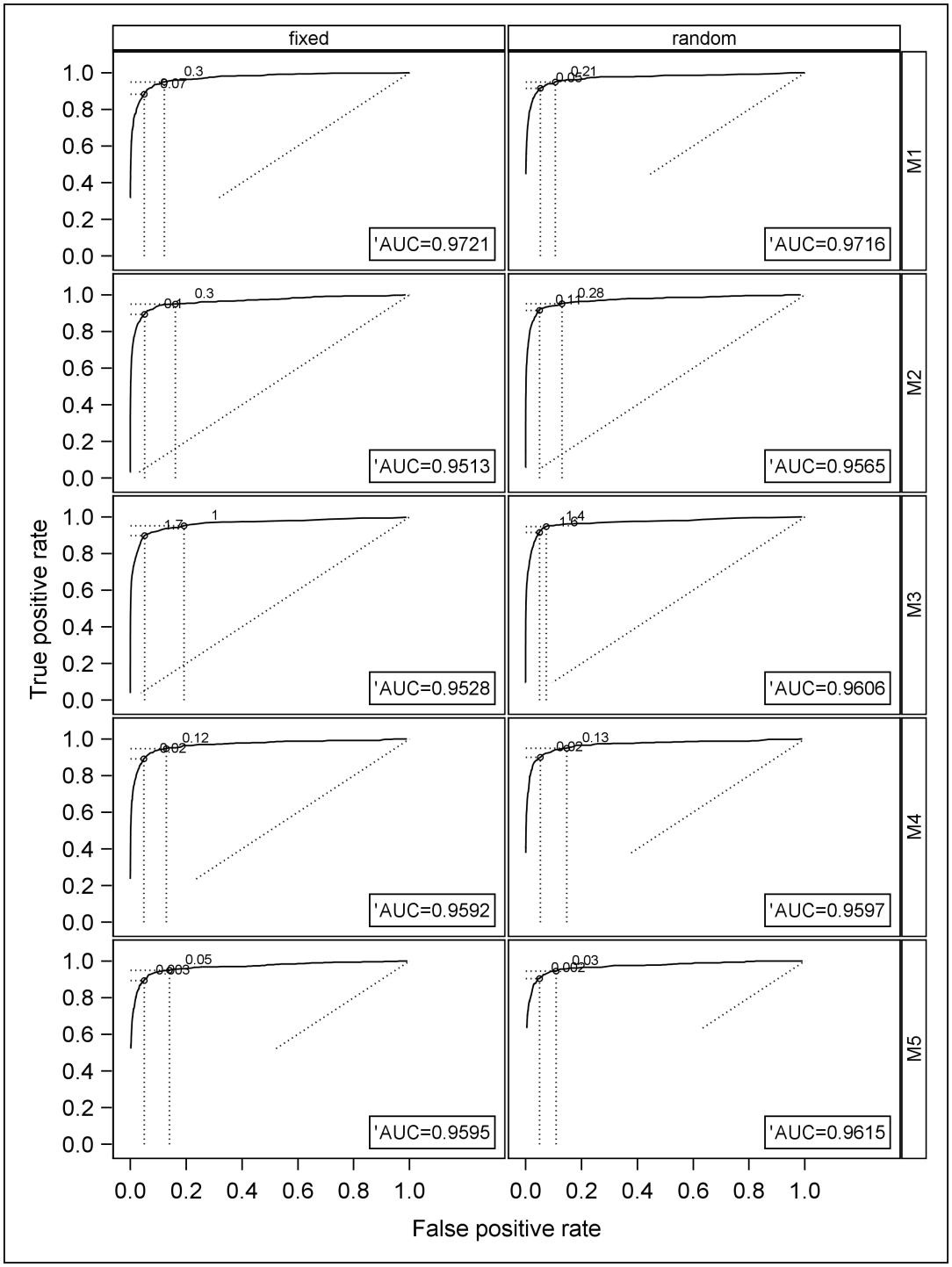


Figure A.6: ROC curves of all methods using fixed (first column) and random (second column) incomplete block effects under a scenario with 2% contamination and 4 deviation units from the mean (Scenario 1). Methods used were PlabStat (M1) outlier detection method, Bonferroni-Holm test using studentized residuals (M2), Studentized residual razor (M3), Bonferroni-Holm test using re-scaled MAD to standardize residuals (M4) and Bonferroni-Holm test using robust studentized residuals (M5).

Table A.6: Area under the curve (AUC) for low (Scenario 1), medium (Scenario 2) and high (Scenario 3) contamination scenarios for the methods: PlabStat (M1f and M1r) with fixed and random block effects, Bonferroni-Holm using studentized residuals (M2f and M2r) with fixed and random block effects, studentized residual razor (M3f and M3r) with fixed and random block effects, Bonferroni-Holm using re-scaled MAD (M4f and M4r) with fixed and random block effects, and Bonferroni-Holm using robust studentized residuals (M5f and M5r) with fixed and random block effects.

Method	Scenario 1	Scenario 2	Scenario 3
M1f	0.9721	0.9798	0.9524
M1r	0.9716	0.9794	0.9543
M2f	0.9513	0.9632	0.9385
M2r	0.9565	0.9627	0.9308
M3f	0.9528	0.9608	0.9413
M3r	0.9606	0.9616	0.9369
M4f	0.9592	0.9657	0.9356
M4r	0.9597	0.9707	0.9493
M5f	0.9595	0.9704	0.9429
M5r	0.9615	0.9573	0.9285

A.4.1 TPR and FPR with fixed rates

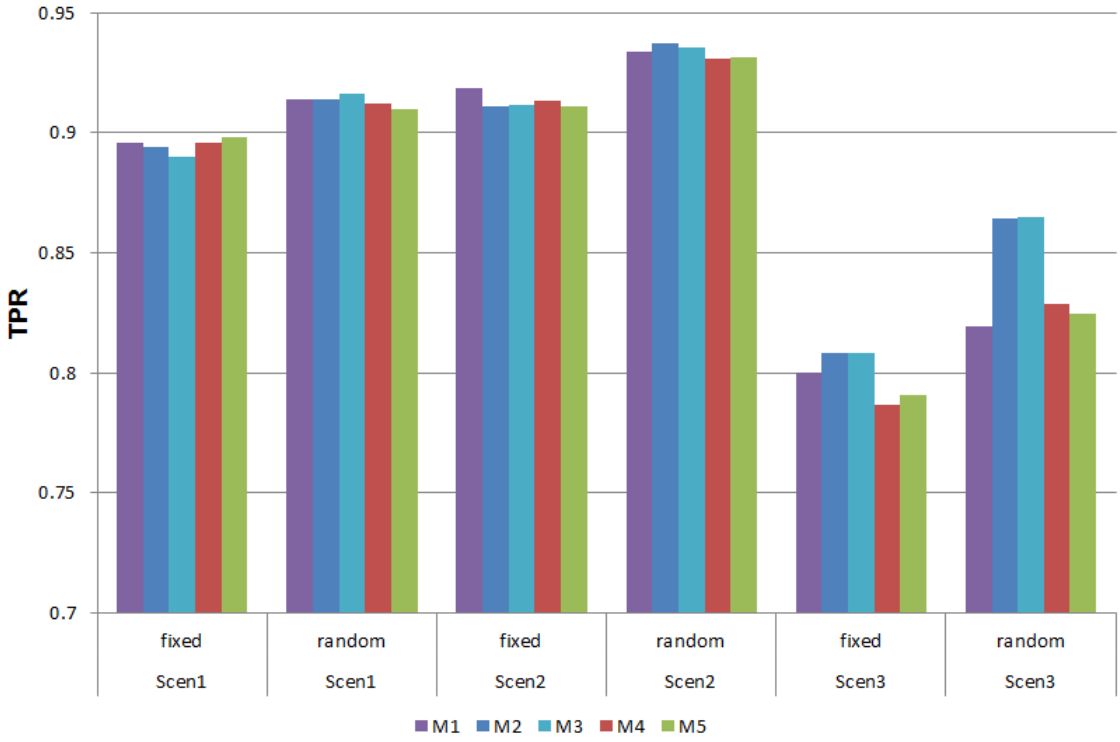


Figure A.7: Expected true positive rate (TPR) fixing FPR=5% of five outlier detection methods across low (Scenario 1), medium (Scenario 2) and high (Scenario 3) contamination scenarios of simulated outliers using a triple lattice experiment and assuming incomplete blocks as random and fixed for methods: PlabStat (M1) outlier detection method, Bonferroni-Holm test using studentized residuals (M2), Studentized residual razor (M3), Bonferroni-Holm test using re-scaled MAD to standardize residuals (M4) and Bonferroni-Holm test using robust studentized residuals (M5).

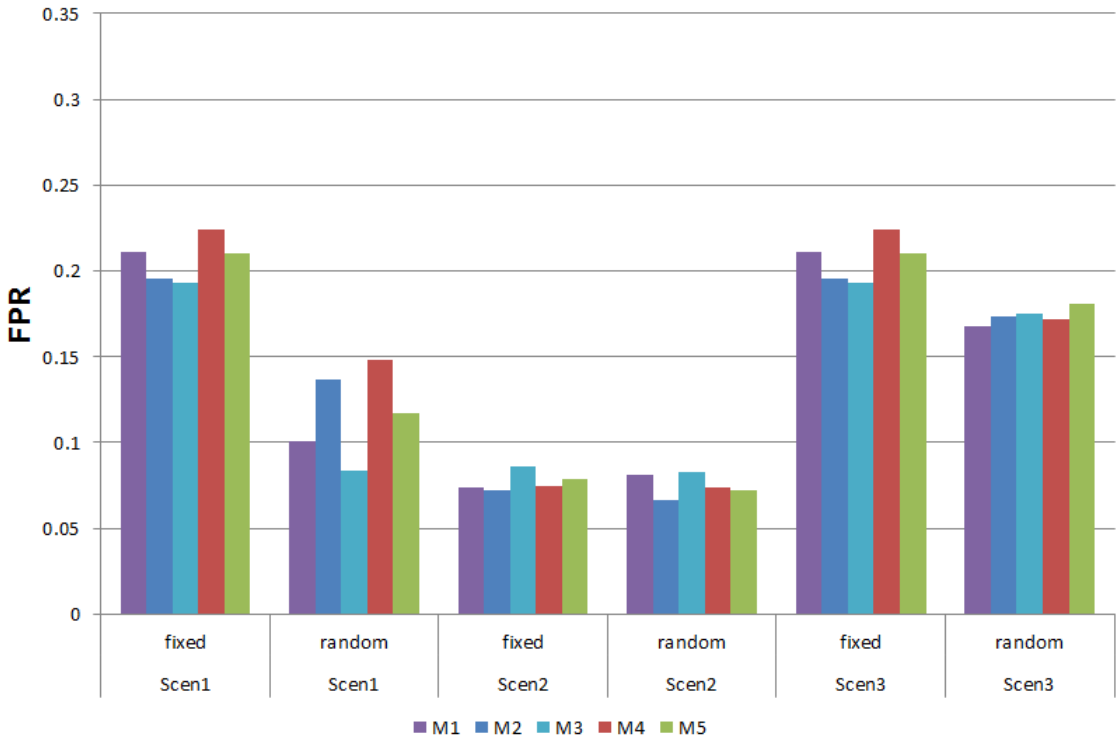


Figure A.8: Expected false positive rate (FPR) fixing TPR=95% of five outlier detection methods across low (Scenario 1), medium (Scenario 2) and high (Scenario 3) contamination scenarios of simulated outliers using a triple lattice experiment and assuming incomplete blocks as random and fixed for method: : PlabStat (M1) outlier detection method, Bonferroni-Holm test using studentized residuals (M2), Studentized residual razor (M3), Bonferroni-Holm test using re-scaled MAD to standardize residuals (M4) and Bonferroni-Holm test using robust studentized residuals (M5).

A.5 Heatmap of an exemplary rye trial

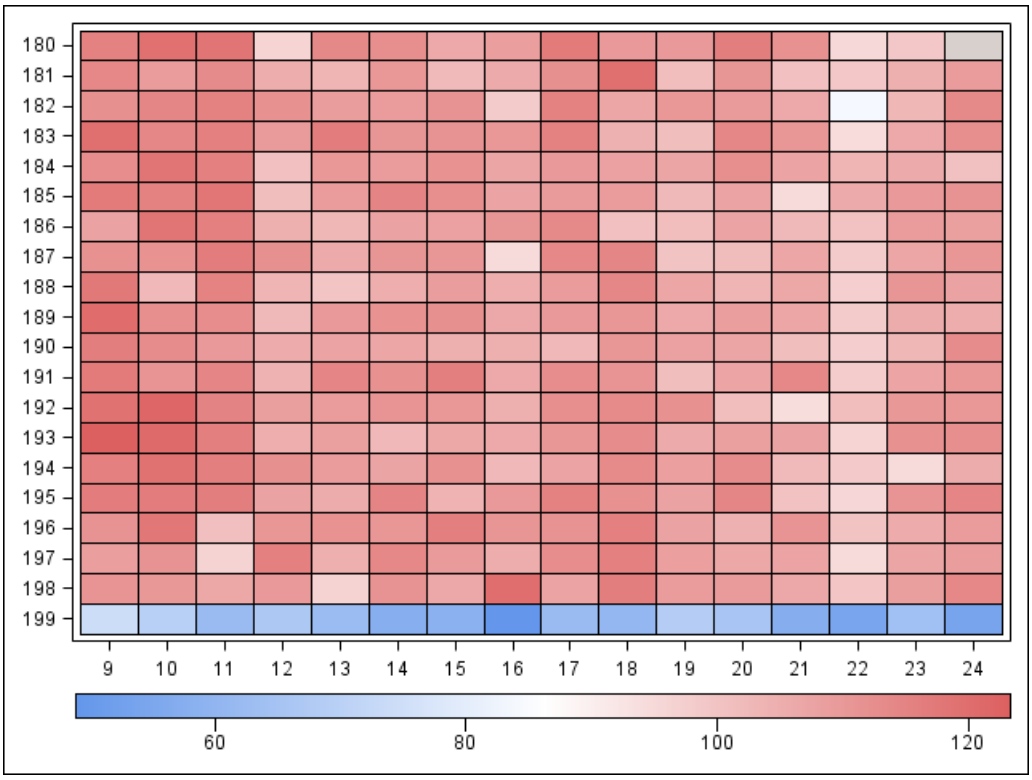


Figure A.9: Heatmap of grain dry matter yield (dt/ha) of a rye trial affected by a herbicide drift. Abscissas represent the rows and ordinates the columns on the field layout.

Appendix B

Supplementary material of Chapter 3

B.1 SAS codes (version 9.3) used to implement first stage of phenotypic analysis referred in Table 3.4

```
/*************************************************************************
********************** Appendix A************************************
********** Some SAS code to fit the models of the first stage *********
**************************referred in Table 4**************************
*************************************************************************/

**** Model 1. Baseline model ------------------------------;

proc mixed data=data1 ;
ods output FitStatistics=fits_M1 lsmeans= adjmeans_M1 covparms=cp_M1;
by year loc;
class year loc genotype tester trial rep block  row column ;
model y1=genotype*tester/ddfm=residual solution ;
random int rep rep*block / sub=trial;
lsmeans genotype*tester /cov;
run;


**** Model 2. Baseline + row + column -----------;

proc mixed data=data1;
ods output FitStatistics=fits_M2 lsmeans= adjmeans_M2 covparms=cp_M2;
by year loc;
class year loc genotype tester trial rep block row column ;
model y1=genotype*tester/ddfm=residual solution ;
random int rep rep*block / sub=trial;
random row / sub=trial*rep;
random column /sub=trial*rep;
lsmeans genotype*tester /cov;
run;


**** Model 3. Baseline --------- AR(1) ;

proc mixed data=data1;
ods output FitStatistics=fits_M3 lsmeans= adjmeans_M3 covparms=cp_M3;
by year loc;
```

```
class year loc genotype tester trial rep block row column plot ;
model y1=genotype*tester/ddfm=residual solution ;
random int rep rep*block / sub=trial;
repeated plot / sub=trial*rep*block type=AR(1);
lsmeans genotype*tester /cov;
parms 1 1 1 0.1 1 / lowerb=  . , . , . , 1e-8 , . ;
run;


*** Model 4. Baseline --------- LV + nugget ;

proc mixed data=data1;
ods output FitStatistics=fits_M4 lsmeans= adjmeans_M4 covparms=cp_M4;
by year loc;
class year loc genotype tester trial rep block row column plot ;
model y1=genotype*tester/ddfm=residual solution ;
random int rep rep*block / sub=trial;
random plot / sub=trial*rep*block type=LIN(1)  ldata=LV_matrix;
lsmeans genotype*tester /cov;
parms 1 1 1 0.1 1 / lowerb=. , . , . , 1e-8, . ;
run;


*** Model 5. Baseline --------- AR(1) x AR(1);

proc mixed data=data1;
ods output FitStatistics=fits_M5 lsmeans= adjmeans_M5 covparms=cp_M5;
by year loc;
class year loc genotype tester trial rep block row column plot ;
model y1=genotype*tester/ddfm=residual solution ;
random int rep rep*block / sub=trial;
repeated row*column / sub=trial*rep type=SP(POWA) (ro co);
lsmeans genotype*tester /cov;
parms 1 1 1 0.1 0.1 1 / lowerb=  . , . , . , 1e-8 , 1e-8 , . ;
run;


*** Model 6. Baseline + row + column --------- AR(1) x AR(1);

proc mixed data=data1;
ods output FitStatistics=fits_M6 lsmeans= adjmeans_M6 covparms=cp_M6;
by year loc;
class year loc genotype tester trial rep block row column plot ;
model y1=genotype*tester/ddfm=residual solution ;
random int rep rep*block / sub=trial;
random row / sub=trial*rep;
random column/ sub=trial*rep;
repeated row*column / sub=trial*rep type=SP(POWA) (ro co);
* ro and co are numerical variables with row and column coordinates;
lsmeans genotype*tester /cov;
parms 1 1 1 1 1 0.1 0.1 1 / lowerb= . , . , . , . , . , 1e-8 , 1e-8 , . ;
run;


*** Model 7. Baseline  --------- AR(1) + nugget ;

proc mixed data=data1;
ods output FitStatistics=fits_M7 lsmeans= adjmeans_M7 covparms=cp_M7;
```

```
by year loc;
class year loc genotype tester trial rep block row column plot ;
model y1=genotype*tester/ddfm=residual solution ;
random int rep rep*block / sub=trial;
repeated plot / sub=trial*rep*block local type=AR(1);
lsmeans genotype*tester /cov;
parms (0 1) (1) (1) (1) (0.1) (1) / lowerb= . , . , . , . , 1e-8 , . ;
run;


*** Model 8. Baseline  --------- AR(1) x AR(1) + nugget;

proc mixed data=data1;
ods output FitStatistics=fits_M8 lsmeans= adjmeans_M8 covparms=cp_M8;
by year loc;
class year loc genotype tester trial rep block row column plot ;
model y1=genotype*tester/ddfm=residual solution ;
random int rep rep*block / sub=trial;
random row / sub=trial*rep;
repeated row*column / sub=trial*rep local type=SP(POWA) (ro co);
* ro and co are numerical variables with row and column coordinates;
lsmeans genotype*tester /cov;
parms 1 1 1 1 0.1 0.1 1 / lowerb=  . , . , . , . , 1e-8 , 1e-8 , . ;
run;


*** Model 9. Baseline + row + column --------- AR(1) x AR(1) + nugget;

proc mixed data=data1;
ods output FitStatistics=fits_M8 lsmeans= adjmeans_M8 covparms=cp_M8;
by year loc;
class year loc genotype tester trial rep block row column plot ;
model y1=genotype*tester/ddfm=residual solution ;
random int rep rep*block / sub=trial;
random row / sub=trial*rep;
random column/ sub=trial*rep;
repeated row*column / sub=trial*rep local type=SP(POWA) (ro co);
* ro and co are numerical variables with row and column coordinates;
lsmeans genotype*tester /cov;
parms 1 1 1 1 1 1 0.1 0.1 1 / lowerb= .,.,.,.,.,1e-8,1e-8,.;*hold=6,7;
run;
```

B.2 Analysis of bias of genomic prediction

One reviewer pointed out that genomic prediction is biased and suggested to investigate the potential bias by regressing observations on the predictions from the same cross validation procedure. Other authors (Le Roy et al., 2012; Wang et al., 2012) have used the method to compare genomic prediction models that make use of different penalization tools, such as RR-BLUP, Bayes, Lasso or any other machine learning method. Although we use here only RR-BLUP, we computed the bias of each error spatial model.

Results Below the results for bias of the models and the mixed datasets for both sampling strategies (WC and AC).

Table B.1: Bias (regression coefficient between observations and predictions) for 9 spatial and non-spatial models (M1, $\cdots$, M9) and mixed datasets using the best locations given AIC (Mix1) and ρ-GP-CV (Mix2). Comparisons were performed using the absolute deviation of the regression coefficient from one. Same letters within rows indicate no significant differences (α = 5%) according to a paired t-test. Sampling strategies were: Within crosses (WC) and across crosses (AC).

	M1	M2	M3	M4	M5	M6	M7	M8	M9	Mix1	Mix2
WC	0.958	0.959	0.944	0.916	0.94	0.939	0.947	0.95	0.933	0.933	1.138
	ab	a	ab	c	abc	abc	ab	ab	bc	bc	d
AC	0.553	0.556	0.546	0.545	0.551	0.549	0.544	0.553	0.548	0.546	0.539
	ab	a	def	ef	bc	cd	f	bc	de	def	g

For both strategies, it turned out that the less biased model was M2, confirming the conclusions throughout the paper that the model with the simplest row-column adjustment had in overall the best results.

Appendix C

Supplementary material of Chapter 4

C.1 Complete selection breeding program

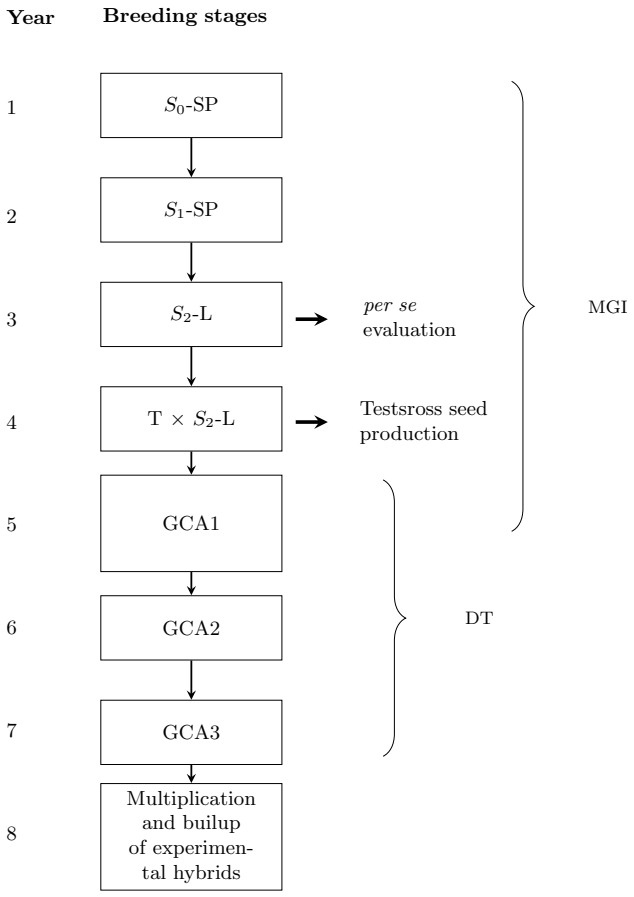


Figure C.1: **Flow diagram of a complete selection cycle of the pollen parent pool.** S_x = selfing generation x, SP = single plant, L = line, T = tester, GCAX = general combining ability X trial, MGI = minimum generation interval, DT = datasets used.

C.2 Diagrams of prediction scenarios

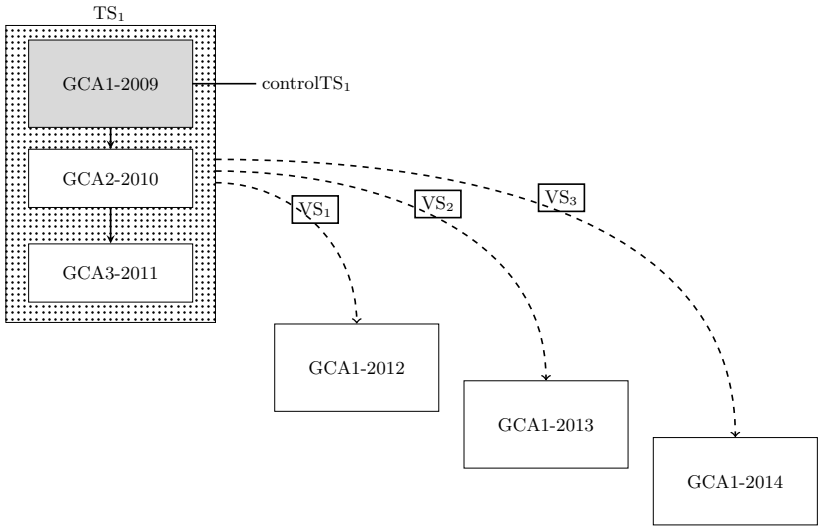


Figure C.2: **Diagram of the first scenario.** TS_1 with dotted background and control set ($controlTS_1$) filled in gray. Arrows represent the prediction goals VS_1, VS_2 and VS_3.

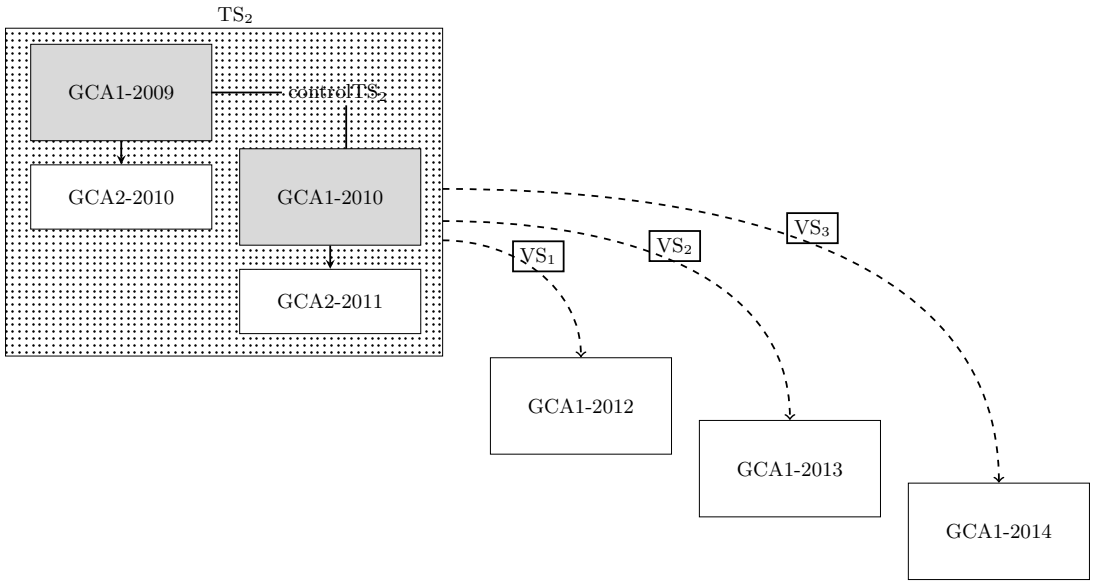


Figure C.3: **Diagram of the second scenario.** TS_2 with dotted background and control set ($controlTS_2$) filled in gray. Arrows represent the prediction goals VS_1, VS_2 and VS_3.

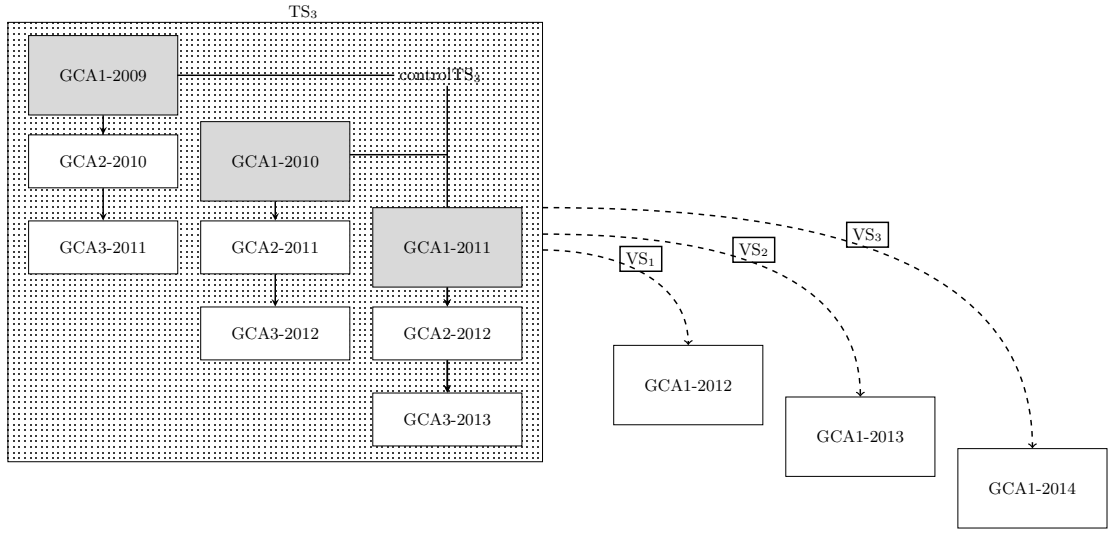


Figure C.4: **Diagram of the third scenario.** TS_3 with dotted background and control set (controlTS_3) filled in gray. Arrows represent the prediction goals VS_1, VS_2 and VS_3.

C.3 Number of locations and Location-year combinations

Table C.1: Number of locations (L) and year-location combinations (YL) for the German (GER) and the Polish (PL) datasets. Last column shows the ratio between YL and L.

Program	TS	VS	L	YL	YL/L
GER	TS_1	VS_1	25	28	1.1
GER	TS_1	VS_2	24	31	1.3
GER	TS_1	VS_3	23	32	1.4
PL	TS_1	VS_1	23	23	1.0
PL	TS_1	VS_2	23	25	1.1
PL	TS_1	VS_3	22	26	1.2
GER	TS_2	VS_1	22	31	1.4
GER	TS_2	VS_2	21	29	1.4
GER	TS_2	VS_3	20	30	1.5
PL	TS_2	VS_1	21	25	1.2
PL	TS_2	VS_2	20	23	1.2
PL	TS_2	VS_3	19	24	1.3
GER	TS_3	VS_1	30	55	1.8
GER	TS_3	VS_2	29	56	1.9
GER	TS_3	VS_3	29	65	2.2
PL	TS_3	VS_1	29	49	1.7
PL	TS_3	VS_2	28	50	1.8
PL	TS_3	VS_3	28	59	2.1

TS_1: GCA1-2009 + GCA2-2010 + GCA3-2011, TS_2: GCA1-2009 + GCA2-2010 + GCA1-2010 + GCA2-2011, TS_3: GCA1-2009 + GCA2-2010 + GCA3-2011 + GCA1-2010 + GCA2-2011 + GCA3-2012 + GCA1-2011 + GCA2-2012 + GCA3-2013, VS_1: GCA1-2012, VS_2: GCA1-2013, VS_3: GCA1-2014.

C.4 Asymptotic correlation and variance-covariance matrices

Table C.2: Asymptotic correlation (lower diagonal) and covariance (upper diagonal) matrix for TS_1-VS_3 German and Polish (GER&PL) dataset. TS_1: GCA1-2009 + GCA2-2010 + GCA3-2011, VS_3: GCA1-2014. The factors are genotypes (G), testers (T), years (Y) and locations (L).

	G	*GY*	*Y*	*L*	*GL*	*TL*	*GTL*	*GT*	*TY*	*GTY*	*YL*	*GYL*	*TYL*	*GTYL*
G	1	-0.4388	0.0102	0.0048	-0.0258	0	0.0016	-0.03	0	0.0279	-0.0008	0.0241	-0.0003	-0.0024
GY	**-0.8747**	1	-0.0101	-0.0069	0.0247	0	-0.002	0.027	0	-0.0314	0.0023	-0.0291	0.0005	0.0049
Y	0.4842	**-0.4283**	1	-103.93	0.0367	0	-0.0074	0.0442	0	-0.0447	-73.5891	-0.0388	0.0507	0.0079
L	0.0001	-0.0001	-0.0219	1	-0.0241	0	-0.0002	-0.0025	0	0.0025	-682.792	0.0226	-1.7145	0.0013
GL	-0.0553	0.0596	0.0009	-0.0005	1	0	-0.163	0.0094	0	-0.0093	-0.015	-0.3634	-0.0003	0.1556
TL	-	-	-	-	-	1	0	0	0	0	0	0	0	0
GTL	0.0022	-0.0032	-0.0001	0	**-0.2809**	-	1	-0.064	0	0.0629	0.0004	0.1544	0.0014	-0.8584
GT	-0.1106	0.112	0.0019	-0.0001	0.7852	-	**-0.1898**	1	0	-0.1247	-0.0056	-0.0095	0.0016	0.0629
TY	-	-	-	-	-	-	-	-	1	0	0	0	0	0
GTY	0.1022	-0.129	-0.0019	0.0001	-0.0413	-	0.1849	**-0.9491**	0	1	0.006	0.012	-0.0016	-0.0682
YL	-3E-05	0.0001	-0.0326	**-0.2556**	-0.0007	-	1.2E-05	-0.0004	0	0.0005	1	0.0135	-10.6168	-0.0001
GYL	0.0505	-0.0686	-0.001	0.0005	**-0.9229**	-	0.2601	-0.0414	0	0.0519	0.0006	1	0.0008	-0.1798
TYL	-0.0001	0.0002	0.0002	-0.0058	-0.0001	-	0.0004	0.0012	0	-0.0011	-0.0759	0.0003	1	-0.0021
GTYL	-0.0033	0.0078	0.0001	1.8E-05	0.2665	-	**-0.9758**	0.1851	0	**-0.1991**	0	**-0.3010**	-0.0006	1

Table C.3: Asymptotic correlation (lower diagonal) and covariance (upper diagonal) matrix for TS_1-VS_2 German and Polish (GER&PL) dataset. TS_1: GCA1-2009 + GCA2-2010 + GCA3-2011, VS_2: GCA1-2013. The factors are genotypes (G), testers (T), years (Y) and locations (L).

	G	*GY*	*Y*	*L*	*GL*	*TL*	*GTL*	*GT*	*TY*	*GTY*	*YL*	*GYL*	*TYL*	*GTYL*
G	1	-0.2593	0.0102	-0.0003	-0.0131	0	0.0045	-0.0564	0	0.0509	0.0033	0.0125	-0.0008	-0.0043
GY	**-0.9082**	1	-0.0097	-0.0005	0.0131	0	-0.0042	0.0499	0	-0.0546	-0.0039	-0.0167	0.0009	0.0066
Y	0.3735	**-0.3420**	1	-200.189	0.0207	0	-0.0004	0.0607	0	-0.0639	-71.0256	-0.023	0.0358	0.0018
L	0.0000	0.0000	-0.0454	1	-0.0231	0	-0.0026	0.0027	0	-0.0021	-1206.04	0.0211	0.0789	0.0038
GL	-0.051	0.0531	0.0007	-0.0007	1	0	-0.1314	0.0081	0	-0.0084	-0.0095	-0.2118	-0.0003	0.1268
TL	-	-	-	-	-	1	0	0	0	0	0	0	0	0
GTL	0.0095	-0.0092	0.0000	0.0000	**-0.3215**	-	1	-0.0539	0	0.0529	0.0016	0.128	0.0012	-0.7415
GT	-0.238	0.2191	0.0021	0.0001	0.6708	-	-0.1431	1	0	-0.1772	-0.0112	-0.0095	0.0028	0.0536
TY	-	-	-	-	-	-	-	-	1	0	0	0	0	0
GTY	0.2171	-0.2420	-0.0022	-0.0001	-0.0415	-	0.1419	**-0.9470**	-	1	0.0112	0.0107	-0.0027	-0.0578
YL	0.0001	-0.0002	-0.0217	**-0.3748**	-0.0004	-	3.8E-05	-0.0005	-	0.0005	1	0.0094	-14.4578	-0.0015
GYL	0.0459	-0.0638	-0.0007	0.0006	**-0.8986**	-	0.2956	-0.0436	-	0.0497	0.0004	1	0.0008	-0.1544
TYL	-0.0003	0.0004	0.0001	0.0003	-0.0001	-	0.0003	0.0014	-	-0.0014	-0.0664	0.0003	1	-0.0019
GTYL	-0.009	0.0144	0.0000	0.0001	0.3059	-	**-0.9736**	0.1403	-	-0.1527	0.0000	-0.3515	-0.0005	1

Table C.4: Asymptotic correlation (lower diagonal) and covariance (upper diagonal) matrix for TS_1-VS_1:GCA1-2012 German and Polish (GER&PL) dataset. TS_1: GCA1-2009 + GCA2-2010 + GCA3-2011, VS_1: GCA1-2012. The factors are genotypes (G), testers (T), years (Y) and locations (L).

	G	GY	Y	L	GL	TL	GTL	GT	TY	GTY	YL	GYL	TYL	$GTYL$
G	1	0	0	0	0	0	0	0	0	0	0	0	0	0
GY	-	1	-0.0042	0.0062	0	-0.0006	0.0035	-0.0008	0	-0.0074	-0.0003	-0.0076	0.0004	0
Y	-	-0.0002	1	-71.8508	0	11.8454	0.0008	0.1002	0	-0.0979	-40.1089	-0.0006	-11.9180	0
L	-	0.0003	-0.0224	1	0	4.5978	-0.0015	-0.0023	0	0.0026	-357.1928	-0.0009	-5.6630	0
GL	-	-	-	-	1	0	0	0	0	0	0	0	0	0
TL	-	-0.0001	0.0158	0.0055	-	1	-0.0028	-0.0194	0	0.0202	-23.6140	0.0018	-182.2474	0
GTL	-	0.0546	0.0001	-0.0001	-	-0.0010	1	-0.0027	0	-0.0044	0.0023	-0.0240	0.0005	0
GT	-	-0.0063	0.0049	-0.0001	-	-0.0036	-0.0371	1	0	-0.1423	-0.0021	0.0016	0.0207	0
TY	-	-	-	-	-	-	-	-	1	0	0	0	0	0
GTY	-	-0.0557	-0.0046	0.0001	-	0.0037	-0.0571	**-0.9440**	-	1	0.0018	0.0015	-0.0210	0
YL	-	0.0000	-0.0268	**-0.2138**	-	-0.0605	0.0004	-0.0002	-	0.0002	1	-0.0027	12.0806	0
GYL	-	-0.1057	-0.0001	-0.0001	-	0.0006	**-0.5756**	0.0196	-	0.0173	-0.0004	1	-0.0003	0
TYL	-	0.0001	-0.0163	-0.0069	-	**-0.9551**	0.0002	0.0040	-	-0.0039	0.0317	-0.0001	1	0
$GTYL$	-	-	-	-	-	-	-	-	-	-	-	-	-	1

Table C.5: Asymptotic correlation (lower diagonal) and covariance (upper diagonal) matrix for TS_2-VS_3:GCA1-2014 German and Polish (GER&PL) dataset. TS_2: GCA1-2009 + GCA2-2010 + GCA1-2010 + GCA2-2011, VS_3: GCA1-2014. The factors are genotypes (G), testers (T), years (Y) and locations (L).

	G	GY	Y	L	GL	TL	GTL	GT	TY	GTY	YL	GYL	TYL	$GTYL$
G	1	-0.1791	0.0089	-0.0124	-0.0373	0	0.0115	-0.0365	0	0.0336	0.0102	0.0373	-0.0081	-0.0125
GY	**-0.8111**	1	-0.0090	0.0086	0.0373	0	-0.0119	0.0338	0	-0.0369	-0.0069	-0.0420	0.0069	0.0154
Y	0.2371	**-0.1994**	1	12.2686	0.0398	0	-0.0149	0.0045	0	-0.0064	-74.2372	-0.0397	-0.3763	0.0140
L	-0.0003	0.0003	0.0032	1	-0.0291	0	-0.0058	0.0031	0	-0.0025	-458.0833	0.0284	-0.3409	0.0062
GL	-0.0915	0.1094	0.0010	-0.0005	1	0	-0.3417	0.0147	0	-0.0142	0.0090	-0.6169	-0.0015	0.3356
TL	-	-	-	-	-	1	0	0	0	0	0	0	0	0
GTL	0.0275	-0.0342	-0.0004	-0.0001	**-0.5319**	-	1	-0.0385	0	0.0378	-0.0097	0.3364	0.0023	-0.6483
GT	**-0.2562**	0.2843	0.0003	0.0002	1.1782	-	-0.1716	1	0	-0.0710	-0.0042	-0.0143	0.0001	0.0373
TY	-	-	-	-	-	-	-	-	1	0	0	0	0	0
GTY	0.2345	**-0.3084**	-0.0004	-0.0001	-0.0644	-	0.1676	**-0.9205**	-	1	0.0044	0.0170	-0.0002	-0.0433
YL	0.0006	-0.0005	-0.0471	**-0.2043**	0.0004	-	-3.9E-04	-0.0005	-	0.0005	1	-0.0095	-3.1882	0.0095
GYL	0.0900	-0.1214	-0.0010	0.0005	**-0.9657**	-	0.5157	-0.0641	-	0.0758	-0.0004	1	0.0013	-0.3594
TYL	-0.0075	0.0076	-0.0035	-0.0022	-0.0009	-	0.0013	0.0002	-	-0.0003	-0.0496	0.0008	1	-0.0021
$GTYL$	-0.0294	0.0436	0.0003	0.0001	0.5128	-	**-0.9701**	0.1633	-	**-0.1884**	0.0004	**-0.5409**	-0.0012	1

Table C.6: Asymptotic correlation (lower diagonal) and covariance (upper diagonal) matrix for TS_2-VS_2:GCA1-2013 German and Polish (GER&PL) dataset. TS_2: GCA1-2009 + GCA2-2010 + GCA1-2010 + GCA2-2011, VS_2: GCA1-2013. The factors are genotypes (G), testers (T), years (Y) and locations (L).

	Y	G	GY	L	GL	TL	GTL	GT	TY	GTY	YL	GYL	TYL	$GTYL$
G	1	-0.1341	0.0084	-0.0123	-0.0243	0	0.0106	-0.0475	0	0.0425	0.0110	0.0240	-0.0060	-0.0106
GY	**-0.8469**	1	-0.0080	0.0097	0.0245	0	-0.0107	0.0429	0	-0.0460	-0.0093	-0.0281	0.0052	0.0137
Y	0.1754	-0.1511	1	3.5213	0.0199	0	-0.0064	0.0044	0	-0.0078	-120.6435	-0.0203	-0.3188	0.0063
L	-0.0005	0.0004	0.0011	1	-0.0195	0	-0.0107	0.0045	0	-0.0036	-837.9200	0.0187	0.0361	0.0112
GL	-0.0913	0.1020	0.0006	-0.0005	1	0	-0.2598	0.0117	0	-0.0115	0.0059	-0.3946	-0.0008	0.2552
TL	-	-	-	-	-	1	0	0	0	0	0	0	0	0
GTL	0.0345	-0.0388	-0.0002	-0.0002	-0.5584	-	1	-0.0325	0	0.0320	-0.0063	0.2563	0.0018	-0.5317
GT	**-0.3540**	0.3555	0.0003	0.0002	0.9334	-	-0.1385	1	0	-0.0933	-0.0064	-0.0117	0.0004	0.0316
TY	-	-	-	-	-	-	-	-	1	0	0	0	0	0
GTY	0.3204	**-0.3847**	-0.0005	-0.0002	-0.0570	-	0.1378	**-0.9229**	-	1	0.0064	0.0137	-0.0004	-0.0367
YL	0.0006	-0.0006	-0.0572	**-0.3189**	0.0002	-	-2.0E-04	-0.0005	-	0.0005	1	-0.0063	-4.4431	0.0063
GYL	0.0878	-0.1147	-0.0006	0.0005	**-0.9563**	-	0.5378	-0.0564	-	0.0666	-0.0002	1	0.0007	-0.2779
TYL	-0.0058	0.0057	-0.0026	0.0002	-0.0005	-	0.0010	0.0004	-	-0.0005	-0.0430	0.0004	1	-0.0016
$GTYL$	-0.0337	0.0484	0.0002	0.0002	0.5361	-	**-0.9671**	0.1319	-	-0.1547	0.0002	-0.5700	-0.0009	1

Table C.7: Asymptotic correlation (lower diagonal) and covariance (upper diagonal) matrix for TS_2-VS_1:GCA1-2012 German and Polish (GER&PL) dataset. TS_2: GCA1-2009 + GCA2-2010 + GCA1-2010 + GCA2-2011, VS_1: GCA1-2012. The factors are genotypes (G), testers (T), years (Y) and locations (L).

	G	GY	Y	L	GL	TL	GTL	GT	TY	GTY	YL	GYL	TYL	$GTYL$
G	1	-0.1526	0.2375	0.0016	-0.0222	0	0.0086	-0.0326	0	0.0294	0.0011	0.0210	-0.0067	-0.0083
GY	**-0.7924**	1	-0.1947	0.0024	0.0222	0	-0.0089	0.0296	0	-0.0330	-0.0013	-0.0258	0.0053	0.0116
Y	0.0089	-0.0087	1	0.7995	0.0208	0	-0.0048	0.0057	0	-0.0084	-46.5272	-0.0219	-0.3627	0.0045
L	0.0001	0.0001	0.0003	1	-0.0164	0	-0.0050	0.0007	0	-0.0003	-243.4461	0.0143	-0.2184	0.0051
GL	-0.0764	0.0908	0.0006	-0.0005	1	0	-0.2234	0.0107	0	-0.0103	0.0042	-0.3623	-0.0006	0.2200
TL	-	-	-	-	-	1	0	0	0	0	0	0	0	0
GTL	0.0259	-0.0319	-0.0001	-0.0001	**-0.5329**	-	1	-0.0288	0	0.0282	-0.0069	0.2200	0.0016	-0.4705
GT	-0.2527	0.2739	0.0004	0.0000	0.0652	-	-0.1548	1	0	-0.0667	-0.0021	-0.0099	0.0002	0.0276
TY	-	-	-	-	-	-	-	-	1	0	0	0	0	0
GTY	0.2259	**-0.3026**	-0.0006	0.0000	-0.0624	-	0.1507	**-0.9122**	-	1	0.0020	0.0124	-0.0001	-0.0329
YL	0.0001	-0.0001	-0.0343	-0.1802	0.0003	-	-0.0004	-0.0003	-	0.0003	1	-0.0043	-3.3215	0.0071
GYL	0.0703	-0.1033	-0.0006	0.0004	**-0.9589**	-	0.5123	-0.0593	-	0.0737	-0.0003	1	0.0005	-0.2364
TYL	-0.0063	0.0060	-0.0029	-0.0018	-0.0004	-	0.0011	0.0003	-	-0.0002	-0.0615	0.0004	1	-0.0015
$GTYL$	-0.0247	0.0409	0.0001	0.0001	0.5149	-	**-0.9685**	0.1455	-	-0.1719	0.0004	**-0.5399**	-0.0009	1

Table C.8: Asymptotic correlation (lower diagonal) and covariance (upper diagonal) matrix for TS_3-VS_3:GCA1-2014 German and Polish (GER&PL) dataset. TS_3: GCA1-2009 + GCA2-2010 + GCA3-2011 + GCA1-2010 + GCA2-2011 + GCA3-2012 + GCA1-2011 + GCA2-2012 + GCA3-2013, VS_3: GCA1-2014. The factors are genotypes (G), testers (T), years (Y) and locations (L).

	G	GY	Y	L	GL	TL	GTL	GT	TY	GTY	YL	GYL	TYL	$GTYL$
G	1	-0.0427	-0.0394	-0.0046	-0.0045	0.0017	-0.0004	-0.0108	0	0.0091	0.0021	0.0039	-0.0011	0.0001
GY	**-0.6425**	1	0.0300	0.0028	0.0038	-0.0009	-0.0001	0.0089	0	-0.0108	-0.0012	-0.0058	0.0006	0.0015
Y	-0.0027	0.0030	1	-11.3578	0.0015	-3.6709	0.0090	-0.0039	0	0.0029	-13.0843	-0.0009	2.9638	-0.0091
L	-0.0003	0.0002	-0.0042	1	-0.0005	-2.4570	-0.0026	0.0003	0	0.0000	-95.7799	0.0015	1.7234	0.0019
GL	-0.0544	0.0687	0.0001	0.0000	1	0.0005	-0.0342	0.0018	0	-0.0019	0.0042	-0.0649	-0.0004	0.0333
TL	0.0010	-0.0008	-0.0150	-0.0082	0.0003	1	0.0006	-0.0010	0	0.0000	1.3346	0.0015	-22.8760	-0.0018
GTL	-0.0028	-0.0013	0.0004	-0.0001	**-0.2839**	0.0003	1	-0.0130	0	0.0129	-0.0020	0.0329	0.0001	-0.2099
GT	-0.1873	0.2286	-0.0005	0.0000	0.0371	-0.0010	-0.1531	1	0	-0.0297	-0.0007	-0.0018	0.0009	0.0126
TY	-	-	-	-	-	-	-	-	1	0	0	0	0	0
GTY	0.1601	**-0.2815**	0.0003	0.0000	-0.0398	0.0000	0.1538	**-0.8880**	-	1	0.0007	0.0029	0.0004	-0.0152
YL	0.0004	-0.0003	-0.0165	-0.0981	0.0009	0.0150	-0.0003	-0.0002	-	0.0002	1	-0.0044	-4.4463	0.0020
GYL	0.0430	-0.0940	-0.0001	0.0001	**-0.8576**	0.0010	0.2468	-0.0333	-	0.0553	-0.0009	1	-0.0016	-0.0445
TYL	-0.0007	0.0006	0.0127	0.0060	-0.0003	**-0.8751**	0.0000	0.0010	-	0.0004	-0.0525	-0.0011	1	0.0007
$GTYL$	0.0006	0.0149	-0.0004	0.0001	0.2688	-0.0007	**-0.9604**	0.1435	-	-0.1768	0.0003	-0.3239	0.0003	1

Table C.9: Asymptotic correlation (lower diagonal) and covariance (upper diagonal) matrix for TS_3-VS_2:GCA1-2013 German and Polish (GER&PL) dataset. TS_3: GCA1-2009 + GCA2-2010 + GCA3-2011 + GCA1-2010 + GCA2-2011 + GCA3-2012 + GCA1-2011 + GCA2-2012 + GCA3-2013, VS_2: GCA1-2013. The factors are genotypes (G), testers (T), years (Y) and locations (L).

	G	GY	Y	L	GL	TL	GTL	GT	TY	GTY	YL	GYL	TYL	$GTYL$
G	1	-0.0489	-0.0526	0.0053	-0.0033	0.0038	-0.0001	-0.0116	0	0.0093	0.0018	0.0027	-0.0001	0.0001
GY	**-0.6967**	1	0.0438	-0.0054	0.0028	-0.0028	-0.0003	0.0087	0	-0.0110	-0.0020	-0.0047	-0.0003	0.0016
Y	-0.0037	0.0040	1	-10.9361	0.0002	-4.7179	0.0079	-0.0066	0	0.0051	-17.3793	-0.0001	3.8107	-0.0080
L	0.0003	-0.0004	-0.0042	1	-0.0001	-3.3662	-0.0035	-0.0011	0	0.0012	-131.2160	0.0006	2.6769	0.0034
GL	-0.0506	0.0550	0.0000	0.0000	1	0.0016	-0.0254	0.0013	0	-0.0015	0.0011	-0.0434	-0.0024	0.0248
TL	0.0023	-0.0021	-0.0176	-0.0110	0.0013	1	-0.0001	-0.0019	0	0.0005	2.5194	0.0002	-25.7527	-0.0008
GTL	-0.0010	-0.0031	0.0004	-0.0001	**-0.2712**	0.0000	1	-0.0118	0	0.0116	-0.0002	0.0246	0.0012	-0.1867
GT	**-0.1988**	0.1937	-0.0007	-0.0001	0.0316	-0.0017	-0.1408	1	0	-0.0311	-0.0016	-0.0014	0.0012	0.0113
TY	-	-	-	-	-	-	-	-	1	0	0	0	0	0
GTY	0.1659	**-0.2552**	0.0006	0.0001	-0.0366	0.0005	0.1447	**-0.8708**	-	1	0.0016	0.0025	0.0003	-0.0140
YL	0.0003	-0.0004	-0.0187	-0.1234	0.0003	0.0230	0.0000	-0.0004	-	0.0004	1	0.0040	-5.4092	-0.0036
GYL	0.0365	-0.0819	0.0000	0.0000	**-0.8136**	0.0002	0.2281	-0.0290	-	0.0536	0.0008	1	-0.0001	-0.0359
TYL	-0.0001	-0.0002	0.0152	0.0094	-0.0021	**-0.8720**	0.0005	0.0012	-	0.0004	-0.0528	-0.0001	1	-0.0001
$GTYL$	0.0011	0.0155	-0.0004	0.0001	0.2561	-0.0003	**-0.9560**	0.1306	-	-0.1677	-0.0004	**-0.3222**	0.0000	1

Table C.10: Asymptotic correlation (lower diagonal) and covariance (upper diagonal) matrix for TS_3-VS_1:GCA1-2012 German and Polish (GER&PL) dataset. TS_3: GCA1-2009 + GCA2-2010 + GCA3-2011 + GCA1-2010 + GCA2-2011 + GCA3-2012 + GCA1-2011 + GCA2-2012 + GCA3-2013, VS_1: GCA1-2012. The factors are genotypes (G), testers (T), years (Y) and locations (L).

	G	GY	Y	L	GL	TL	GTL	GT	TY	GTY	YL	GYL	TYL	$GTYL$
G	1	-0.2622	-0.4926	0.0179	-0.0056	-0.0058	-0.0007	-0.0140	0.0002	0.0122	-0.0096	0.0052	0.0050	0.0005
GY	**-0.8915**	1	0.4666	-0.0146	0.0049	0.0054	0.0007	0.0118	-0.0007	-0.0148	0.0088	-0.0082	-0.0050	0.0012
Y	-0.0158	0.0153	1	-8.4482	-0.0009	-7.1051	0.0157	-0.0194	-2.3148	0.0149	-15.8806	-0.0016	6.4270	-0.0130
L	0.0006	-0.0005	-0.0027	1	-0.0126	-2.8175	0.0006	0.0003	-0.2087	0.0002	-118.7310	0.0130	2.1704	-0.0017
GL	-0.0351	0.0310	-0.0001	-0.0008	1	-0.0009	-0.0455	0.0023	-0.0009	-0.0025	0.0114	-0.0821	-0.0012	0.0445
TL	-0.0022	0.0020	-0.0255	-0.0106	-0.0006	1	-0.0006	-0.0001	1.8393	0.0005	1.8097	0.0032	-19.5968	-0.0018
GTL	-0.0027	0.0028	0.0006	0.0000	-0.3169	-0.0002	1	-0.0160	-0.0007	0.0159	-0.0048	0.0437	0.0027	-0.2373
GT	-0.1163	0.1001	-0.0016	0.0000	0.0365	-0.0001	-0.1481	1	-0.0017	-0.0431	-0.0015	-0.0021	0.0001	0.0151
TY	0.0003	-0.0009	-0.0275	-0.1586	-0.0020	0.2532	-0.0010	-0.0051	1	0.0029	0.1432	0.0006	-2.0138	0.0002
GTY	0.1014	-0.1257	0.0012	0.0000	-0.0392	0.0004	0.1472	**-0.8951**	0.0088	1	0.0011	0.0038	0.0000	-0.0190
YL	-0.0010	0.0009	-0.0153	-0.1190	0.0021	0.0201	-0.0005	-0.0004	0.0053	0.0003	1	-0.0106	-4.2475	0.0045
GYL	0.0296	-0.0473	-0.0001	0.0007	**-0.8693**	0.0020	0.2753	-0.0293	0.0012	0.0542	-0.0018	1	-0.0020	-0.0568
TYL	0.0020	-0.0020	0.0244	0.0086	-0.0009	**-0.8614**	0.0012	0.0001	**-0.2934**	0.0000	-0.0500	-0.0013	1	-0.0006
$GTYL$	0.0019	0.0043	-0.0005	-0.0001	0.3016	-0.0007	**-0.9553**	0.1362	0.0003	-0.1720	0.0005	**-0.3480**	-0.0002	1

C.5 Predictive abilities of sampling scenarios

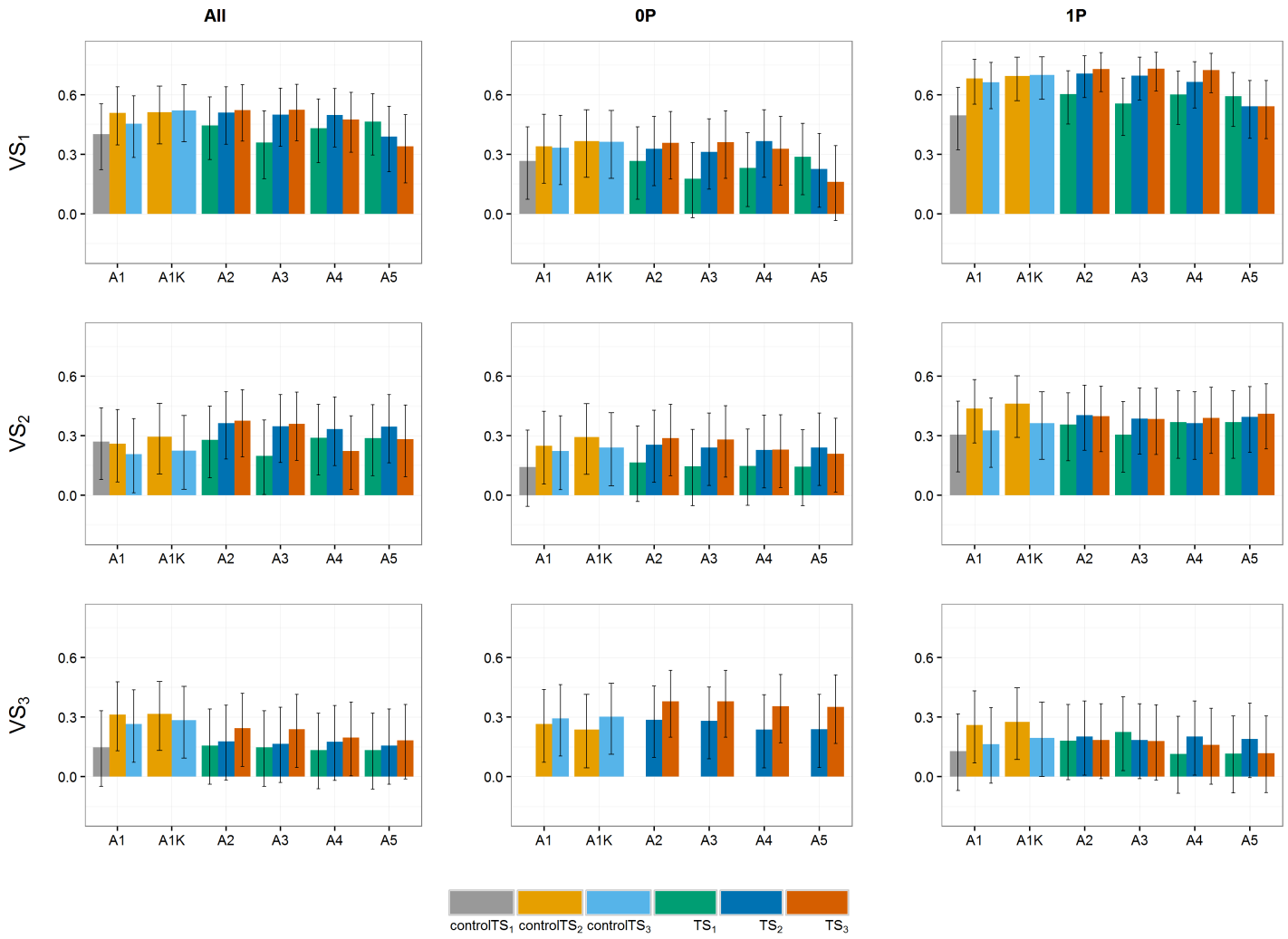


Figure C.5: **Predictive abilities (y-axis) of the German dataset using VS-size of 100 genotypes for the three scenarios.** TS_1 and controlTS_1, TS_2 and controlTS_2, and TS_3 and controlTS_3 to predict the validation sets VS_1, VS_2 and VS_3 with All-, 0P- and 1P-scenarios. Black lines for each bar represent the 95% confidence intervals of the mean predictive ability. Year-wise approach (A1) and year-wise with kinship approach (A1K) were fitted to the control sets, approaches 2-stg-Kin (A2), 2-stg-Kin-het (A3), 3-stg-NoKin (A4) and 3-stg-Kin (A5) to the complete sets. TS_1: GCA1-2009 + GCA2-2010 + GCA3-2011, controlTS_1: GCA1-2009, TS_2: GCA1-2009 + GCA2-2010 + GCA1-2010 + GCA2-2011, controlTS_2: GCA1-2009 + GCA1-2010, TS_3: GCA1-2009 + GCA2-2010 + GCA3-2011 + GCA1-2010 + GCA2-2011 + GCA3-2012 + GCA1-2011 + GCA2-2012 + GCA3-2013, controlTS_3: GCA1-2009 + GCA1-2010 + GCA1-2011, VS_1: GCA1-2012, VS_2: GCA1-2013, VS_3: GCA1-2014.

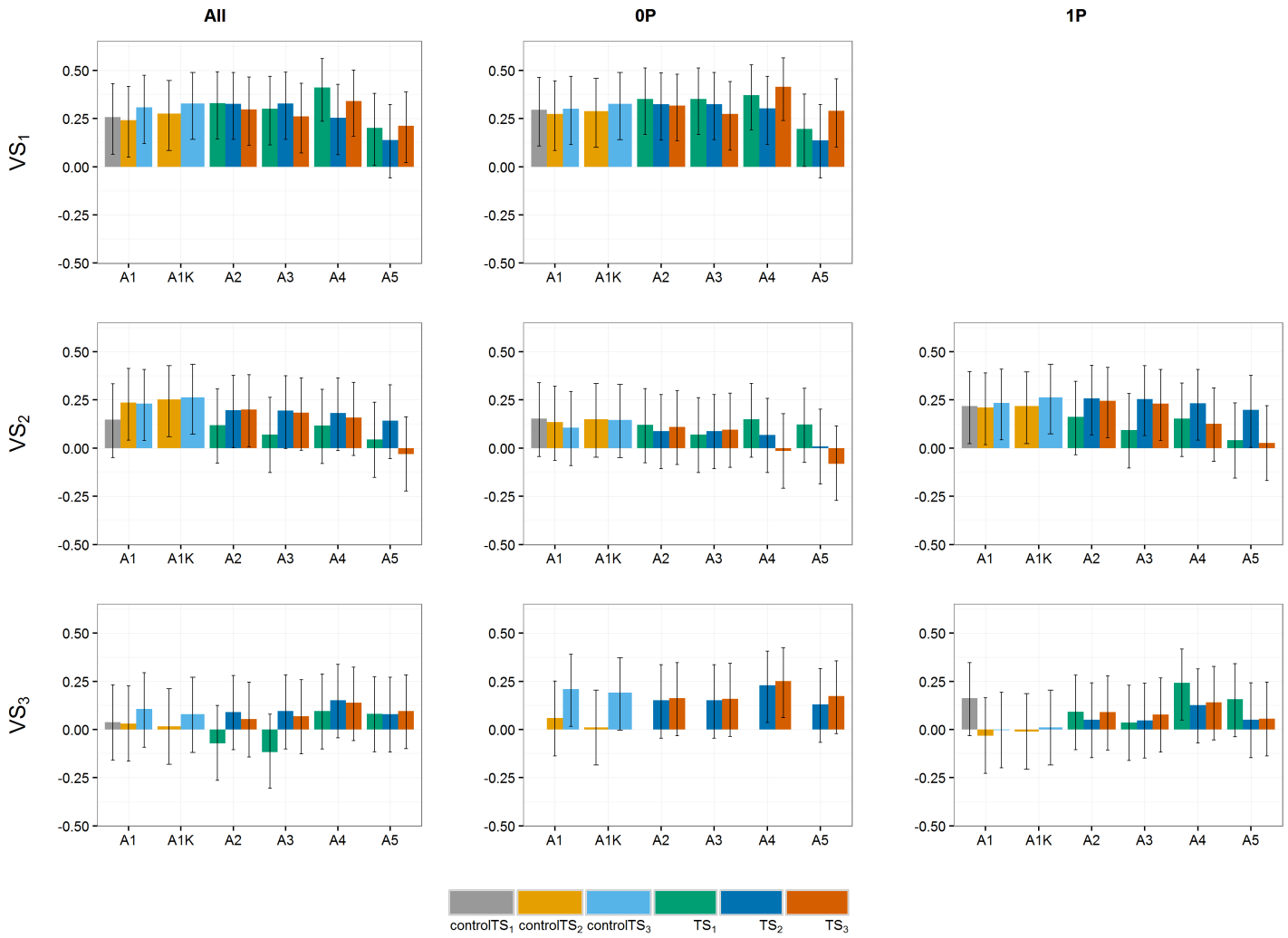


Figure C.6: **Predictive abilities (y-axis) of the Polish dataset using VS-size of 100 genotypes for the three scenarios.** TS_1 and controlTS$_1$, TS_2 and controlTS$_2$, and TS_3 and controlTS$_3$ to predict the validation sets VS_1, VS_2 and VS_3 with All-, 0P- and 1P-scenarios. Black lines for each bar represent the 95% confidence intervals of the mean predictive ability. Year-wise approach (A1) and year-wise with kinship approach (A1K) were fitted to the control sets, approaches 2-stg-Kin (A2), 2-stg-Kin-het (A3), 3-stg-NoKin (A4) and 3-stg-Kin (A5) to the complete sets. TS_1: GCA1-2009 + GCA2-2010 + GCA3-2011, controlTS$_1$: GCA1-2009, TS_2: GCA1-2009 + GCA2-2010 + GCA1-2010 + GCA2-2011, controlTS$_2$: GCA1-2009 + GCA1-2010, TS_3: GCA1-2009 + GCA2-2010 + GCA3-2011 + GCA1-2010 + GCA2-2011 + GCA3-2012 + GCA1-2011 + GCA2-2012 + GCA3-2013, controlTS$_3$: GCA1-2009 + GCA1-2010 + GCA1-2011, VS_1: GCA1-2012, VS_2: GCA1-2013, VS_3: GCA1-2014.

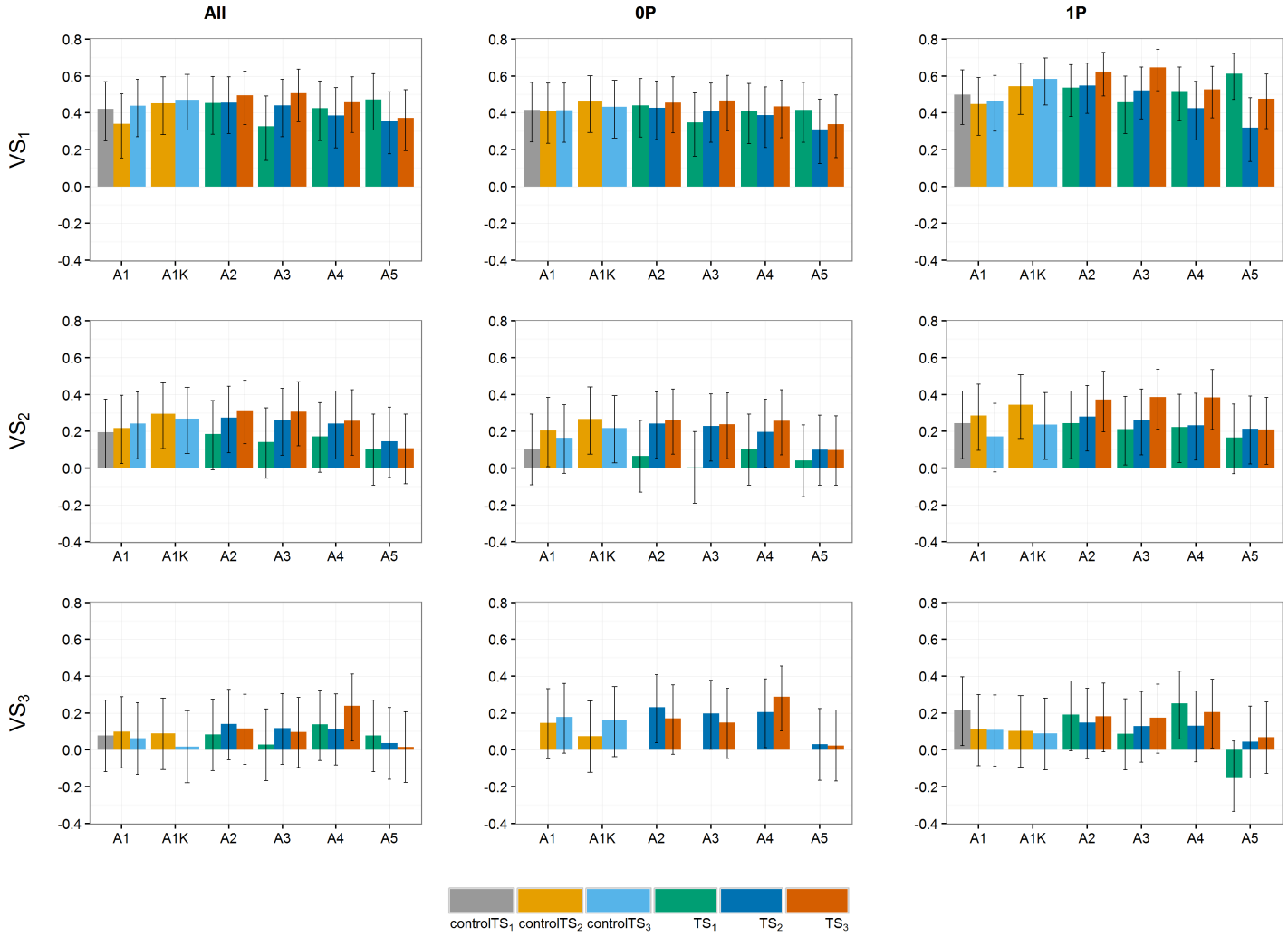


Figure C.7: **Predictive abilities (y-axis) of the German and Polish dataset using VS-size of 100 genotypes for the three scenarios.** TS_1 and controlTS_1, TS_2 and controlTS_2, and TS_3 and controlTS_3 to predict the validation sets VS_1, VS_2 and VS_3 with All-, 0P- and 1P-scenarios. Black lines for each bar represent the 95% confidence intervals of the mean predictive ability. Year-wise approach (A1) and year-wise with kinship approach (A1K) were fitted to the control sets, approaches 2-stg-Kin (A2), 2-stg-Kin-het (A3), 3-stg-NoKin (A4) and 3-stg-Kin (A5) to the complete sets. TS_1: GCA1-2009 + GCA2-2010 + GCA3-2011, controlTS_1: GCA1-2009, TS_2: GCA1-2009 + GCA2-2010 + GCA1-2010 + GCA2-2011, controlTS_2: GCA1-2009 + GCA1-2010, TS_3: GCA1-2009 + GCA2-2010 + GCA3-2011 + GCA1-2010 + GCA2-2011 + GCA3-2012 + GCA1-2011 + GCA2-2012 + GCA3-2013, controlTS_3: GCA1-2009 + GCA1-2010 + GCA1-2011, VS_1: GCA1-2012, VS_2: GCA1-2013, VS_3: GCA1-2014.

C.6 PCA plots

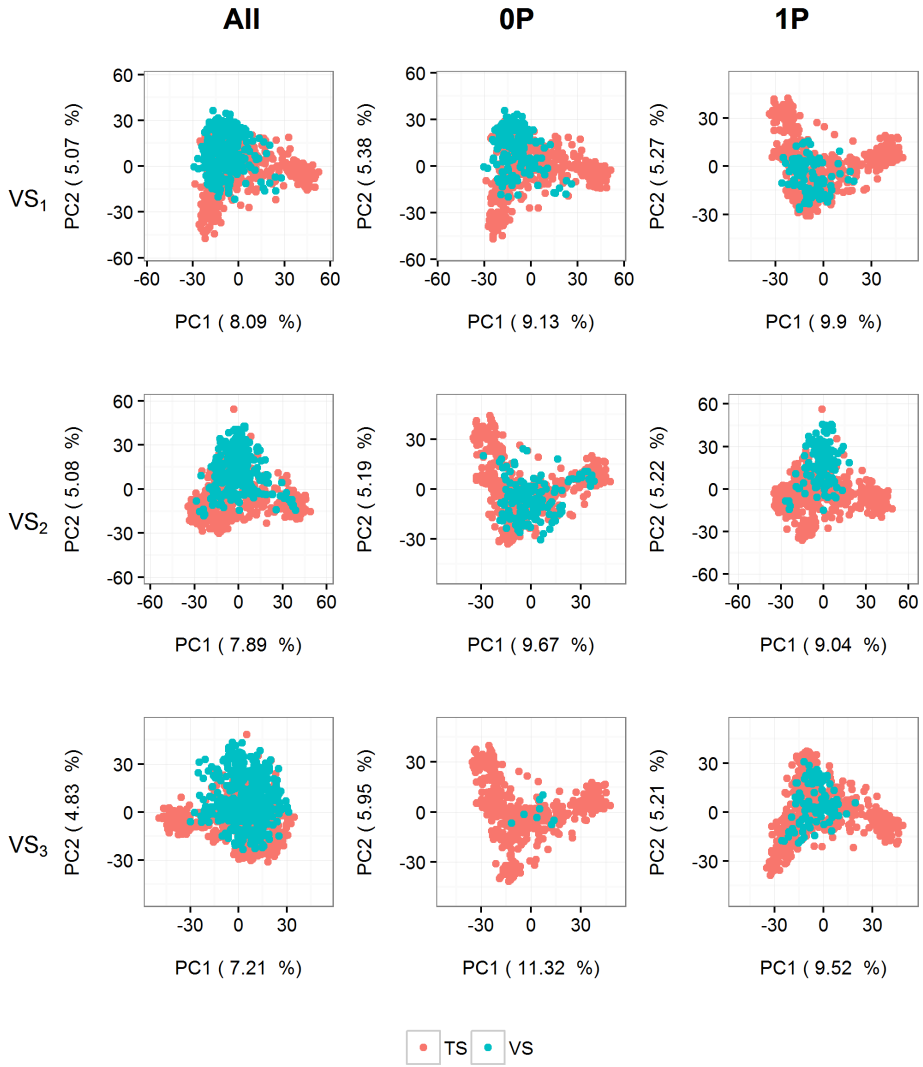


Figure C.8: **Principal component (PC) plots for the German dataset between TS_1 and all VS.** TS_1 and relatedness scenarios (All-, 0P- and 1P-scenarios) for VS_1, VS_2 and VS_3. TS_1:GCA1-2009 + GCA2-2010 + GCA3-2011, VS_1:GCA1-2012, VS_2:GCA1-2013, VS_3:GCA1-2014.

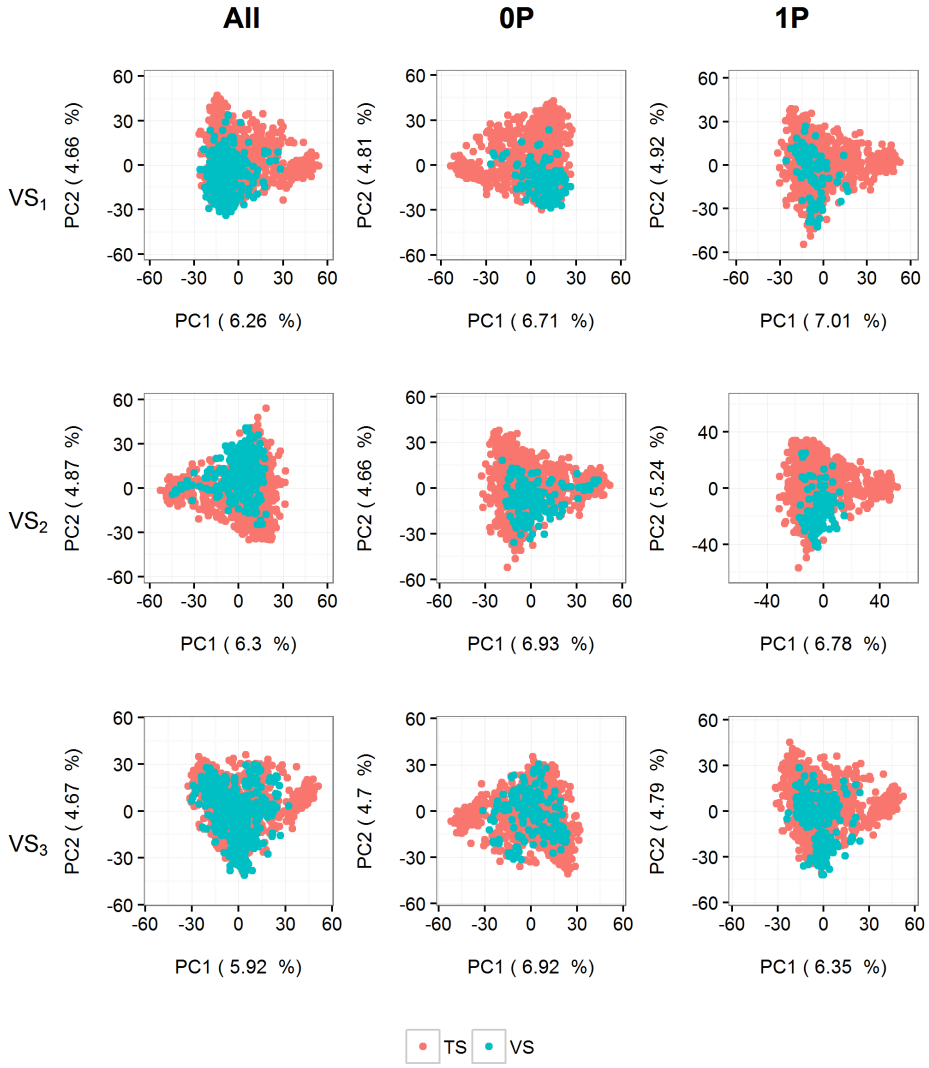


Figure C.9: **Principal component (PC) plots for the German dataset between TS_2 and all VS.** TS_2 and relatedness scenarios (All-, 0P- and 1P-scenarios) for VS_1, VS_2 and VS_3. TS_2:GCA1-2009 + GCA2-2010 + GCA1-2010 + GCA2-2011, VS_1:GCA1-2012, VS_2:GCA1-2013, VS_3:GCA1-2014.

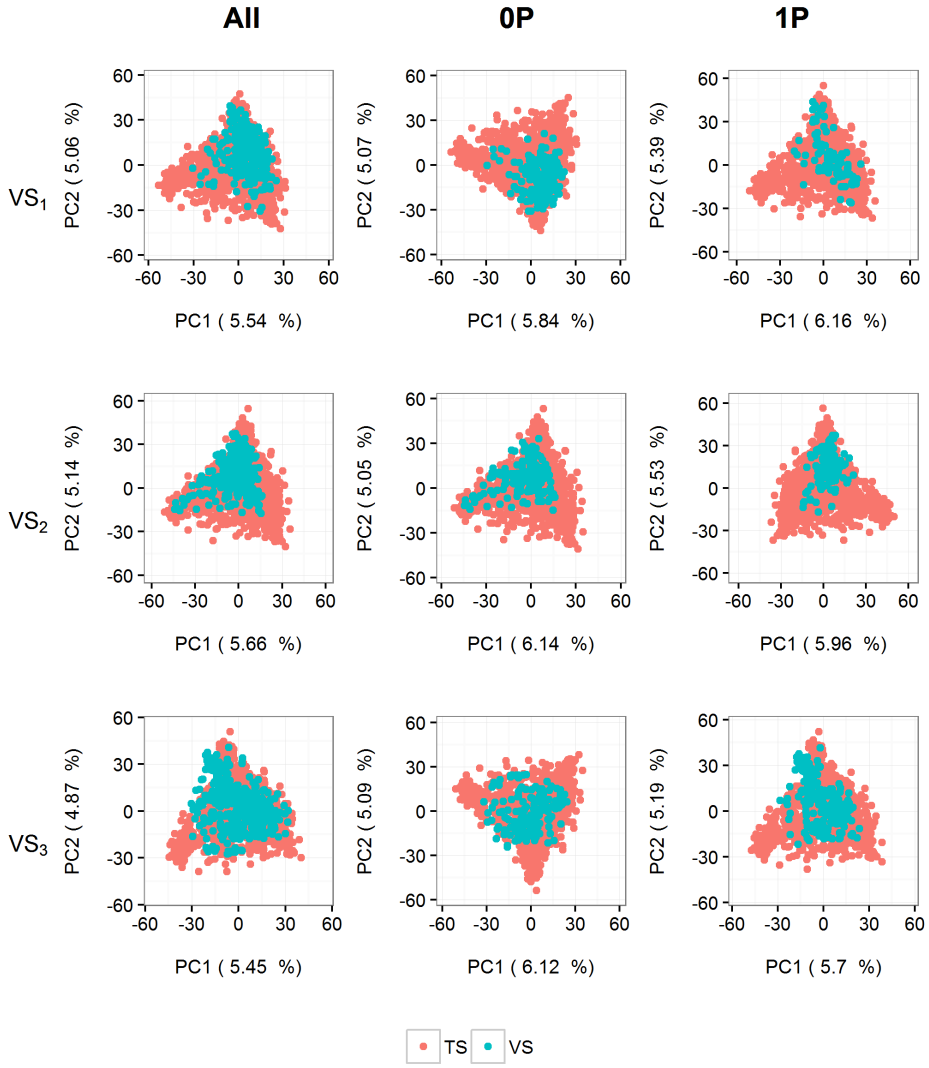


Figure C.10: **Principal component (PC) plots for the German dataset between TS_3 and all VS.** TS_3 and relatedness scenarios (All-, 0P- and 1P-scenarios) for VS_1, VS_2 and VS_3. TS_3:GCA1-2009 + GCA2-2010 + GCA3-2011 + GCA1-2010 + GCA2-2011 + GCA3-2012 + GCA1-2011 + GCA2-2012 + GCA3-2013, VS_1:GCA1-2012, VS_2:GCA1-2013, VS_3:GCA1-2014.

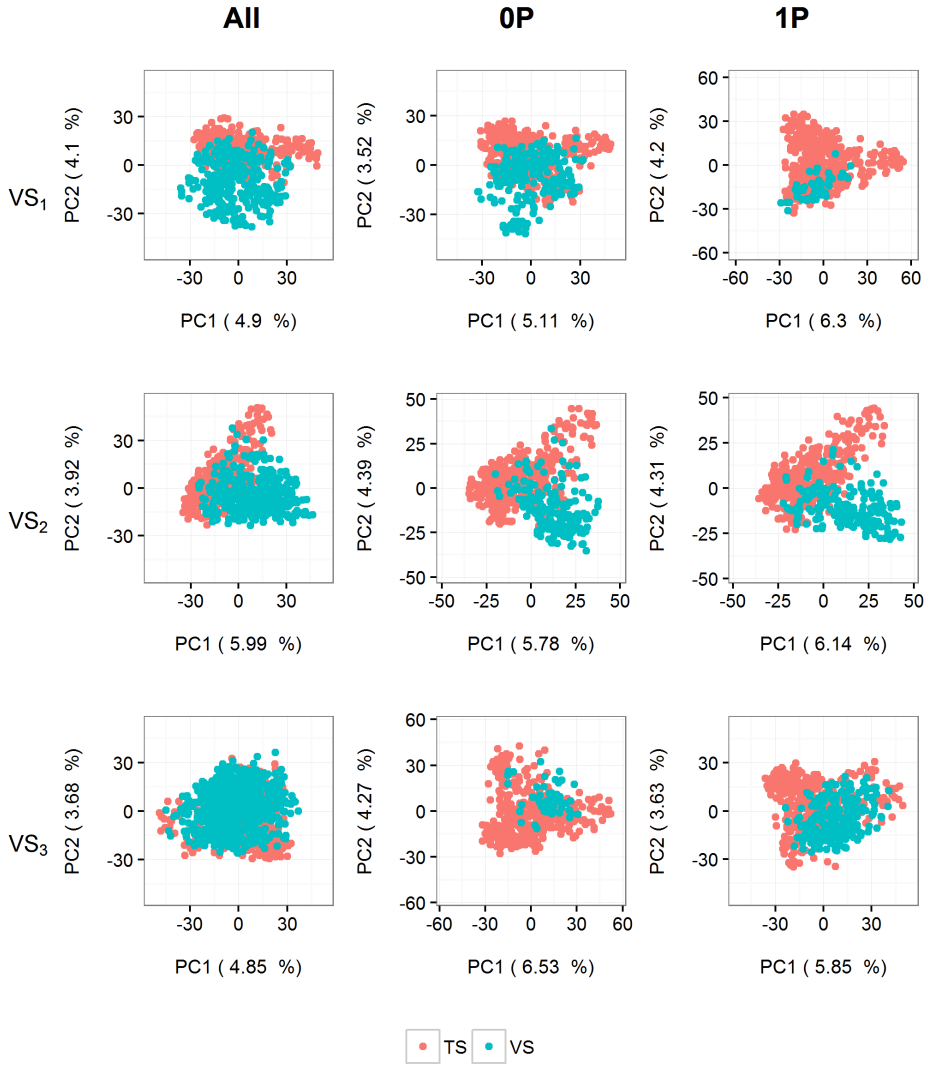


Figure C.11: **Principal component (PC) plots for the Polish dataset between TS_1 and all VS.** TS_1 and relatedness scenarios (All-, 0P- and 1P-scenarios) for VS_1, VS_2 and VS_3. TS_1:GCA1-2009 + GCA2-2010 + GCA3-2011, VS_1:GCA1-2012, VS_2:GCA1-2013, VS_3:GCA1-2014.

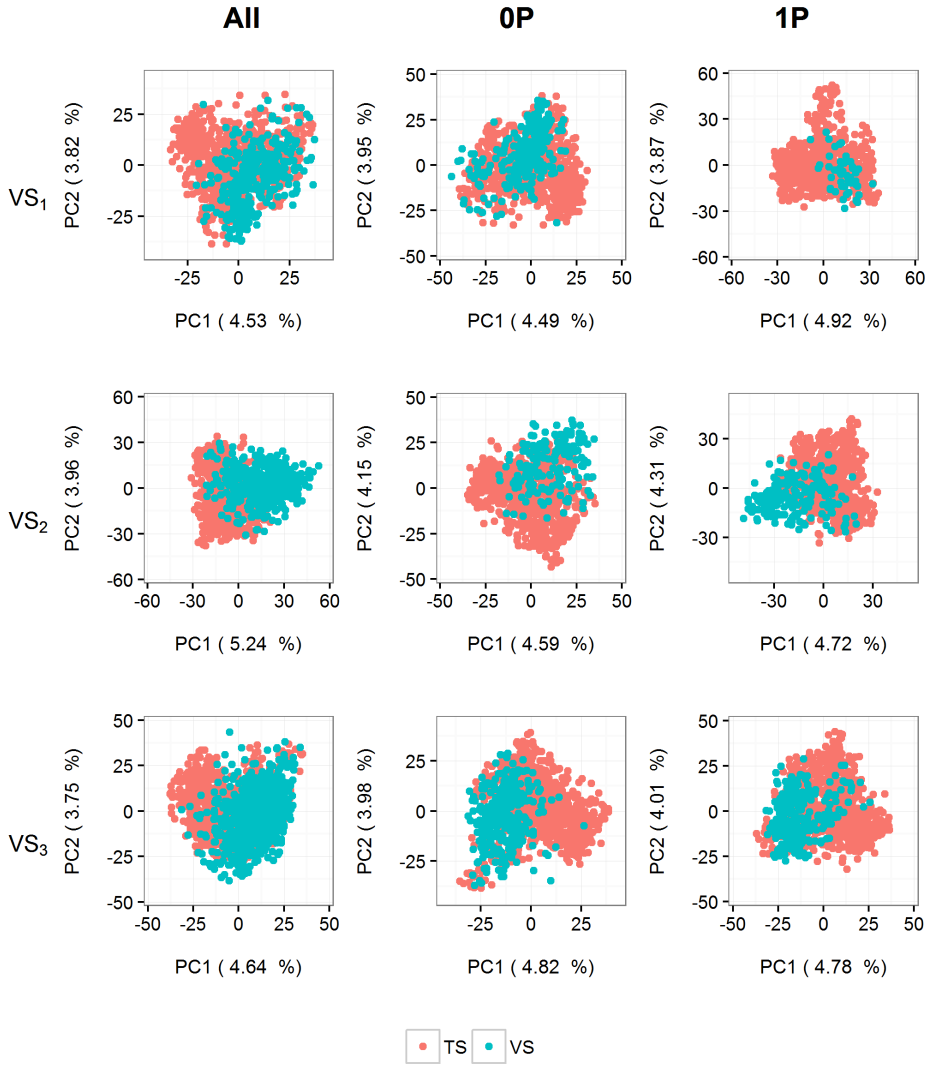


Figure C.12: **Principal component (PC) plots for the Polish dataset between TS_2 and all VS. TS_2 and relatedness scenarios (All-, 0P- and 1P-scenarios) for VS_1, VS_2 and VS_3.** TS_2:GCA1-2009 + GCA2-2010 + GCA1-2010 + GCA2-2011, VS_1:GCA1-2012, VS_2:GCA1-2013, VS_3:GCA1-2014.

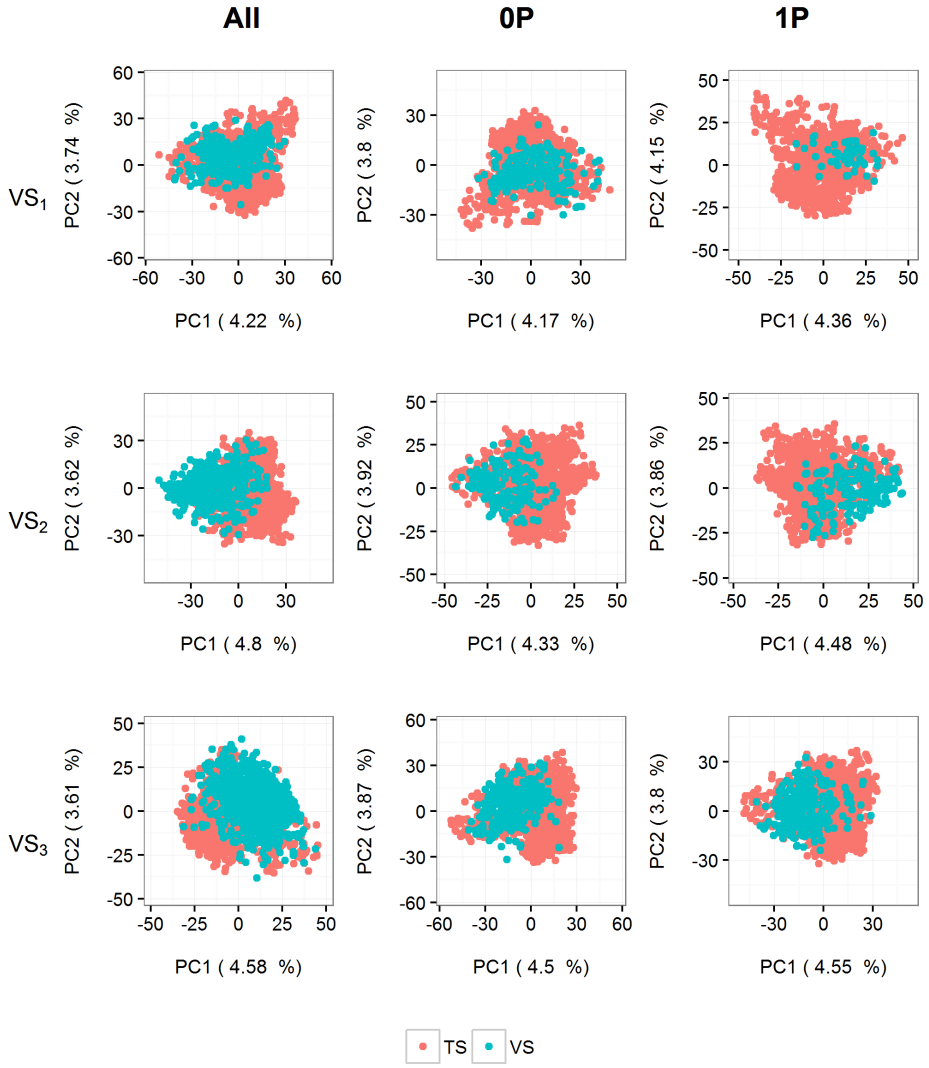


Figure C.13: **Principal component (PC) plots for the Polish dataset between TS_3 and all VS.** TS_3 and relatedness scenarios (All-, 0P- and 1P-scenarios) for VS_1, VS_2 and VS_3. TS_3:GCA1-2009 + GCA2-2010 + GCA3-2011 + GCA1-2010 + GCA2-2011 + GCA3-2012 + GCA1-2011 + GCA2-2012 + GCA3-2013, VS_1:GCA1-2012, VS_2:GCA1-2013, VS_3:GCA1-2014.

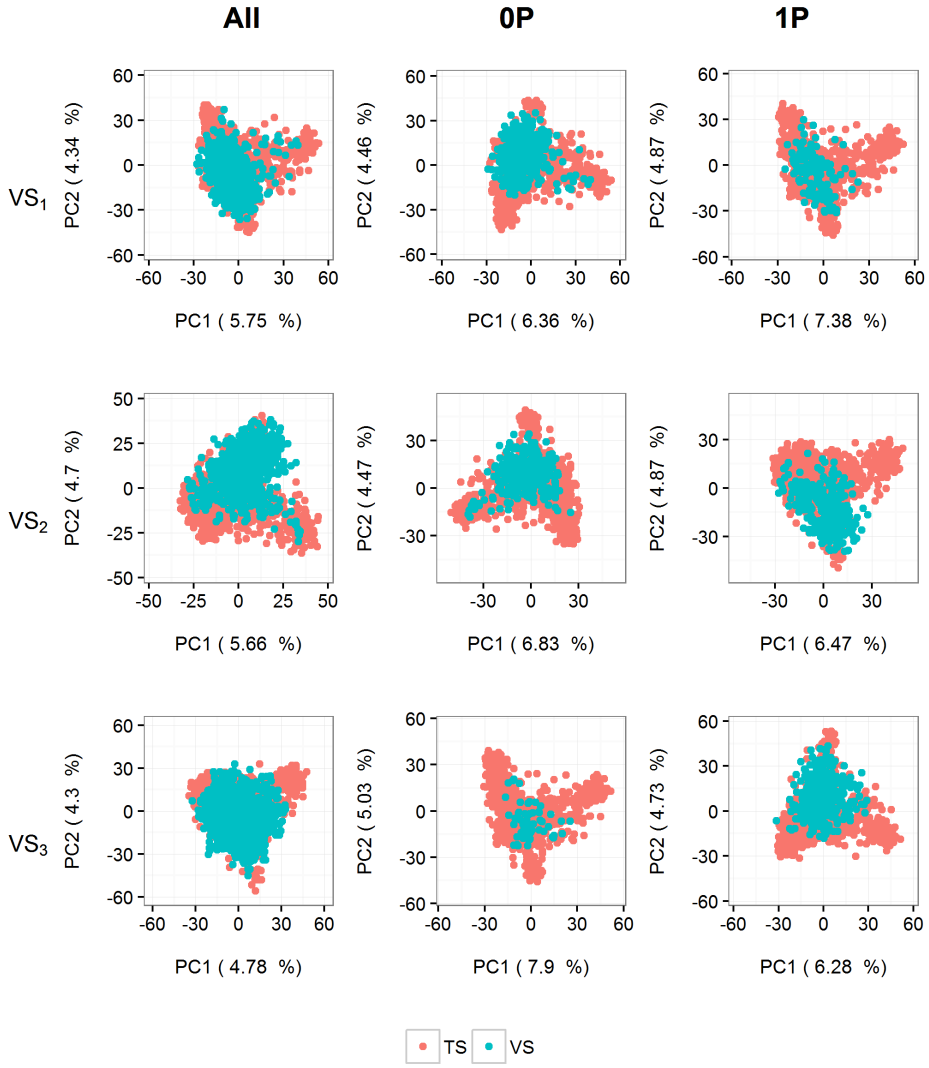


Figure C.14: Principal component (PC) plots for the German and Polish dataset between TS_1 and all VS. TS_1 and relatedness scenarios (All-, 0P- and 1P-scenarios) for VS_1, VS_2 and VS_3.

TS_1:GCA1-2009 + GCA2-2010 + GCA3-2011, VS_1:GCA1-2012, VS_2:GCA1-2013, VS_3:GCA1-2014.

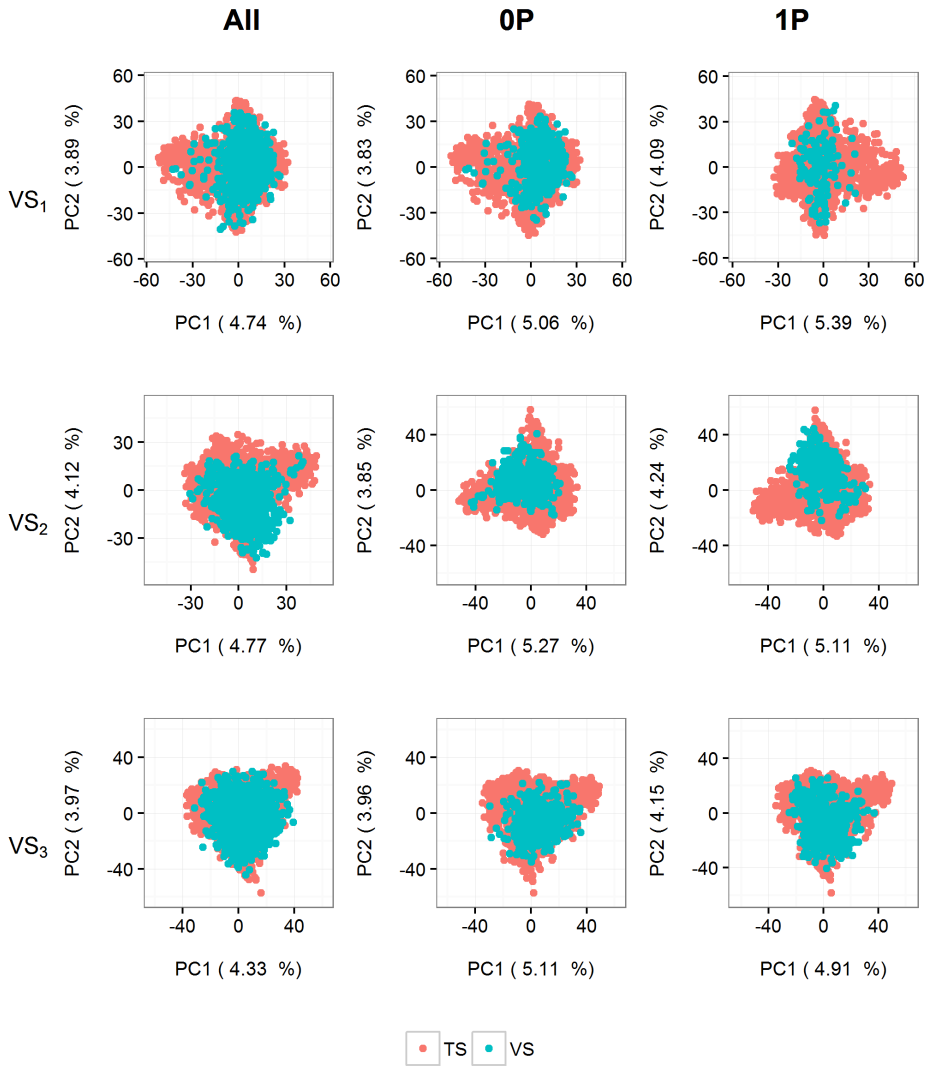


Figure C.15: **Principal component (PC) plots for the German and Polish dataset between TS_2 and all VS.** TS_2 and relatedness scenarios (All-, 0P- and 1P-scenarios) for VS_1, VS_2 and VS_3. TS_2:GCA1-2009 + GCA2-2010 + GCA1-2010 + GCA2-2011, VS_1:GCA1-2012, VS_2:GCA1-2013, VS_3:GCA1-2014.

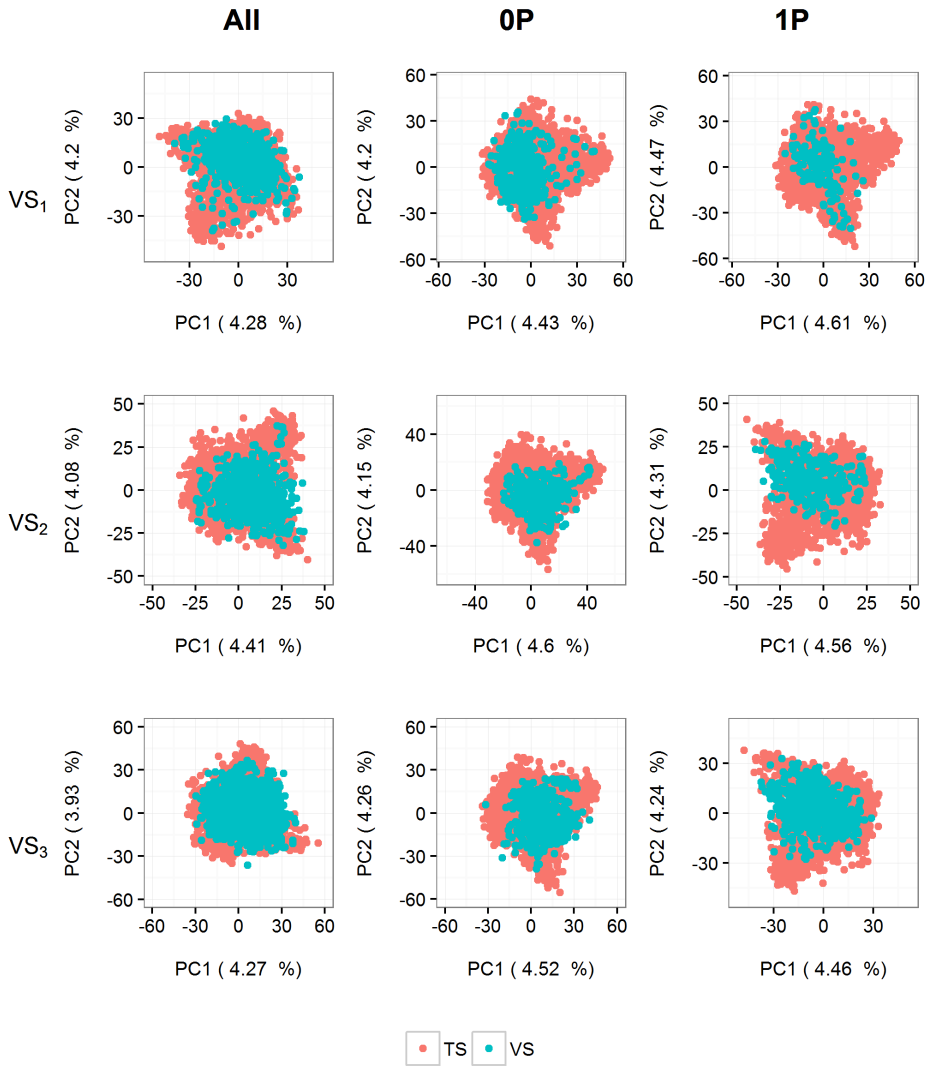


Figure C.16: **Principal component (PC) plots for the German and Polish dataset between TS_3 and all VS.** TS_3 and relatedness scenarios (All-, 0P- and 1P-scenarios) for VS_1, VS_2 and VS_3. TS_3:GCA1-2009 + GCA2-2010 + GCA3-2011 + GCA1-2010 + GCA2-2011 + GCA3-2012 + GCA1-2011 + GCA2-2012 + GCA3-2013, VS_1:GCA1-2012, VS_2:GCA1-2013, VS_3:GCA1-2014.

C.7 Euclidean distance

Table C.11: Means of Euclidean distance between all TS and VS combinations of the three datasets: German (GER), Polish (PL) and German and Polish (GER&PL), with All-, 0P- and 1P-scenarios.

		GER&PL			GER			PL		
		TS_1	TS_2	TS_3	TS_1	TS_2	TS_3	TS_1	TS_2	TS_3
VS_1	All	101.089	101.596	101.847	99.173	99.524	100.114	101.977	102.388	102.533
	0P	100.923	101.551	101.854	98.226	99.073	99.862	102.339	102.616	102.749
	1P	100.080	101.175	101.590	97.247	98.536	99.460	101.774	102.347	102.609
VS_2	All	100.839	101.453	101.764	98.168	98.826	99.624	102.170	102.726	102.883
	0P	100.265	101.187	101.613	96.860	98.365	99.368	102.238	102.654	102.840
	1P	100.657	101.428	101.771	97.543	98.509	99.452	102.321	102.781	102.928
VS_3	All	100.162	100.868	101.222	98.271	98.815	99.511	100.783	101.792	102.039
	0P	99.758	100.944	101.387	95.601	98.076	99.126	101.879	102.358	102.562
	1P	100.417	101.131	101.511	97.548	98.800	99.593	101.687	102.349	102.544