2009

Prediction of cultivar performance and heterogeneity of genotype variance, correlation and error variance in the Iowa Crop Performance Tests−Corn (Zea mays L.)

Yoon-sup So
Iowa State University

Follow this and additional works at: http://lib.dr.iastate.edu/etd

Part of the Agronomy and Crop Sciences Commons

Recommended Citation
http://lib.dr.iastate.edu/etd/10569

This Dissertation is brought to you for free and open access by the Graduate College at Iowa State University Digital Repository. It has been accepted for inclusion in Graduate Theses and Dissertations by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
Prediction of cultivar performance and heterogeneity of genotype variance, correlation and error variance in the Iowa Crop Performance Tests—Corn (Zea mays L.)

by

Yoon-Sup So

A dissertation submitted to the graduate faculty in partial fulfillment of the requirements for the degree of DOCTOR OF PHILOSOPHY

Major: Plant Breeding

Program of Study Committee:
Jode Edwards, Co-Major professor
Kendall R. Lamkey, Co-Major professor
Philip Dixon
Candice Gardner
Jack C. M. Dekkers
Roger Elmore

Iowa State University
Ames, Iowa
2009
# TABLE OF CONTENTS

**LIST OF FIGURES** ........................................................................................................ iii
**LIST OF TABLES** ........................................................................................................ iv
**ACKNOWLEDGMENTS** ............................................................................................... v
**ABSTRACT** .................................................................................................................. vi

**CHAPTER 1. GENERAL INTRODUCTION** ................................................................. 1
  - Introduction ............................................................................................................. 1
  - Thesis Organization ............................................................................................... 6
  - References ............................................................................................................ 6

  - Abstract .................................................................................................................. 9
  - Introduction .......................................................................................................... 10
  - Materials and Methods ....................................................................................... 12
  - Results and Discussion ...................................................................................... 19
  - Conclusions ......................................................................................................... 24
  - References .......................................................................................................... 24

**CHAPTER 3. PREDICTIVE ABILITY ASSESSMENT OF LINEAR MIXED MODELS IN MULTI-ENVIRONMENT TRIALS IN CORN (Zea mays L.)** ........ 33
  - Abstract ................................................................................................................ 33
  - Introduction .......................................................................................................... 34
  - Materials and Methods ....................................................................................... 36
  - Results .................................................................................................................. 43
  - Discussion .......................................................................................................... 48
  - Appendix ............................................................................................................. 53
  - References .......................................................................................................... 56

**CHAPTER 4. GENERAL CONCLUSIONS** ................................................................. 66
  - Conclusions .......................................................................................................... 66
  - Recommendations for Future Research ............................................................. 68
LIST OF FIGURES

Figure 1. Number of hybrids in a data set versus the best model. ................................ 32

Figure 2. Histogram of difference between maximum and minimum \( \text{RMSPD}_{ALL} \) in a data set. ................................................................. 61

Figure 3. Correlation between difference in maximum and minimum \( \text{RMSPD}_{ALL} \) among the 24 models in a data set and difference in a) environment-specific genotype variances, b) genotype correlations among pairs of environments and c) environment-specific error variances from the unstructured model with heterogeneous error variance assumption. ................................................. 62

Figure 4. Correlation between the difference in the maximum and minimum \( \text{RMSPD}_{ALL} \) among the 24 models in a data set from cross-validation analyses and the standard deviation of the differences in pooled GE interaction variance among the 24 models in a data set from full data set analyses. ....................... 63

Figure 5. Correlation between number of common hybrids and difference between maximum and minimum \( \text{RMSPD}_{ALL} \) among models in a data set. ................................. 64

Figure 6. Boxplots for the deviation of pooled GE interaction variance of heterogeneous variance models from that of compound symmetry model based on 200 simulated data. ................................................................. 65
LIST OF TABLES

Table 1. Number of hybrids evaluated in the Iowa Crop Performance Tests–Corn from 1995 to 2005. ................................................................. 27

Table 2. Model description and number of parameter. ........................................ 28

Table 3. Number of datasets with each model as the best fitting model by AIC and BIC. ................................................................. 29

Table 4. Comparisons between models M1-M7 and model M8. ......................... 30

Table 5. Parameters used for simulating unstructured variance covariance structure. 58

Table 6. Number of data sets with each model as the best fitting model by cross-validation, AIC and BIC. ......................................................... 59
ACKNOWLEDGMENTS

I would like to dedicate this space to acknowledge all those who I am deeply indebted for the completion of my Ph.D. study at Iowa State University. My major co-advisor, Dr. Jode Edwards must be recognized for his supervision and advice. There is no doubt that I could not finish the degree without him and he deserves to be appreciated first among all.

I wish to express sincere appreciation to all my committee members, Drs. Kendall Lamkey, Philip Dixon, Candice Gardner, Jack Dekkers, and Roger Elmore. Their expertise in plant and animal breeding and quantitative genetics, maize production and statistics made a great deal of improvement on the quality of the research outcomes.

My special thanks go to Drs. Jean-Luc Jannink and Mike Lee who provided me an opportunity to continue my Ph.D. at Iowa State University. I should not forget to mention Dr. Jim Rouse of Iowa State University Extension who is currently in charge of the Iowa Crop Performance Tests. I learned a lot about the test program through the discussion with Dr. Rouse.

Speaking of the Iowa Crop Performance Tests, I am highly grateful to all those people who were involved in the test program where the data for my research project came from. Their contribution for the test program is greatly appreciated here.

My research assistantship came from the Baker Center of the Agronomy Department. I thank the financial support too.

I also extend my gratitude to all my parents and my family. Thank you Hyunjung and Zed for being my wife and my son.
ABSTRACT

Multi-environment trials generally have highly unbalanced data structures in which a particular cultivar is only observed in a subset of all environments for which data are available. A very common approach to reporting data from such unbalanced data is to subset the data into balanced sets and restrict comparisons within balanced sets. Such an approach results in much information being ignored. In an attempt to make use of all available information, a likelihood-based mixed linear model approach can be chosen since unbalanced data can be analyzed in a straightforward manner. Two studies were undertaken to determine the complexity of heterogeneity of genotype variance, correlation and error variance and to investigate predictive ability of multivariate mixed linear models with varying levels of heterogeneity of those variance components for hybrid performance in unobserved environment in the data sets of the Iowa Crop Performance Tests—Corn. In the first study, a likelihood-based model selection approach identified evidence of heterogeneity of error variances among 58 of 65 single-year and single-district balanced data sets for two model selection criteria, AIC and BIC. Heterogeneity of genotypic variances and correlations between pairs of environments was found in about half of the data sets analyzed. In the second study where two years of data within a district formed 51 highly unbalanced data sets, there was no substantial difference between the best and worst prediction models among all 24 models considered using cross validation, although the best models were generally simpler and parsimonious models. When there was a relatively large difference between the best and worst prediction model, the magnitude of the difference appeared to be highly positively associated with the difference in pooled GE interaction variance among models and to be negatively associated with number of common hybrids between two years in the data sets. There seemed to be a negative association between the difference in pooled GE interaction variance among models and the number of common hybrids in the data sets. A simulation study indicated that the cause of the deviation of pooled GE interaction variance that was obtained from heterogeneous models from that obtained from the homogeneous genotype variance covariance model was due mainly to poor estimation of
some of the variance components by very small number of common hybrids across two years. Because the prediction ability based on an average BLUPs across environments are about the same for models with varying degrees of heterogeneity in genotype variance, correlation and error variance, we may still need to find a statistical model with the best fit of the observed data which would give the most appropriate shrinkage estimator for each environment.
CHAPTER 1. GENERAL INTRODUCTION

Introduction

The Iowa Crop Performance Tests–Corn (Iowa Crop Improvement Association, 2005), conducted annually since 1920, aims to provide growers unbiased information of cultivar performance. During the period of 1995~2005, the annual test had divided the State of Iowa into 7 districts with 3 test locations in each district. Commercial seed companies and Iowa State University entered hybrids into the district(s) appropriate for the hybrid. The majority of hybrids were only entered in the test for one or two years within each district. The lack of common hybrids across years and districts created a highly unbalanced data set for combined analyses, making it hard to provide comparisons among hybrids based on analyses designed for balanced data. The test report during the 11 years only made comparisons among hybrids grown in exactly the same set of environments. This has led to a situation in which almost all comparisons must be based on single-year, single-district estimators even if additional data is available for particular hybrids. The sub setting of the data into single-year, single-district sets results in much potentially relevant data being ignored for many comparisons. In order to make use of all data available, a likelihood-based mixed model approaches seems to be a good choice since with such a model unbalanced data can be analyzed in a straightforward manner (Piepho and Möhring, 2006).

Despite the advantage of likelihood-based mixed model approaches, statistical analyses must also consider inherent and important features of multi-environment trial (MET) data sets: (1) there is the frequent presence of genotype by environment (GE) interactions (Gauch, 1988; Signor et al., 2001), which become of practical significance especially when the GE interactions involve changes in rank of cultivars in different environments, so called crossover interactions (Baker, 1988; Crossa and Cornelius, 1997; Bernardo, 2002), and (2) the statistical assumption in typical MET data analyses, the homogeneity of error and GE interaction variances is often violated (Crossa and Cornelius, 1997; Edwards and Jannink, 2006).
Presence of GE interaction is due to heterogeneity of genotype variance among environments and the lack of perfect correlation of genotype among pairs of environments (Bernardo, 2002; Farconer, 1952). Hence, the aforementioned two features in MET data analyses should not be considered separately. Cooper and DeLacy (1994) stated that “any study which investigates the impact of GE interaction on response to selection should distinguish between these two components”. This statement implies the need of investigating the presence of heterogeneity of genotype variance and correlation among environments in MET data analyses.

In classic quantitative genetics, Falconer and Mackay (1996) indicate “a character measured in two different environments is to be regarded not as one character but as two.” This insight bears an important aspect in investigating GE interactions as a genetic correlation of the trait measured in two different environments (Lynch and Walsh, 1998). Extending it to more than two different test locations, their insight can be viewed in a multivariate mixed model context. Modeling genetic variance covariance structure then becomes central to the analysis of MET data.

Restricted (or residual) maximum likelihood estimation (REML) has been the preferred method in estimation of variance component parameters in MET data analyses especially with unbalanced data sets (Piepho and Möhring, 2006; Holland, 2006). Although efficient use of information from all experimental units when data are unbalanced is one advantage in REML estimation of variance components (Meyer, 1985), Piepho and Möhring (2006) stressed the potential bias of variance component estimates depending on the pattern of missing observation in unbalanced data sets. There are three types of missing-data patterns in unbalanced data sets: (1) missing completely at random, (2) missing at random and (3) missing not at random (Little and Rubin, 2002; Verbeke and Mohlenberghs, 2002). In a simple case of total of $n$ hybrids tested in one location at two years, we can have $n_1$ and $n_2$ missing hybrids in the first and second year, respectively with $n_c$ common hybrids across two years ($n=n_1+n_2+n_c$). When missing observation of $n_1$ and $n_2$ hybrids occurs at random within each year and independently to each other (e.g., hailstorm in the first year and drought in the second year), the missing observations in each year are independent of both data being obtained and being missing.
This is referred to as missing completely at random. In some cases of cultivar evaluation programs, first year entries are subject to selection for second year test. When all \( n \) hybrids are tested in the first year and \( n_1 \) hybrids are selected out, the remaining \( n-n_1 \) hybrids are tested in the second year. A combined data set of this case has \( n_1 \) hybrids missing in the second year. The missing \( n_1 \) hybrids in the second year are based on the performance in the first year hence are dependent of observed data from the first year. This data drop-out pattern is referred to as missing at random and is dependent on observed data but independent of data previously missing if any. In these two cases, a likelihood-based inference is valid (Little and Rubin, 2002). When \( n_1 \) out of \( n \) hybrids are missing in the first year by missing completely at random and all \( n \) hybrids are observed in the second year, one of the common practices in combining data and estimating variance components by analysis of variance method was to drop \( n_1 \) hybrids observed in the second year forming a balanced subset of data with \( n-n_1 \) hybrids across two years. Hence the missing \( n_1 \) hybrids in the second year are dependent on missing data in the first year. This pattern is referred to as missing not at random. Variance component estimates in this case are not valid with respect to all hybrids entered in the test program and \( n_1 \) hybrids observed in the second year can not be ignored in terms of unbiased estimates of variance components (Piepho and Möhring, 2006). Selection creates unbalancedness in combined data sets but the unbalanced structure is not much of a concern for variance component estimates as far as all available data are to be used in mixed linear model analyses (Piepho and Möhring, 2006). The Iowa Crop Performance Tests–Corn receives new hybrids from seed companies each year at the same time some of hybrids tested in one or two previous years are often dropped out of the test program, creating unbalanced data sets when combined. It is not clear if the missing data pattern is by missing completely at random or missing at random since the test program is not the one that performs selection but it is the participating companies who decide which one to drop and which one to enter. The unbalancedness, however, is obviously not because of missing not random which leads to produce biased estimates of variance components (Little and Rubin, 2002).
Nevertheless, behavior of REML estimators of parameter components can still suffer from smaller sample size (Swallow and Monahan, 1984; Holland, 2006). If we are to exploit the presence of heterogeneity of genotype variance, correlation and error variance, unbalanced data sets can pose a problem that not all variance components are estimated with equal precision. In recent years, the Iowa Crop Performance Tests—Corn has seen a decline in number of hybrids entered in for evaluation and there have been as few as 25 common hybrids between 2004 and 2005 in a district. Although REML based mixed linear model has been selected a choice of methods in such MET data analyses (Piepho and Möhring, 2006), there have been very few empirical studies that have examined the impact of such an unbalanced data sets with small common hybrids between some pairs of environments on variance component estimates and predictive ability of mixed linear models with varying degrees of heterogeneity in treatment variance components.

When prediction of hybrid performance is the primary goal in multi-environment trials in plant breeding, predictive precision of yield trials can be improved cost-effectively by effective statistical analysis (Gauch and Zobel, 1988). The usual MET data analysis which has been termed “postdictive” evaluation seeks a statistical model that captures as much of the variation as possible in the observed data (Gauch and Zobel, 1988). They argued that the best postdictively successful model is not usually the best prediction model. Instead, they proposed a model selection criterion for, what they called, “predictive” success based on the assessment of models’ ability for predictions with data not included in modeling with the statement: “Yield trials actually measure past yields on experimental plots, but the purpose is to improve future yields on farmer’s fields.” Data not included in the model construction is considered as simulation of future performance not yet measured and this type of data dropping and evaluation is done through cross validation process.

Many researchers took the idea of Gauch and Zobel (1988) and developed data splitting schemes of their own. Gauch (1988) and Crossa et al. (1991) used a method in which a set of \( m \) random replications of genotypes from a total of \( n \) replications within each environment was selected ignoring blocking scheme of the experiment, leaving the
remaining set of $n-m$ replications of genotypes for validation. Piepho (1994) argued that when the experiment is blocked, the block structure should be preserved in data partition procedure by random selection of the whole block. On the other hand, Gauch and Zober (1996) and Dias and Krzanowski (2003) suggested leave-one-out methods that take only one observation for validation, arguing that the modeling data with $n-1$ data point looses the minimum information from the full set of data with $n$ data point and makes the most efficient use of the full data set. None of the aforementioned studies, however, appear to predict yields of hybrids evaluated in untested or unobserved environments, such as farmer’s fields. Rather they predicted yield response on experimental plots in a way that the results of analyses were to be extrapolated to the entire target environments after the validation of prediction was done on the sample environments. It, however, seems to reflect more of the real world situation that extrapolation of estimators from the analysis on sample environments, hence “predictor” should be done prior to validation on another sample of unobserved environments. This naturally leads us to a data splitting scheme that a set of validation data should include at least one environment worth data set. In this regards, studies reflecting this idea are scarce except one study by Bernardo (1992) in which the arithmetic mean of corn hybrids was compared to two types of weighted mean yields via cross validation that involved a series of 3 to 10 environment data sets assigned as validation data.

In an attempt to make use of all data and avoid the need to subset data in the Iowa Crop Performance Tests—Corn, two research projects were conducted. The first project was carried out to address a question how much heterogeneity exists in genetic variances, error variances and genetic covariances between environments for grain yield. To do so, single-year and single-district data sets were taken to form balanced data sets in order to avoid confounding effects of potential bias in variance component estimates from combined unbalanced multi-year and multi-district data sets. The second project was carried out with the following objectives 1) to assess predictive ability of multivariate mixed models with varying degrees of heterogeneity of genotype variance, correlation, and error variance via cross validation for the prediction of hybrid performance in unobserved environments, and 2) to investigate the impact of small number of common
hybrids on variance component estimates and the prediction ability of the models using simulation. Based on the findings from the two research projects, an attempt was also made for a general recommendation for choice of a best prediction model for long term use in the Iowa Crop Performance Tests—Corn and for future research in such MET data analysis in general.

**Thesis organization**

The following two chapters follow the University requirements for journal papers. Chapter 2 entitled “A comparison of mixed model analyses of the Iowa Crop Performance Tests—Corn (*Zea mays* L.)” has been accepted for publication in Crop Science and is in press. Chapter 3 entitled “Predictive ability assessment of linear mixed models in multi-environment trials in corn (*Zea mays* L.)” has been prepared for publication in Crop Science. The two chapters were modified for insertion in this dissertation to meet the format requirements by Iowa State University.

**References**


CHAPTER 2. A COMPARISON OF MIXED MODEL ANALYSES OF THE IOWA CROP PERFORMANCE TESTS—CORN (Zea mays L.)

Modified from a paper to be published in *Crop Science*

Yoon-Sup So and Jode Edwards

Abstract

Multi-environment trials generally have highly unbalanced data structures in which a particular cultivar is only observed in a subset of all environments for which data are available. A very common approach to reporting data from such unbalanced data is to subset the data into balanced subsets and restrict comparisons within balanced subsets. Such an approach results in much information being ignored. We undertook an empirical study of 65 individual sets of data from the Iowa Crop Performance Tests—Corn to compare eight different mixed linear models in order to determine what features in the data need to be considered in developing approaches to make use of all available information. We used a model selection approach to identify the best model based on the presence or absence of heterogeneity of error variances among environments, heterogeneity of genotypic variances among environments, and heterogeneity of genotypic correlations between pairs of environments. The trait analyzed was grain yield. We found evidence of heterogeneity of error variances among locations in 58 of 65 sets of data for two model selection criteria. Heterogeneity of genotypic variances and correlations between pairs of environments was found in about half of the sets we analyzed. A general recommendation for model selection cannot be made from this analysis. In general we found that heterogeneity of variances and correlations was prominent in many datasets. Identification of the best statistical model for a particular dataset may be dependent on application of a model selection approach.
Introduction

The Iowa Crop Performance Tests—Corn (Iowa Crop Improvement Association, 2005) is conducted in order to provide unbiased third party information on commercial hybrid performance to growers. Until 2005, the annual test divided the State of Iowa into 7 districts with 3 test locations in each district. Because of the short time that a given hybrid is sold commercially, the majority of hybrids are only entered in the test for one or two years. Additionally, there are few common hybrids across districts. The lack of common hybrids across years and districts creates a highly unbalanced data set increasing the difficulty in providing comparisons among hybrids based on analyses designed for balanced data. The approach taken in the Iowa Crop Performance Test has been to report only comparisons among hybrids grown in exactly the same set of environments. This has led to a situation in which almost all comparisons must be based on single-year, single-district estimators even if additional data is available for particular hybrids. The subsetting of the data into single-year, single-district sets results in much potentially relevant data being ignored for many comparisons.

From a purely statistical perspective, it is hard to imagine any situation in which ignoring information leads to a better estimator than using all available information. There are many inherent statistical features of multi-environment trial data sets that may make it difficult to use all available information under classical models developed for balanced data: (i) data are often incomplete (not all cultivars grown in all trials) by selection and unbalanced due to missing observations by unforeseen circumstances (Kelly et al., 2007; Spilke et al., 2005), (ii) there is the frequent presence of genotype by environment (GE) interactions (Gauch, 1988; Signor et al., 2001), which become of practical importance especially when the GE interactions involve changes in rank of cultivars, so called crossover interactions, in different environments (Baker, 1988; Crossa and Cornelius, 1997; Bernardo, 2002), and (iii) the statistical assumption of homogeneity of variance for random effects is often violated (Crossa and Cornelius, 1997; Edwards and Jannink, 2006). Given the typical statistical complexities of multi-environment trial data, there are two options available to the analyst, i.) Subset the data into balanced subsets and only report comparisons within balanced subsets, or ii.) Choose a statistical
approach to make comparisons between hybrids with unequal information (e.g., not necessarily tested in exactly the same environments). The Iowa Crop Performance Tests—Corn has chosen the first approach, where all comparisons are confined to balanced subsets of the data. From a statistical perspective, if in fact an appropriate statistical method can be identified to provide unbiased comparisons from unequally tested cultivars, the second approach could produce more precise comparisons than the first. Many statistical methods for the analyses of multi-environment trial data have been proposed that do not depend on balanced data (Smith et al., 2005). Various methods rely on varying assumptions and variance-covariance structures in the data. Hence, the choice of the best model depends on the particular structure and statistical properties of the data, and a statistical approach to selecting the best model (Gurka, 2006; Kelly et al., 2007; Smith et al., 2005; West et al., 2007). Cooper and Delacy (1994) observed that a great deal has been written about analysis of multi-environment yield trials from a theoretical perspective. However, very little has been done to compare broad classes of models empirically. In order to explore what approaches might be appropriate for combining all available information in unbalanced sets of Iowa Crop Performance Test data for corn, we first need to understand what complexities in the data, other than the unbalanced nature of the data, need to be considered in choosing and testing potential models for highly unbalanced sets of data. For example, if data within districts has little heterogeneity of variance or covariances, methods for subdivided target regions (Piepho and Mohring, 2005; Atlin et al., 2000) could be applied quite readily with little need for modeling heterogeneity of covariance structure within individual regions (in our case districts). However, if there is substantial heterogeneity within districts, other approaches may be needed. We undertook a survey of eleven years of data collected in the Iowa State Crop Performance Test for corn with the specific objectives of (i) to determining how much heterogeneity exists in genetic variances, error variances, and genetic covariances between environments for grain yield and (ii) to use model selection techniques to determine if any general recommendations could be made for choice of a best statistical model for describing data collected in the Iowa Crop Performance Test.
Materials and Methods

**Iowa Crop Performance Test:** Yield data from the Iowa Crop Performance Tests – Corn from 1995 to 2005 were used in this study. Commercial seed companies and Iowa State University entered hybrids into the district(s) appropriate for the hybrid. The number of hybrids tested varied among districts and years because it was up to the companies to decide what hybrids to enter in which districts (Table 1). The evaluation was conducted in an $\alpha$-lattice design from 1995 to 2003 and in a row-column design during the last two years with 4 replications across all trials. Each entry was machine-planted in a four-row plot and the center two rows of each plot were collected for grain yield. Yields were adjusted to 15.5 percent moisture content and converted to the metric unit of ton per hectare for this study. Details on the test locations and general management practice can be found at http://www.croptesting.iastate.edu.

Each analysis was restricted to a single year, single district combination with three locations in the district in order to obtain precise covariance estimates. Had data been combined across districts or across years, in most possible subsets of data there were very few common hybrids across years or across districts, meaning that covariances between environments in different years or different districts would have been estimated with very low precision. Thus, we chose to restrict these analyses to balanced subsets in order to focus our study on heterogeneity of covariance components and avoid potential confounding effects of unbalanced data on our conclusions. There were total of 77 separate data sets (11 years x 7 districts) but 12 data sets were discarded because they only had yield data from 2 locations available, leaving 65 data sets with three locations of data for analysis in this study.

**Statistical models:** The statistical models evaluated in this study were based on a mixed linear model of the form:

$$Y_{erbg} = \mu_e + R_{re} + B(R)_{bre} + G_{ge} + \epsilon_{erbg},$$

(1)

where $Y_{erbg}$ = grain yield of $g^{th}$ genotype in $b^{th}$ lattice block of $r^{th}$ replication at $e^{th}$ environment, $\mu_e$ is the mean of $e^{th}$ environment, $R_{re}$ is the effect of $r^{th}$ replication at $e^{th}$ environment, $B(R)_{bre}$ is the effect of $b^{th}$ lattice block within $r^{th}$ replication at $e^{th}$ environment.
environment, $G_{ge}$ is the effect of $g^{\text{th}}$ genotype at $e^{\text{th}}$ environment and $e_{rbg}$ is the residual associated with $e^{\text{th}}$ environment, $r^{\text{th}}$ replication, $b^{\text{th}}$ lattice block and $g^{\text{th}}$ genotype. For the years 2004-2005, an additional term was added to the linear model to account for lattice blocks in two directions, rows and columns. Environment means, replication effects, and lattice block effects were considered fixed, whereas genotype effects ($G_{ge}$) and residuals ($e_{rbg}$) were considered random and normally distributed. The data were nearly balanced with respect to environments and replications (the only imbalance resulting from missing data), so inferences on genotypes and their (co)variance components would have been essentially the same whether these variables were fixed or random. Thus we chose fixed to simplify computation and avoid issues with precision of estimating variance components with these effects given that there were very few environments and replications for obtaining precise variance component estimates. In the case of lattice blocks, fitting lattice blocks as random effects would have resulted in some recovery of interblock information. However, recovery of interblock information is also dependent on estimation of the variance of block effects. In likelihood-based mixed model approaches, it is almost impossible to account for the error of estimation of variance components associated with block effects because distributions of variance component estimators are only known in very simple cases (Searle et al., 1992). Thus, we chose to fit blocks as fixed effects in our work to avoid introducing an unquantifiable source of error due to nuisance parameters.

We compared models in which error variances were allowed to be heterogeneous among environments ($\sigma_{e_1}^2 \neq \sigma_{e_2}^2 \neq \sigma_{e_3}^2$) and in which error variances were homogeneous ($\sigma_{e_1}^2 = \sigma_{e_2}^2 = \sigma_{e_3}^2$). The genotype effects for the $g^{\text{th}}$ cultivar in each of the three environments, $G_{g1}$, $G_{g2}$, and $G_{g3}$, were assumed to be random variables with a multivariate normal distribution with a $3 \times 3$ variance-covariance matrix $\Sigma$ of the general form:
\[
\sum_{G_{h3}} = \begin{bmatrix}
\sigma^2_{G_1} & \sigma_{G_12} & \sigma_{G_13} \\
\sigma_{G_12} & \sigma^2_{G_2} & \sigma_{G_23} \\
\sigma_{G_13} & \sigma_{G_23} & \sigma^2_{G_3}
\end{bmatrix}.
\]

The three diagonal elements \((\sigma^2_{G_1}, \sigma^2_{G_2}, \sigma^2_{G_3})\) are genotypic variances of cultivars within environments and off-diagonal elements \((\sigma_{G_{12}}, \sigma_{G_{13}}, \sigma_{G_{23}})\) are the genotypic covariances between pairs of environments. The combined variance-covariance matrix of all genotype effects for the three locations and all genotypes in a district was a block diagonal matrix of the form \([\textbf{I}_g \otimes \sum_{G_{h3}}]\), where \([\textbf{I}_g]\) is an \(g\)-dimensional identity matrix and \(\otimes\) is the Kronecker product of two matrices. The block diagonal structure results from the assumptions that all cultivars had identical variance-covariance matrices and that zero covariances between all pairs of cultivars were assumed.

Based on the general form of the variance-covariance matrix for the three environments, we examined four different variance-covariance structures based on different levels of restrictions on the six parameters in the 3 x 3 variance-covariance matrix. The unstructured variance-covariance matrix \(\sum_{G_{h3}}\) with distinct genotypic variances in each environment \((\sigma^2_{G_1}, \sigma^2_{G_2}, \sigma^2_{G_3})\) and three distinct genotypic covariances between the three possible pairs of environments \((\sigma_{G_{12}}, \sigma_{G_{13}}, \sigma_{G_{23}})\) was referred to as the unstructured model (SAS Institute Inc., 2004) as was the most general. From the unstructured model, three reduced models were examined, each with different restrictions that reduce the total number of genotypic variance-covariance parameters to less than the six distinct parameters in the unstructured variance-covariance matrix.

The first restriction examined was homogeneous genotypic variances across environments with heterogeneous correlations between pairs of environments \((\sigma^2_{G_1} = \sigma^2_{G_2} = \sigma^2_{G_3} = \sigma^2_G)\):

\[
\sum_{G_{h3}} = \sigma^2_G \begin{bmatrix}
1 & \rho_{12} & \rho_{13} \\
\rho_{12} & 1 & \rho_{23} \\
\rho_{13} & \rho_{23} & 1
\end{bmatrix}.
\]
This structure was referred to as a heterogeneous correlation structure because the genotypic variances were constant among environments, but correlations (and by direct extension, covariances) between pairs of environments remained heterogeneous. With one genetic variance and three correlations, this variance-covariance matrix has four parameters. The third structure was heterogeneous compound symmetry (SAS Institute Inc., 2004) which had constant correlations among the three pairs of environments, but heterogeneous genotypic variances. This model also had four parameters. The specific form of the covariance matrix is:

\[
\Sigma_{G_{ns}} = \begin{bmatrix}
\sigma^2_{G_1} & \sigma_{G_1}\sigma_{G_2}\rho & \sigma_{G_1}\sigma_{G_3}\rho \\
\sigma_{G_1}\sigma_{G_2}\rho & \sigma^2_{G_2} & \sigma_{G_2}\sigma_{G_3}\rho \\
\sigma_{G_1}\sigma_{G_3}\rho & \sigma_{G_2}\sigma_{G_3}\rho & \sigma^2_{G_3}
\end{bmatrix}.
\]

The correlation between pairs of environments, \( \rho \), was held constant among all pairs of environments, but variances were allowed to be heterogeneous among environments. Covariances between environments were thus expressed as functions of genotypic standard deviations and the constant correlation between environments. This model had a total of four parameters, the common correlation and three variances.

The last structure of interest was compound symmetry (SAS Institute Inc., 2004) in which both genotypic variances and covariances between pairs of environments were homogeneous:

\[
\Sigma_{G_{ns}} = \begin{bmatrix}
\sigma^2_G & \sigma & \sigma \\
\sigma & \sigma^2_G & \sigma \\
\sigma & \sigma & \sigma^2_G
\end{bmatrix}.
\]

The compound symmetry structure was of interest because it is equivalent to the classical linear model with an average genotypic effect and a genotype x environment interaction effect as in the linear model (Littell et al., 1998):

\[
Y_{erg} = \mu + R_{xe} + B(R)_{bre} + G_g + GE_{ge} + \varepsilon_{erg}
\]

where the term \( G_{ge} \) in linear model equation (1) is replaced with the terms \( G_g + GE_{ge} \) in linear model equation (2) (historically, the substitution was made in reverse, \( G_g + GE_{ge} \)).
has been replaced by $G_{ge}$ with development and application of multivariate mixed linear models). If the terms $G_g$ and $GE_{ge}$ in model (2) were assumed to be random variables and normally distributed with variances $V(G_g) = \sigma_{GU}^2$ and $V(GE_{ge}) = \sigma_{GEU}^2$ (U in the subscripts denotes the univariate mixed linear model equation (2)), then covariances between replicate observations on the same genotype were:

$$\text{Cov}(Y_{ger}, Y_{ger'}) = \sigma_{GU}^2 + \sigma_{GEU}^2 = \text{covariance between observations from the same environment}$$

$$\text{Cov}(Y_{ge}, Y_{ge'}) = \sigma_{ge}^2 = \text{covariance between observations from different environments}$$

Expressions for the same covariances with compound symmetry covariance structure in linear model equation (1) were:

$$\text{Cov}(Y_{ger}, Y_{ger'}) = \sigma_G^2 = \text{covariance between observations from the same environment}$$

$$\text{Cov}(Y_{ge}, Y_{ge'}) = \sigma = \text{covariance between observations from different environments}$$

With regard to random effects, models (1) and (2) are thus defined as equivalent models, meaning they describe same variance-covariance structure among observations with the following equivalencies among parameters: $\sigma = \sigma_G^2$, $\sigma_G^2 = \sigma_{GU}^2 + \sigma_{GEU}^2$, and thus $\sigma_{GEU}^2 = \sigma_G^2 - \sigma$ (Henderson, 1984).

Notation and names for eight mixed linear models resulting from linear model equation (1) with four genotypic covariance structures with two error-variance assumptions, homogeneous and heterogeneous among locations, are summarized in Table 2. Names of genotypic covariance structures were chosen to be consistent with those used in the MIXED procedure of SAS (SAS Institute, 2004) with the exception of the heterogeneous correlation model.

**Analysis procedures:** A total of 65 data sets were analyzed with the eight models for a total of 520 total analyses using the MIXED procedure in SAS version 9.1.3 (SAS
Institute Inc., 2004). The heterogeneous (environment-specific) error variance structure was achieved using a GROUP option in the REPEATED statement. Genotypic variance-covariance structures were specified with the TYPE option in the random statement. The types unstructured, heterogeneous compound symmetry, and compound symmetry are built into the procedure, whereas for heterogeneous correlation structure, type “LIN(q)” was used with input of the variance-covariance matrix. The heterogeneous correlation structure was named heterogeneous correlation in part for convenience of description and notation, but it should be pointed out that the model was parameterized in the MIXED procedure using heterogeneous covariances. Hybrid cultivars were designated as a subject effect in the random statement, which creates the block diagonal structure in the variance-covariance matrix of all random effects. Variance components were estimated by restricted maximum likelihood (REML) with solutions obtained by Newton-Raphson iteration in the MIXED procedure.

If the Newton-Raphson iteration in the MIXED procedure did not converge or resulted in an infinite likelihood, alternate starting values were provided. If new starting values did not solve the problem, the iteration method was changed to the Fisher scoring algorithm instead of ridge-stabilized Newton-Raphson algorithm (SAS Institute Inc., 2004). If new starting values or change of iterative method did not result in convergence to a non-infinite likelihood solution, covariance structures other than the unstructured model were discarded for the data set. We also discarded models whose estimated genotype variance covariance matrix was not positive definite. For the unstructured covariance structure, two alternative methods of specifying the unstructured covariance matrix were attempted. First, an unstructured correlation matrix was attempted because the unstructured correlation structure in the MIXED procedure constrains correlations to $-1 < \rho_{ij} < 1$ ($i \neq j$) for all correlations, whereas the unstructured covariance matrix can result in estimates outside the parameter space (SAS Institute Inc., 2004). If the unstructured correlation matrix did not converge to a finite likelihood and a positive semi-definite covariance matrix, a factor analytic decomposition of the covariance matrix was used, specified by TYPE=FA0(3), in the MIXED procedure (SAS Institute Inc., 2004). The factor analytic structure is a mixed model version of the additive main effects
and multiplicative interaction model in which the unstructured covariance matrix is represented by a Cholesky decomposition which guarantees that the estimated covariance matrix is at least positive semi-definite (Kelly et al., 2007; Piepho, 1998; Crossa et al., 2006; Smith et al., 2001).

**Model comparison:** The 8 models in each dataset were compared using AIC and BIC:

\[
AIC = (-2 \times \text{REML Loglikelihood}) + 2p,
\]

\[
BIC = (-2 \times \text{REML Loglikelihood}) + (p \times \ln(n)),
\]

where \(p\) is the effective number of estimated parameters and \(n\) is the effective number of observations used in estimation of the model. The effective number of estimated parameters \((p)\) is the total number of parameters to be estimated (number of parameters in \(\Sigma_{G_{ij}}\) and residual error parameters) minus the number of parameters with estimates on a boundary constraint, e.g., estimated variance component equal to zero. The effective number of observations \((n)\) equals the level of genotype effect in our example. In general, when the SUBJECT option is used in RANDOM statement to construct a block diagonal structure, the effective number of observations \((n)\) is the dimension of the effect set in SUBJECT option (SAS Institute Inc., 2004). Both statistics are in the smaller-is-better form. The “2\(p\)” in AIC and “\(p \times \ln(n)\)” in BIC can be viewed as a penalty that is applied to a comparison of models with different number of parameters and (or) different number of observations.

**Outlier elimination and computational issues:** The REML-based estimation of variance components is known to be sensitive to outliers (West, 2007; SAS Institute Inc., 2004). Prior to main analyses, all data sets were analyzed using the unstructured genotype covariance and heterogeneous error variance model, \(\text{M1}(\sigma_G^2, \rho_G, \sigma_e^2)\). Studentized residuals were plotted using quantile-quantile (QQ) plots to visually check the distribution of residuals. In addition, \(p\)-values for studentized residuals were computed from a \(t\)-distribution with degrees of freedom being equal to the error degrees of freedom and were adjusted using a Bonferroni correction based on the total number of observations in the data set (individual plots in this case, not genotypes as in the penalty
in BIC). An adjusted \( p \)-value of 0.03 was chosen for removing outliers and was checked visually by examining QQ plots. There were 69 outliers removed out of a total of 132,443 observations (0.052%).

**Results and Discussion**

The number of hybrids evaluated in each trial ranged from 90 to 256 (Table 1). Estimated variance components by the model with the classic statistical assumption (=M8) had an average of 0.240 (range=0.034–0.942) for the genotype variance (\( \sigma_{G}^{2} \)), an average of 0.151 (\( \sigma_{G/E}^{2} \), range=0.045–0.505) for the GE interaction variance and an average of 0.584 (\( \sigma_{e}^{2} \), range=0.333–1.486) for the error variance.

In four cases, we obtained infinite likelihoods which required the addition of staring values with the PARMS statement to obtain solutions with non-infinite likelihoods. M2(\( \rho_{G}, \sigma_{e}^{2} \)), M5(\( \sigma_{G}^{2}, \rho_{G} \)) and M6(\( \rho_{G} \)). In an additional 21 analyses in models containing heterogeneous correlations between environments, we obtained non-positive definite estimators of the genotypic covariance matrix, \( \Sigma_{G} \). In the case of models M1(\( \sigma_{G}^{2}, \rho_{G}, \sigma_{e}^{2} \)) (4 analyses with non-positive definite matrices) and M5(\( \sigma_{G}^{2}, \rho_{G} \)) (11 non-positive definite matrices), which had unstructured genotypic covariance matrices, this problem could be solved by replacing the unstructured genotypic variance-covariance matrix with a factor analytic decomposition (type=FA0). In models M2 (\( \rho_{G}, \sigma_{e}^{2} \)) (2 cases of non-positive definiteness) and M6 (\( \rho_{G} \)) (4 cases of non-positive definiteness), which had heterogeneous correlations but not genetic variances, the factor analytic decomposition could not be used, so fit statistics from these analyses were not used in model comparisons.

In a few cases, the genetic model converged to a solution with an estimated genetic variance of zero, but in a majority of analyses that produced a non-positive definite \( \Sigma_{G} \) had one or more estimated genetic correlation coefficients between environments greater than one. Estimated correlation coefficients greater than one are
to be expected in relatively homogeneous target regions in which correlations between environments are expected to be close to one. In such cases, the use of factor analytic decomposition of the covariance matrix which is mathematically restricted to have estimators within the parameter space will likely be a necessity to avoid these issues. The occurrence of non-positive definite genetic covariance matrix is a common occurrence (Hill and Thompson, 1978; Demidenko, 2004). Nevertheless, this has been rarely discussed in likelihood-based mixed model approach to plant breeding trial data analyses. Holland (2006) reported estimated genotypic correlation greater than one in his simulation study on multivariate REML estimation of genotypic correlation via SAS PROC MIXED. Kelly et al. (2007) in their study on the accuracy of varietal selection indicated difficulties in fitting an unstructured model with 80 genotypes in a simulation study using the ASReml software. Positive definiteness of the genotypic variance-covariance matrix is only a concern for the optimization algorithms that make use of the information matrix from an inverse of the second derivative (=Hessian matrix) of the objective function from REML (Demidenko, 2004). The ridge-stabilized Newton-Raphson (observed information matrix) and the Fisher scoring (expected information matrix) algorithm used in the MIXED procedure in SAS and the Average Information (average of observed and expected information matrix) algorithm by ASReml use the inverse of the Hessian. However, convergence is still possible with a non-positive definite covariance matrix in the expectation-maximization (EM) algorithm (Demidenko, 2004; Gilmour et al., 2006). As implemented by Kelly et al. (2007), ASReml has an option to invoke the expectation-maximization during the iteration process whenever the Average Information algorithm produces estimates outside the parameter space. However, the use of the EM algorithm may not be the best solution, because convergence can be extremely slow. In the MIXED procedure in SAS, which does not have an EM option, an unstructured covariance matrix can be reparameterized using a factor analytic parameterization (TYPE=FA0(e), where e is the dimension of the matrix). The factor analytic parameterization guarantees the estimate of $\sum_{G_{x3}}$ will be at least positive semi-definite. Unfortunately, the factor-analytic reparameterization was only possible for the
unstructured model in our work and not for the other three models in cases of non-positive definite genotypic variance-covariance matrices.

**Comparison among models:** Model M8, which represented the classical assumption of homogeneous genotypic, genotype x environment interaction, and error variances was identified as the best model 7 times by BIC and not identified as the best model in any of the 65 datasets analyzed according to AIC (Table 3). In other words, by either criterion, in at least 58 of 65 cases, the best model was a model with some level of heterogeneity in the covariance structure. The most important level of heterogeneity was the error variance judging by the fact that the best model included homogeneous error variances among environments in only 6 datasets according to AIC and 15 of 65 datasets according to BIC (Table 3). Overall, the models identified as the best model in the largest number of datasets were model M1($\sigma^2_G, \rho_G, \sigma^2_\varepsilon$), which allowed all components to be heterogeneous, by AIC (19 data sets - 29.2%) and M4($\sigma^2_\varepsilon$), which included heterogeneity of error variance only, by BIC (24 data sets – 36.9%). Model M4($\sigma^2_\varepsilon$), the overall best fit model by BIC, was the second best model by AIC (17 data sets – 26.2%). Considering that the best overall model for either criterion was the best model in at most 36.9% of datasets, no single model could be identified as the best overall model. However, considering that the vast majority of cases included heterogeneous error variances, we might possibly conclude that analysis of corn yield trial data sets like those we analyzed should include heterogeneous error variances. As a means of quantifying differences among models, we used the simplest mode, M8, as a control and computed the average difference in likelihood, AIC, and BIC between the remaining seven models and model M8 (Table 4). We then computed a probability of obtaining a value from a chi-square distribution with degrees of freedom equivalent to the difference in number of parameters between each of models M1 through M7 and model M8 and the average difference in the likelihood. This constituted a rough ‘average’ of the significance of the difference between each of the seven models containing heterogeneity and the simplest model. The four models containing heterogeneous error had stunningly small p-values, all of them less than 1.1E-13. Such small values have little quantitative meaning, other
than to suggest these are very large differences in the likelihood. The likelihood differences for models with homogeneous error variance were much less, but the average difference in likelihood values still corresponded to highly significant differences suggesting the models with heterogeneous (co)variance components were on average much better than models with homogeneous components.

Because of the large influence of the error-variance assumption, i.e., heterogeneous versus homogeneous error variances, genotypic variance-covariance structures were compared with the error-variance assumption restricted to either homogeneous or heterogeneous among environments in order to provide a better comparison among genotypic variance-covariance matrix structures. If the error variance was assumed to be heterogeneous a priori, (models M1-M4, Table 3), the unstructured model was identified as the best model in the most cases according to AIC (23 data sets – 35.4%) and compound symmetry according to BIC (30 data sets – 46.2%). Very similar results were observed under the homogeneous error variance assumption with the unstructured model being the best model according to AIC and the compound symmetry the best model according to BIC. Restricting comparisons according to the error variance assumption did not provide any clear patterns in terms of choosing a best overall genotypic variance-covariance structure, nor did a clear pattern emerge as to whether heterogeneity in correlations or genetic variances was more commonly a feature of the best model. According to AIC with heterogeneous error variances, the best model included heterogeneous genetic variances in 39 cases and heterogeneous correlations in 32 cases (Table 3). Thus, these analyses suggest that in multi-environment yield trial analysis for corn hybrids in Iowa, heterogeneous error variance is a feature very likely to be important. Heterogeneity of genotypic components may be important, but it is not possible from our analyses to conclude generally that such heterogeneity will be important, or that including heterogeneity at either level would necessarily be required in all circumstances. If one desires the best statistical model, a model selection approach considering the levels of heterogeneity we evaluated would be the best course of action. However, in utilizing model selection approaches, two additional issues arise, namely, the power of detection of particular features, such as heterogeneity of error variance, and
choice of model selection criteria. We used two model selection criteria, and consistent with what is known about them theoretically, we found that BIC identified simpler models and AIC identified more complex models (Akaike, 1973; Gurka, 2006). The BIC depends on the number of observations, or in this context, the number of hybrids in the analysis, raising the question as to whether more or less complex models were found by either criterion with larger or smaller datasets. The number of hybrids in each dataset versus the best identified model is shown in Figure 1, in which there does not seem to be a strong relationship for either information criteria.

This work was undertaken in part to determine what levels of heterogeneity necessarily should be included in a more comprehensive study of analytical approaches to obtain the best predictions of hybrid performance. The simple answer to this question is that all levels of heterogeneity of variances and covariances need to be addressed in future evaluations of prediction models and experimental design for hybrid evaluation trials. The present work does not address the important question of which model will provide the best predictions of hybrid performance. The best statistical model according to a statistical test or model selection criterion may not necessarily produce the best prediction of performance in a future environment. Our longer term objective is to develop analytical strategies for the Iowa Crop Performance test that make use of all available information for all comparisons rather than relying on sub setting the data into many individual analyses that ignore large portions of available data. Given that the average genetic variance was 0.24, or just 25% of the total variance on a plot basis (total variance = 0.97), the relative signal is fairly small, underscoring the importance of effectively using all available information. The results of the present study suggest that any analytical approaches to make use of all available information will need to have two features. First, heterogeneity of variances and covariances will need to be evaluated at all levels, and secondly, a single-best approach may not exist, but rather, we may very well need to develop model selection approaches based on cross-validation of predictions to choose the best model for particular datasets.

Much of the theoretical and analytical work preceding the present paper have clearly established that heterogeneous variances may be important. Nonetheless,
classical approaches, in which homogeneity of variance is assumed, are still too often the starting point, especially among academic researchers and introductory plant breeding courses. This approach is backwards. We argue that the statistical tools are available in which plant breeders should start with the assumption that (co)variances may be heterogeneous and use the tools of statistics to determine whether or not the heterogeneity may or may not be ignored.

Conclusions

We found overwhelming evidence of the need to include heterogeneous error variances among locations in corn hybrid performance evaluation trials in Iowa. With respect to genotypic components of variance, we found that the choice of the best statistical model varied with datasets and information criteria such that one cannot conclude generally whether heterogeneity of genotypic (co)variance components should be included in analyses of multi-environment trial data. If one desires the best statistical model, an appropriate model selection approach should be chosen.

References


Table 1. Number of hybrids evaluated in the Iowa Crop Performance Tests—Corn from 1995 to 2005.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>196</td>
<td>169</td>
<td>144</td>
<td>156</td>
<td>156</td>
<td>144</td>
<td>121</td>
<td>110</td>
<td>110</td>
<td>110</td>
<td>100</td>
</tr>
<tr>
<td>2</td>
<td>196</td>
<td>169</td>
<td>132</td>
<td>132</td>
<td>†</td>
<td>132</td>
<td>110</td>
<td>110</td>
<td>110</td>
<td>110</td>
<td>110</td>
</tr>
<tr>
<td>3</td>
<td>256</td>
<td>240</td>
<td>196</td>
<td>†</td>
<td>210</td>
<td>210</td>
<td>169</td>
<td>156</td>
<td>144</td>
<td>140</td>
<td>100</td>
</tr>
<tr>
<td>4</td>
<td>196</td>
<td>169</td>
<td>144</td>
<td>156</td>
<td>182</td>
<td>182</td>
<td>144</td>
<td>132</td>
<td>132</td>
<td>99</td>
<td>100</td>
</tr>
<tr>
<td>5</td>
<td>256</td>
<td>225</td>
<td>169</td>
<td>169</td>
<td>196</td>
<td>169</td>
<td>156</td>
<td>144</td>
<td>156</td>
<td>90</td>
<td>90</td>
</tr>
<tr>
<td>6</td>
<td>144</td>
<td>121</td>
<td>169</td>
<td>144</td>
<td>156</td>
<td>132</td>
<td>132</td>
<td>132</td>
<td>132</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>7</td>
<td>169</td>
<td>144</td>
<td>156</td>
<td>156</td>
<td>132</td>
<td>132</td>
<td>132</td>
<td>132</td>
<td>132</td>
<td>90</td>
<td>90</td>
</tr>
</tbody>
</table>

† Data sets in the empty cells were not used since yield data were only available from 2 locations.
Table 2. Model description and number of parameter.

<table>
<thead>
<tr>
<th>Genotypic Covariance structure</th>
<th>Genetic Variances</th>
<th>Genetic Correlations</th>
<th>Model designation</th>
<th>Genotypic parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unstructured</td>
<td>Heterogeneous</td>
<td>Heterogeneous</td>
<td>M1($\sigma_G^2, \rho_G, \sigma_e^2$)</td>
<td>6</td>
</tr>
<tr>
<td>Heterogeneous correlation</td>
<td>Homogeneous</td>
<td>Heterogeneous</td>
<td>M2($\rho_G, \sigma_e^2$)</td>
<td>4</td>
</tr>
<tr>
<td>Heterogeneous compound symmetry</td>
<td>Heterogeneous</td>
<td>Homogeneous</td>
<td>M3($\sigma_G^2, \sigma_e^2$)</td>
<td>4</td>
</tr>
<tr>
<td>Compound symmetry</td>
<td>Homogeneous</td>
<td>Homogeneous</td>
<td>M4($\sigma_e^2$)</td>
<td>2</td>
</tr>
</tbody>
</table>

† Symbols in parenthesis indicate heterogeneous parameters: $\sigma_G^2$ = genotype variance, $\rho_G$ = genotype correlation, and $\sigma_e^2$ = error variance.
Table 3. Number of datasets with each model as the best fitting model by AIC and BIC. The comparison was made in three ways: across all eight models, four models under heterogeneous error variance assumption and the other four models under homogeneous error variance assumption within each dataset.

<table>
<thead>
<tr>
<th>Model</th>
<th>AIC</th>
<th>BIC</th>
<th>AIC</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>19</td>
<td>23</td>
<td>6</td>
<td>10</td>
</tr>
<tr>
<td>M2</td>
<td>9</td>
<td>9</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>M3</td>
<td>14</td>
<td>16</td>
<td>12</td>
<td>16</td>
</tr>
<tr>
<td>M4</td>
<td>17</td>
<td>17</td>
<td>24</td>
<td>30</td>
</tr>
<tr>
<td>M5</td>
<td>4</td>
<td>22</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>M6</td>
<td>0</td>
<td>15</td>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td>M7</td>
<td>2</td>
<td>17</td>
<td>4</td>
<td>23</td>
</tr>
<tr>
<td>M8</td>
<td>0</td>
<td>11</td>
<td>7</td>
<td>25</td>
</tr>
</tbody>
</table>
Table 4. Comparisons between models M1-M7 and model M8, which assumed homogeneous error variance, genotypic variance, and genotypic covariances between pairs of environments. For the Likelihood, AIC and BIC, the difference between each of models M1 though M7 and model M8 was computed for each of 65 datasets and averaged across the 65 datasets. To quantify the magnitude of the difference in likelihood ratios, the probability of obtaining a larger value from a chi-squared distribution with degrees of freedom corresponding to the difference in number of parameters between the models in the in Table and Model M8, which had three parameters.

<table>
<thead>
<tr>
<th>Model</th>
<th>Number of parameters</th>
<th>Average difference in likelihood from M8</th>
<th>Probability of a larger ( \chi^2 )-variate</th>
<th>Average AIC difference from M8</th>
<th>Average BIC difference from M8</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1 ( (\sigma_G^2, \rho_G, \sigma_e^2) )</td>
<td>9</td>
<td>77.71</td>
<td>5.3E-15</td>
<td>65.74</td>
<td>47.69</td>
</tr>
<tr>
<td>M2 ( (\rho_G, \sigma_e^2) )</td>
<td>7</td>
<td>65.42</td>
<td>1.1E-13</td>
<td>57.42</td>
<td>45.36</td>
</tr>
<tr>
<td>M3 ( (\sigma_G^2, \sigma_e^2) )</td>
<td>7</td>
<td>69.39</td>
<td>1.5E-14</td>
<td>61.39</td>
<td>49.32</td>
</tr>
<tr>
<td>M4 ( (\sigma_e^2) )</td>
<td>5</td>
<td>56.70</td>
<td>2.4E-13</td>
<td>52.70</td>
<td>46.67</td>
</tr>
<tr>
<td>M5 ( (\sigma_G^2, \rho_G) )</td>
<td>7</td>
<td>28.45</td>
<td>5.1E-6</td>
<td>20.48</td>
<td>8.46</td>
</tr>
<tr>
<td>M6 ( (\rho_G) )</td>
<td>5</td>
<td>10.57</td>
<td>0.0025</td>
<td>6.57</td>
<td>0.53</td>
</tr>
<tr>
<td>M7 ( (\sigma_G^2) )</td>
<td>5</td>
<td>18.84</td>
<td>4.1E-5</td>
<td>14.84</td>
<td>8.80</td>
</tr>
</tbody>
</table>
Figure Caption

Figure 1. Number of hybrids in a data set versus the best model. Each dot or triangle represents datasets of a particular size, with some dots or triangles representing more than one dataset. The total number of datasets with each model identified as the best model is given in parentheses.
Figure 1
CHAPTER 3. PREDICTIVE ABILITY ASSESSMENT OF LINEAR MIXED MODELS IN MULTI-ENVIRONMENT TRIALS IN CORN (Zea mays L.)

To be submitted to Crop Science

Yoon-Sup So and Jode Edwards

Abstract

Prediction of future performance of cultivars is an important objective of multi-environment trials. To achieve this goal, predictive ability of statistical models is often examined using cross validation that partition a data set into modeling and validation data. In the present paper, a series of linear mixed models with varying degrees of heterogeneous genotype variance, correlation and error variance structure was compared for their ability to predict performance in an untested environment in 51 unbalanced data sets from the Iowa Crop Performance Tests–Corn. Simulation studies were conducted to investigate the impact of small number of common hybrids on variance component estimates. Our cross validation study indicated that in most cases there was no substantial difference in predictive ability among models that covered from none to the most general types of heterogeneity of genotype variance, correlation, and error variance structure. When there was a relatively large difference among models, it appeared that the difference was due to poor estimation of variance components in models with large heterogeneity because of very small sample size caused by unbalancedness from combining two years of data sets. Simulation study confirmed the observation from cross validation. Because the prediction ability based on an average BLUPs across environments are about the same for models with varying degrees of heterogeneity in genotype variance, correlation and error variance, we may still need to find a statistical model with the best fit of the observed data which would give the most appropriate shrinkage estimator for each environment.
Introduction

Prediction of hybrid performance is the primary goal in multi-environment trials in plant breeding. Predictive precision of yield trials can be improved cost-effectively by means of effective statistical analysis (Gauch and Zobel, 1988). The usual MET data analysis which has been termed “postdictive” evaluation seeks a statistical model that captures as much of the variation as possible in the observed data (Gauch and Zobel, 1988). They argued that the best postdictively successful model is not usually the best prediction model. Instead, they proposed a model selection criterion for what they termed “predictive” success based on the assessment of a model’s ability for prediction of data not included in model construction with the statement: “Yield trials actually measure past yields on experimental plots, but the purpose is to improve future yields on farmer’s fields”. Data not included in model construction is considered as simulation of future performance not yet measured and this type of data dropping and evaluation is done through cross validation process.

Many researchers took this idea and developed data splitting schemes of their own. Gauch (1988) and Crossa et al. (1991) used a method that selected a set of \( m \) random replications of genotypes from total of \( n \) replications within each environment ignoring the blocking scheme of the experiment, leaving the rest set of \( m-n \) replications of genotypes for validation. Piepho (1994) argued that when the experiment is blocked, the block structure should be preserved in data partition procedure. Gauch and Zober (1996) and Dias and Krzanowski (2003) suggested leave-one-out methods that take only one observation for validation, arguing that model construction with \( n-1 \) observation looses the minimum information from the full set of data with \( n \) observations and makes the most efficient use of the full data set. None of the aforementioned studies, however, appear to predict yields of hybrids evaluated in untested or unobserved environments, such as farmer’s fields. Rather they predicted yield response on experimental plots in a way that the results of analyses were to be extrapolated to the entire target environments after validation of prediction was done on the sample environments. It is the authors’ belief that data splitting strategy and validation of prediction should reflect the real world procedure as much as possible in that extrapolation of estimators from the analysis on
sample environments, hence “prediction” should be done prior to validation on another sample of unobserved environments. This naturally leads us to a data splitting scheme in which a set of validation data should include at least one environment worth of data set. In this regards, we found in the literature only one study by Bernardo (1992) in which arithmetic mean of corn hybrids were compared to two types of weighted mean yields via cross validation that involved a series of 3 to 10 environment data sets assigned as validation data.

Besides the choice of model selection criteria, MET data analyses present plant breeders with the recurrent problem of genotype by environment (GE) interaction that often involves a change in rank which is referred to as cross-over interaction (Bernardo, 2002). Presence of GE interaction is due to heterogeneity of genotype variance among environments and the lack of perfect correlation of genotype among pairs of environments (Bernardo, 2002; Farconer, 1952). Cooper and DeLacy (1994) stated that “any study which investigates the impact of GE interaction on response to selection should distinguish between these two components”. This statement implies the need to investigating the presence of heterogeneity of genotype variance and correlation among environments in MET data analyses.

Another aspect of MET data is that plant breeders often deal with highly unbalanced data sets by hybrid turn-over or selection (Piepho and Möhring, 2006). This means that not all variance components are estimated with equal precision in models in which heterogeneous genotypic variance, correlation and error variances are assumed. Although restricted maximum likelihood (REML) estimation for variance components is the preferred method for unbalanced MET data analysis (Holland, 2006; Piepho and Möhring, 2006), there are very few empirical studies that have examined the impact of unbalancedness on predictive ability of mixed linear models with varying degrees of heterogeneity in treatment variance components.

The objectives of this paper were 1) to assess predictive ability of multivariate mixed models with varying degrees of heterogeneity of genotype variance, correlation, and error variance via cross validation, 2) to investigate the relationship between variance component estimates in data sets with small sample size due to unbalancedness and
prediction ability of the models using simulation and 3) to determine if a general recommendation can be made for choice of a best prediction model for long term use in the Iowa Crop Performance Tests—Corn.

Materials and Methods

Corn yield data were taken from Iowa Crop Performance Tests—Corn (Iowa Crop Improvement Association, 2005) from 1995 to 2005. The test program divided the state of Iowa into 7 districts. There were 3 test locations in each district in which the same entries were evaluated in a given year. Due to rapid hybrid turnover, no consecutive two years had the same entries tested in a district although there were some common hybrids between two continuous years. This cross-validation study combined two years of yield trial data in each district where there were data from all 3 test locations available. Combining three years of data was not feasible because the first and third year rarely had any hybrids in common. In total, 51 data sets were available for this cross-validation study. Each data set was composed of 6 environments (2 years × 3 locations). The total number of hybrids in the 51 data sets ranged from 187 to 386. In a single year, the maximum number of hybrids tested was 240 and the minimum was 90. The number of common hybrids between two years in a data set ranged from 25 to 110. The trials were evaluated in an α-lattice design from 1995 to 2003 and in a row-column design from 2004 to 2005 with 4 replications per location.

Statistical model and genotype variance covariance structures: For the combined analyses, trials evaluated in a row-column design were treated as an α-lattice design, ignoring blocking in columns since row blocks in 2004-2005 data sets were in the corresponding block direction in α-lattice design of the previous trials. The statistical model for trials designed as an α-lattice design can be presented in the form of multivariate mixed linear model:

\[ \begin{align*}
Y_{erg} &= \mu + E_e + R_{re} + B(R)_{brc} + \gamma_{ge} + \epsilon_{erbg},
\end{align*} \]

where \( Y_{erg} \) is the grain yield of \( g^{th} \) genotype in \( b^{th} \) lattice block of \( r^{th} \) replication at \( e^{th} \) environment, \( \mu \) is the grand mean, \( E_e \) is the effect of \( e^{th} \) environment, \( R_{re} \) is the effect
of \( r \)th replication at \( e \)th environment, \( B(R)_{nre} \) is the effect of \( b \)th lattice block within \( r \)th replication at \( e \)th environment, \( \gamma_{ge} \) is the effect of \( g \)th genotype at \( e \)th environment and \( \epsilon_{ergb} \) is the residual associated with \( e \)th environment, \( r \)th replication, \( b \)th lattice block and \( g \)th genotype. The environment effect \( E_e \) can be further decomposed to \( L_l + Y_y + LY_{yl} \), \( l \)th location effect, \( y \)th year effect and their interaction depending on model assumptions. The genotype effect (\( \gamma_{ge} \)) and residual (\( \epsilon_{ergb} \)) were considered random effects and all remaining effects were considered fixed. For convenience, random effects were expressed as Greek characters while fixed effects were in Roman characters. We compared models in which error variances were allowed to be heterogeneous among environments (\( \sigma^2_{e_1} \neq \sigma^2_{e_2} \neq \ldots \neq \sigma^2_{e_6} \)) and in which error variances were homogeneous (\( \sigma^2_{e_1} = \sigma^2_{e_2} = \ldots = \sigma^2_{e_6} \)).

Genotype effects were assumed to follow a multivariate normal distribution with zero mean and variance covariance matrix of the form \( I_g \otimes \Sigma_{G_{ge}} \), where \( I_g \) is an \( g \)-dimensional identity matrix, \( \otimes \) indicates the Kronecker product of two matrices and \( \Sigma_{G_{ge}} \) is an \( e \)-dimensional symmetric variance covariance matrix of a genotype. The block diagonal form was produced by the distributional assumption of independence among genotypes. We further assumed that all genotypes had the same variance covariance matrix, \( \Sigma_{G_{ge}} \). A general form of \( \Sigma_{G_{ge}} \) can be written as follows:

\[
\Sigma_{G_{ge}} = \begin{bmatrix}
\sigma^2_{G_1} & \sigma^2_{G_2} & \cdots & \sigma^2_{G_1} \\
\sigma^2_{G_2} & \sigma^2_{G_3} & \cdots & \sigma^2_{G_2} \\
\vdots & \vdots & \ddots & \vdots \\
\sigma^2_{G_e} & \sigma^2_{G_{e-1}} & \cdots & \sigma^2_{G_e}
\end{bmatrix}
\]

where \( \sigma^2_{G_i} = \sigma^2_{G_j} \) (\( i \neq j \)). \( (1) \)

The order of environments in the \( \Sigma_{G_{ge}} \) matrix was arranged in the order of location within year so that \( e = 1\sim6 \) are \( l = 1, 2, 3 \) in \( y = 1 \) and \( l = 1, 2, 3 \) in \( y = 2 \), respectively. Hence, the diagonal elements, \( \sigma^2_{G_1}, \sigma^2_{G_2}, \sigma^2_{G_3} \) and \( \sigma^2_{G_4}, \sigma^2_{G_5}, \sigma^2_{G_6} \) represent environment-specific genotype variance for location 1, 2 and 3 in the 1st year and location 1, 2 and 3 in the 2nd year, respectively. Based on the general form of genotype variance covariance matrix, we
compared 12 different structures of $\Sigma_{gee}$ under homogeneous and heterogeneous error variance assumption. Hence, there were total of 24 mixed linear models examined for each of 51 data sets.

Our full model took the most general form as shown in (1) with an unstructured genotype variance covariance structure. Genotype variance in all environments and genotype covariances between all pairs of environments (consequently genotype correlations) were allowed to be heterogeneous in this structure. The number of parameters was $\frac{e(e+1)}{2}$ for this structure. The Newton-Raphson procedure of obtaining REML estimators implemented in proc mixed often produced a non-positive definite covariance matrix. In order to restrict the covariance structure to be non-negative definite, the unstructured model was estimated using the factor analytic decomposition available in proc mixed (FA0($q$) where $q=e$) which is equivalent to an unstructured covariance matrix but is restricted to be non-negative definite (SAS Institute Inc., 2004).

Factor analytic decomposition provides a straightforward way to restrict the general unstructured model. The additive main effects Multiplicative Interaction (AMMI) model is a similar concept in a fixed-effects context (Cross et al., 1990). The factor analytic decomposition of an unstructured covariance matrix is expressed as $\Sigma_{gee} = \Lambda \Lambda' + D$, where $\Lambda$ is an $e \times q$ matrix of environment loading with $k$th column for the $q^{th}$ latent factor $(k = 1 \sim q)$ and a $e \times e$ diagonal matrix for the residuals, $D = \text{diag}(\eta_1, \cdots, \eta_e)$. When $q > 1$, the element(s) in the $i$th row $(i = 1 \sim e)$ and $j$th column $(j = 1 \sim q)$ were set to zero for model identifiability. Three different structures can be assumed with respect to the D-matrix: i.) no residual, i.e., no $D$ matrix (FA0($q$) structure), (2) homogeneous residual, $D = \text{diag}(\eta_1, \cdots, \eta_e)$ (FA1($q$) structure) or (3) heterogeneous residuals, $D = \text{diag}(\eta_1, \cdots, \eta_e)$ (FA2($q$) structure).

The number of parameters for three types of the FA model are $\frac{q}{2}(2e-q+1)$, $\frac{q}{2}(2e-q+1) + 1$, $\frac{q}{2}(2e-q+1) + e$ for FA0($q$), FA1($q$) and FA2($q$), respectively. All
possible numbers of $q$ factors were considered in each type based on $e = 5$ since one entire environment data set was assigned as validation data for cross-validation and was not used in actual analyses for variance parameter estimates. Nine FA models were considered in this study: FA$_0$(1), FA$_0$(2), FA$_0$(3), FA$_0$(4), FA$_1$(1), FA$_1$(2), FA$_1$(3), FA$_2$(1), and FA$_2$(2). As described previously, our unstructured model was fit using the FA$_0$(5) because this model is guaranteed to be at least semi-positive definite meaning all eigenvalues of $\Lambda$ are equal to or greater than zero.

Further constraints on parameters were made to produce a 3-way interaction model. An example of this structure in a multivariate form is presented below for a 2 year $\times$ 3 location data set. Development of the model and the relationship with an univariate counterpart is described in Appendix.

$$
\sum_{\text{interaction}} (3-WAY) = \begin{bmatrix}
\sigma^2_G & \sigma_G & \cdots & \sigma_G \\
\sigma_G & \sigma^2_G & \cdots & \sigma_G \\
\vdots & \vdots & \ddots & \vdots \\
\sigma_G & \sigma_G & \cdots & \sigma^2_G
\end{bmatrix}
$$

The last structure of $\sum_{G_{\text{env}}}$ is the compound symmetry structure where genotype variance across all environments and genotype covariance between pairs of environments are assumed homogenous:

$$
\sum_{G_{\text{env}}} (CS) = \begin{bmatrix}
\sigma^2_G & \sigma_G & \cdots & \sigma_G \\
\sigma_G & \sigma^2_G & \cdots & \sigma_G \\
\vdots & \vdots & \ddots & \vdots \\
\sigma_G & \sigma_G & \cdots & \sigma^2_G
\end{bmatrix}
$$

Note that location and year effects were not distinguished in this structure but were combined as the environment effect, which is a common practice in MET data analyses. This is a multivariate form of a classic univariate linear model under homogeneous error.
variance for MET data analyses when a combination of location and year effect is treated as an environment from the univariate 3-way interaction model in Appendix.

**Cross-validation procedure:** We cross-validated by deleting one environment and used data from remaining environments to develop the predictor. With total of 6 environments in a full data set, six rounds of cross-validation were possible. Predictive accuracy of the predictor from the models was examined by root mean squared error of prediction difference ($RMSPD_{ALL}$) as:

$$RMSPD_{ALL} = \sqrt{\sum_{e=1}^{6} w_e MSE(model)_e},$$

where $w_e$ was a weight due to unequal number of genotypes across all cross-validation rounds by unbalanced structure and $MSE(model)_e$ was the mean squared error of the model for $e^{th}$ round of cross-validation in which data from $e^{th}$ environment was set aside as validation data. The $MSE(model)_e$ for $e^{th}$ round of cross-validation was driven as

$$MSE(model)_e = RMSPD_e^2 - Var(validation)_e = \frac{\sum_{g=1}^{n+e} (\hat{y}_{e,g} - \hat{\hat{y}}_{g})^2}{n_v + n_e} - \frac{\sigma^2_e}{r}. \quad (2)$$

The squared $RMSPD_e$ estimates the error in predicting the true value at the validation environment as $RMSPD_e^2 = MSE(model - validation)_e = MSE(model)_e + Var(validation)_e$ where $MSE(model)_e = Var(model) + (Bias)^2$ (For the development of this equation, see Crossa et al. (1990)). Hence, our measure of prediction error, $RMSPD_{ALL}$ was a weighted mean of $RMSPD_e$ adjusted for error in validation data across all six rounds of cross-validation.

The term $n_v$ in equation (2) indicates the number of hybrids grown only in the year containing $e^{th}$ environment of the validation data while the term $n_c$ is the number of common hybrids across two years. When $n_m$ is designated as the number of hybrids grown only in the year not containing the $e^{th}$ environment of the validation data (only present in modeling data), the weight term $w_e$ in $RMSPD_{ALL}$ equals $\frac{n_v + n_c}{n_v + n_c + n_m}$ for $e^{th}$
cross-validation. The weight was due to the unbalancedness of pooled data for two years and \( n_x \neq n_m \) in most cases.

The \( m \hat{\gamma}_g \) in \( RMSPD_e^2 \) in equation (2) represents the predictor and is an average of the environment-specific empirical best linear unbiased predictions (EBLUPs) for the \( g^{th} \) genotype across all environments in the modeling data from the \( e^{th} \) cross-validation. The \( v \hat{\gamma}_g \) is the predictand and is the arithmetic mean yield of the \( g^{th} \) genotype deviated from the grand mean in the validation data from \( e^{th} \) cross-validation. This was based on a linear model for validation data expressed as \( vY_{gr} = v\mu + v\gamma_g + v\epsilon_{gr} \), where \( Y_{gr} \) is the yield of the \( g^{th} \) genotype at \( r^{th} \) replication, \( \mu \) is the grand mean, \( \gamma_g \) is the \( g^{th} \) genotype effect and \( \epsilon_{gr} \) is the residual of \( g^{th} \) genotype at \( r^{th} \) replication. All effects were fixed except the random residual with \( \text{N}(0, \sigma^2_v) \). Hence, \( v \hat{\gamma}_g \) is equal to \( v \hat{Y}_g - v\mu = v \hat{\gamma}_g + v \hat{\epsilon}_g \), where \( v \hat{Y}_g \) and \( v \hat{\epsilon}_g \) is the simple mean yield of the \( g^{th} \) genotype and the average residual of the \( g^{th} \) genotype over all replications, respectively. While the cross-validation compared the predictor of a model to the predictand \( (m \hat{\gamma}_g, v \hat{\gamma}_g) \), our real interest was a comparison between the predictor \( (m \hat{\gamma}_g) \) and \( v \hat{\gamma}_g \) \( (=v \hat{\gamma}_g - v \hat{\epsilon}_g) \). Equation (2) was thus driven with an assumption that the predictor and residual in the validation data are independent. Model comparison was based on the \( RMSPD_{ALL} \) and smallest \( RMSPD_{ALL} \) of a model indicated that the model predicted performance of hybrids the best over all the others in average across all unobserved environments.

In addition to cross-validation, the same 24 models were fitted to full data sets \( (e = 6) \) and models were also selected based on the most two common information criteria, Akaike’s Information Criterion (AIC) and Bayesian Information Criterion (BIC), for comparison to the result from cross-validation.

**Data analyses and summary:** All models were fit using the mixed procedure in SAS (SAS Institute Inc., 2004). The 12 genotype variance covariance structures were constructed with TYPE option in RANDOM statement. Heterogeneous error variance structure was modeled using GROUP option in REPEATED statement.
Parameters were estimated via REML and initial values were only given in PARMS statement when the first estimation attempt failed because of likelihood problem. We discarded models for data sets when likelihood problem was not resolved with a few different starting values.

We also examined a relative importance of lack of perfect genetic correlation among environments to heterogeneity of genotype variance in GE interaction variance component in the 51 data sets from equations summarized by Cooper and DeLacy (1994). The unstructured variance covariance model under heterogeneous error variance assumption was used to partition heterogeneity of genotype variance and the lack of perfect correlation among environments:

\[
\sigma_{GE}^2 = V(\sigma_{G_1}) + L(r_{G_1}) = \sum_j (\sigma_{G_i} - \bar{\sigma}_{G_i})^2 + \bar{\sigma}_{G_i} \sigma_{G_j} (1 - \frac{\bar{\sigma}_{G_i}}{\sigma_{G_i} \sigma_{G_j}}),
\]

where \(\bar{\sigma}_{G_i}\) is the average of the square roots of the environment-specific genotypic variance, \(\bar{\sigma}_{G_i} \sigma_{G_j}\) is an average of all the pairwise geometric means among the environment-specific genotype variance and \(\bar{\sigma}_{G_j}\) is an average of all genetic covariances.

The term \(\frac{\bar{\sigma}_{G_i}}{\sigma_{G_i} \sigma_{G_j}}\) is also referred to as the pooled genetic correlation among all the environments.

Single year and district data sets were screened for possible outliers using unstructured genotype variance covariance model with heterogeneous error variance assumption for a previous study. There were 69 outliers removed out of a total of 132,443 observations (0.052%).

**Simulation study:** To study the impact of forming unbalanced data sets for combined analysis, we set up a simulation study with two true \(\sum_{G_{ov}}\) structures, compound symmetry and unstructured with true heterogeneous error variance. Variance component estimates for the two scenarios were selected from a data set for which each \(\sum_{G_{ov}}\) structure was determined as the best fitting model by BIC in analyses of full data.
sets. Parameters used for compound symmetry structure were $0.3354$ for $\sigma_G^2 + \sigma_G$ and $0.1393$ for $\sigma_G$ with heterogeneous error variance parameters of $R' = [0.4083, 0.384, 0.5269, 0.8373, 0.6249, 1.0333]$. Parameters used for unstructured genotype variance covariance are in Table 5 with the same error variance parameters as in the compound symmetry model. We then used the estimates to generate a data set with a balanced set of 175 entries per location, 3 locations, two years, and three replications per location. Entries were arranged in a randomized complete block design for simplicity. We took this data set and randomly selected a set of 75 unique entries from the first year, another set of 75 unique entries for the second year and 25 common entries across two years. Hence, in the unbalanced data case, we had total of 100 entries for the first and second year each, with 25 shared entries, while maintaining a total of 175 entries across the two years, as in the balanced data case. There were four scenarios from a combination of two different distributional assumptions (compound symmetry versus unstructured) and two different data structures (balanced versus unbalanced case). Each scenario was simulated 200 times. Fixed effects were set to zero in simulations, but were included in the linear models. There were four models fitted to the data sets, compound symmetry, $FA_1(2)$, $FA_2(2)$ and unstructured genotype variance covariance model under heterogeneous error variance. The pooled GE interaction variance ($\sigma_{GE}^2 = V(G) + L(r_{G})$) was computed from the estimated genotype variance covariance matrix of $FA_1(2)$, $FA_2(2)$ and unstructured models and it was compared to the GE interaction variance obtained from the compound symmetry model as $\sigma_{GE}^2 = \sigma_G^2 - \sigma_G$.

Results

From the combined analyses of all 6 environments by the compound symmetry model under homogeneous error variance, estimates of homogeneous genotype variance ($\sigma_G = \sigma_G^2$) ranged from 0.08 to 0.75 with an average of 0.24, whereas estimates of homogeneous GE interaction variance ($\sigma_G^2 - \sigma_G = \sigma_{GE}^2$) ranged from 0.08 to 0.31, with an average of 0.16. Estimates of homogeneous error variance ($\sigma_e^2$) in each data set
ranged from 0.39 to 1.14 with an average of 0.58. Genetic correlations among
environments in all data sets had a range of 0.26 to 0.83, with an average of 0.58.

The 3-way interaction model under homogeneous error variance fitted to full data
sets provided estimates of three different types of genetic correlations; (1) correlation
between different locations within each year, which ranged from 0.33 to 0.85 with an
average of 0.62, (2) correlation between two years within the same location, which
ranged from -0.43 to 0.85 with an average of 0.38, and (3) correlation between different
locations between two years, which had a range of -0.19 to 0.77 with an average of 0.37
in all 51 data sets. In general, genetic correlation between locations within year was
greater than correlation between years within location except in 4 data sets while it was
mostly greater than correlation between different locations in two different years except
only in one data set.

From the analyses to full data sets, we examined the fit of the aforementioned two
classic MET data models by information criteria. The 3-way interaction model was better
in 42 data sets than the compound symmetry model in 51 data sets by AIC. The
compound symmetry model by BIC, however, had better fit in 27 data sets than the 3-
way interaction model. The result was quite the same for the two genotype variance
structure under heterogeneous error variance assumption.

The GE interaction variance \( \sigma^2_{GE} \) is due to a combination of heterogeneity of
genotype variance and the lack of genotype correlation among environments (Bernardo,
2000; Cooper and DeLacy, 1994). From the parameter estimates of the unstructured
genotype variance covariance structure with heterogeneous error variance assumption,
lack of perfect correlation \( L(r_{Gij}) \) of hybrid performance among pairs of environments
explained from 61.1% to 98.7% of the GE interaction variance with an average of 87.9%
across the 51 data sets. The pooled genotype correlation among all pairs of environments
corrected for heterogeneity of genotypic variance among environments ranged from 0.06
to 0.75 with an average of 0.482.

Table 6 presents a summary of the cross-validation analyses, as well as the full
environment data analyses. Model comparison was made among all 24 models and
among the 12 genotypic covariance-structure models within each error variance structure (homogeneous or heterogeneous). Among all models considered for each data set, compound symmetry structure under heterogeneous error variance had the lowest RMSPD_{ALL} via cross-validation for 9 data sets. Based on the best prediction model from each data set, prediction error ranged from 0.35 ton/ha to 0.65 ton/ha with an average of 0.48 ton/ha in all 51 data sets. The second best overall model was the 3-way interaction model with heterogeneous error variance, which was chosen in 6 data sets, followed by the compound symmetry and 3-way interaction model with homogeneous error variance and FA_{1}(1) with heterogeneous error variance model from 5 data sets. In this overall comparison, the two most parsimonious genotypic covariance structures had the lowest RMSPD_{ALL} in nearly 50% of all 51 data sets evaluated. Heterogeneous error variance models appeared to have better accuracy in predicting hybrid performance for unobserved environments (62.7%, 32 data sets). In general, however, the cross-validation analyses showed that better predictors were obtained from simpler genotype variance covariance structures. This was similarly observed in separate comparisons among the 12 genetic models under the two different error variance structures. The two most parsimonious genetic models combined were picked as the best prediction model for around 50% of the 51 data sets under each error variance structure.

If we were to select a prediction model based on one of the common information criteria, AIC tended to select more complex models while BIC appeared to select toward more parsimonious structures from across all 24 models and the 12 models within different error variance assumption (Table 6). Although there seemed to be no statistical relationship between the two likelihood-based selection criteria and results from cross-validation method in the overall comparison, BIC tended to pick models close to those that were best with the cross-validation in terms of model complexity.

Magnitudes of the difference between the maximum and minimum RMSPD_{ALL} among 24 models were small across the 51 data sets (Figure 2). The difference ranged from 0.004 to 0.127 with an average of 0.033 (ton/ha). The standard deviation of environment-specific error variances from the model with the lowest RMSPD_{ALL} was 0.80 (ton/ha) in the data set where the largest difference in RMSPD_{ALL} (0.127 ton/ha) among
24 models was observed. Hence, although some patterns emerged with respect to the best model, quantitatively, the predictive ability of the 24 model did not vary a great deal.

Although the difference in $RMSPD_{ALL}$ among models, in general, was quite small, there were some data sets in which the difference in $RMSPD_{ALL}$ was relatively larger than others. To understand the cause of the difference in $RMSPD_{ALL}$ among the models, heterogeneity in genotype variance, correlation and error variance was compared to the difference in $RMSPD_{ALL}$ among models (Figure 3). To do so, we took variance component estimates from the most general form of the unstructured model with heterogeneous error variance and compared the difference between the maximum and minimum $RMSPD_{ALL}$ among models to the difference between the maximum and minimum environment-specific genotype variances, genotype correlations among pairs of environments and environment-specific error variances (Figure 3). The difference between maximum and minimum $RMSPD_{ALL}$ among all models in a data set appeared to be correlated with both differences between maximum and minimum genotypic variance ($r=0.47$) and genotypic correlation ($r=0.49$), but not with the difference between maximum and minimum error variance ($r=0.16$). Note that some pairs among the 51 data sets were not mutually independent to each other since a single year data within each district was often shared in two separate combined data sets for two years. For this reason, we didn’t carry out the test of significance for the correlation coefficient. As the difference between maximum and minimum genotypic variance among environments and correlation among pairs of environments increased, we observed an increase of difference in the maximum and minimum $RMSPD_{ALL}$ among all 24 models across the 51 data sets.

The twelve genotype variance covariance structures were set to model different degrees of heterogeneity in genotype variance among environments and correlation among pairs of environments in the data sets. The equation (3) by Cooper and DeLacy (1994) was useful in summarizing the heterogeneity in the genotype variance covariance structure of a model. We computed the pooled GE interaction variance from all 24 models in a data set and the differences between the GE interaction variance from the compound symmetry structure with the homogeneous error variance structure and the pooled GE interaction variance of the rest of the models were obtained. The standard
deviation of the differences in a data set was plotted against the difference in the maximum and minimum $\text{RMSPD}_{\text{ALL}}$ among the 24 models for all 51 data sets (Figure 4). The graph showed that larger difference in prediction ability among the models appeared to be highly correlated with larger standard deviation of the difference in pooled GE interaction variance ($r=0.76$).

Our data sets that combined two years of yield trials within each district had a great deal of unbalancedness in that there were often as few as 25 common hybrids evaluated in two consecutive years. Figure 5 showed the correlation between the number of common hybrids in a data set and the difference between maximum and minimum $\text{RMSPD}_\text{ALL}$ among models in a data set. It appeared that the prediction ability difference among models was negatively correlated with number of common hybrids in a data set ($r=0.33$). Larger differences between maximum and minimum $\text{RMSPD}_\text{ALL}$ among the models were observed with fewer common hybrids. This may be explained by poor estimation of variance components for heterogeneous variance models, especially the genotypic correlation between pairs of environments that had small number of common hybrids. In fact, the worst prediction model with the largest $\text{RMSPD}_\text{ALL}$ in a data set for all 51 data sets had at least one type of heterogeneity assumption. The FA0(1) and FA0(2) under both homogeneous and heterogeneous error variance structures had the largest $\text{RMSPD}_\text{ALL}$ in 45 data sets. Association of the standard deviation of the difference in pooled GE interaction variance to number of common hybrids was very similar to Figure 4 (data not shown).

So far, we observed from the cross-validation and full data analyses that prediction ability difference among the 24 models was highly positively associated with the magnitude of standard deviation of the difference in pooled GE interaction variance. It appeared that understanding the cause of deviation of pooled GE interaction variance of heterogeneous variance models from that of compound symmetry structure would give us an indirect assessment of the difference in prediction ability among the models. Since the standard deviation of the difference in pooled GE interaction variance was found to be negatively correlated with the number of common hybrids in a combined unbalanced data set, we set up a simulation study to compare the difference in GE interaction
variance of compound symmetry structures and a few heterogeneous variance structures under balanced and unbalanced cases.

Figure 6 shows the impact of small number of common hybrids from unbalanced data sets on the difference in GE interaction variance between the homogeneous and heterogeneous variance structures from the simulation study. For all 200 simulated data from balanced and unbalanced case with compound symmetry and unstructured true parameter scenarios, estimates of the GE interaction variance were obtained for all four models from equation (3). The difference between the estimated GE interaction variance from the compound symmetry structure and that of the three heterogeneous variance models was taken and used to plot the variation of the difference. With 175 total hybrids for the balanced data case in both compound symmetry and unstructured true parameter scenarios, estimates of the pooled GE interaction variance from the three heterogeneous variance models were quite the same as that from the compound symmetry structure. It, however, was not the case for the unbalanced data case where there were only 25 common hybrids between two years.

Discussion

Predictive ability of 12 genetic variance covariance structures under homogeneous and heterogeneous error variance assumption was examined via cross-validation. We conducted our cross-validation by a data partition scheme in which entire genotype observations from one single environment were assigned as validation data. This approach has not been commonly used in other cross-validation studies of MET data analyses. One method of cross-validation has been to partition data into n replications within each environment for modeling data and m replications for validation data with n+m replication with or without preserving block structure (Crossa et al., 1990; Piepho, 1994). Others have used a leave-one-out data partition method, arguing that less data points assigned for validation data is closer to analysis of the full set of data (Cornelius and Crossa, 1999; Dias and Krzanowski, 2003). When cultivar selection to recommend for unobserved environments is the primary objective of MET data analyses, an underlying question in the statistical analyses would be how well our estimate of hybrids
predicts performance in untested locations (i.e., farmers’ fields) of the target environment. Hence, we excluded all data from one environment for our validation data. In the literature, we found only one study by Bernardo (1992) where random data partition was based on a unit of environments rather than replication within each environment. Instead of dropping one environment, Bernardo (1992) had a series of 3~10 environment samples as validation data sets from each of 1000 simulated data and obtained a simple mean across all the environments in the validation data as his predictands.

One may argue that it would be more appropriate to use data from year 1 to predict data from year 2 in cross-validation when the prediction of hybrid is for untested environment in the following year that has not been occurred. In our example data sets and cross-validation procedure, the modeling data always included observations from two test locations in the same year in which observations from the third test locations were assigned as the validation data. Moreover, the modeling data also always included observations from one test location in one year and the same location in different year was used as the validation data. Our observation from the 3-way interaction model under homogeneous error variance assumption indicated that the correlation among different locations within each year was generally higher than the correlation between two years within the same location. Including data from test locations that were observed in the same year as validation data could potentially provide a better prediction than excluding them. It, however, was not practical in our data sets which were highly unbalanced between two consecutive years. If we were to use data from year 1 to predict data from year 2, our predictor could have only provided predictor for common hybrids between the two years and there would have been no way to predict new hybrids entered in the second year. Nevertheless, we admit that the prediction error could be underestimated by the way we implemented data partition in our cross-validation.

Cross-validation resulting from 51 data sets covering 11 years of Iowa crop performance trial for corn indicated that there appeared no universally best model to be applied in MET data analyses for years of data sets especially when data sets are highly unbalanced with small sample size. Moreover, the estimated $RMSPD_{ALL}$ among the 24
models in most data sets was not largely different. This may be explained by our choice of predictor, an average of environment-specific BLUPs of a hybrid. Solution for BLUP is obtained from the mixed model equation as

\[ \hat{y} = \hat{G}Z\hat{V}^{-1}(y - X\hat{\beta}) \]

where \( \hat{y} \) is the empirical BLUP of the performance of hybrids for each environments, \( \hat{G} \) is the estimated genotype variance covariance matrix, \( Z' \) is the transpose of the incidence matrix for random effects, \( \hat{V}^{-1} \) is inverse of the estimated phenotypic variance covariance matrix, \( y \) is a vector of observations, \( X \) is incidence matrix for fixed effects and \( \hat{\beta} \) is the estimated fixed effects vector (Henderson, 1984). In this expression, \( \hat{G}Z\hat{V}^{-1} \) can be viewed as a weight matrix. The estimated phenotypic variance covariance matrix is the sum of estimated genotype variance covariance matrix and the estimated error variance matrix. Hence, the weight matrix contains information from variance component estimates for all random effects. Different types of heterogeneous genotype variance, correlation (or covariance) and error variance compared in this study imply different weighting scheme for obtaining environment-specific BLUPs of a hybrid. However, the heterogeneity of genotype variance, correlation and even error variance is balanced out during the process of averaging environment-specific BLUPs of each hybrid to obtain the predictor. Predictors for models with different weighting scheme then become quite the same among all models. Consequently, the efforts of modeling heterogeneity of random effect variance are of little benefit in selection of a best prediction model for unobserved environment since averaging over different weights from heterogeneity of variances is same as the use of equal weights from homogeneous variances.

Bernardo (1992) examined weighted and unweighted mean performance of varieties (fixed effect) across environments using similar data splitting method to ours via cross validation and recommended use of the unweighted mean in estimating average variety performance, even when error and GE interaction variances are heterogeneous. We might draw a similar conclusion based on our cross-validation study from 51 data sets and recommend use of the compound symmetry model under homogeneous error variance assumption if we are to ignore the small observed difference in \( RMSPD_{ALL} \) among the 24 models and only focus on an average across environment-specific BLUPs.
as our predictor when combined multi year data sets are highly unbalanced. This, however, could lead to biased prediction for environment-specific BLUPs per se in cases where true underlying parameter structure is in deed highly heterogeneous. For example, if a hybrid was evaluated in a higher-than-average genotypic variance environment or in a smaller-than-average error variance environment, the empirical BLUP of that hybrid for the particular environment would have too much shrinkage effect by use of the equal weight from compound symmetry model with homogeneous error variance assumption. Although the current study only focused on an average across environment-specific BLUPs, growers who refer to the test report of the Iowa Crop Performance Tests—Corn still pay great attention to the summary result on the location closest to their farms.

When there are sufficiently large samples in a balanced data set, the pooled GE interaction variance \( V \sigma^2_{GE} = V(\sigma_{G}) + L(\tau_{G}) \) from variance components estimates of heterogeneous variance models should equal to GE interaction variance estimated from the compound symmetry structure. Deviation of two estimated GE interaction variances between the compound symmetry structure and any heterogeneous variance model in our simulation study implies a potential bias in variance component estimates due to finite sample size. In balanced cases, all variance components in the three heterogeneous variance models are estimated with equal precision but this is not the case for unbalanced data, especially covariance parameters. For unbalanced data case, there were 100 entries for covariance parameters between pairs of locations within the same year but there were only 25 entries for covariance parameters between two different years. Hence, those covariance parameters estimated from 25 common entries were more poorly estimated. This resulted in wider variation of the difference in the estimate of GE interaction variance between compound symmetry structure and any heterogeneous variance covariance structure when the estimated heterogeneous variance covariance parameters were summarized as the pooled GE interaction variance and compared to the estimate of GE interaction variance of the compound symmetry structure. This observation was similar regardless of different types of true parameter distribution used in the simulation. Our simulation study supports the results from real data sets that the observed difference
in predictive ability among models was due to sampling errors with finite sample size. The difference was more obvious when there were fewer common hybrids in data sets.

One disadvantage of using cross-validation based model selection for MET data is that it takes considerable amount of computation time. For our cross-validation study, it would take about 2 weeks for total of 7344 separate analyses (24 models × 6 cross-validation rounds × 51 data sets) with an ordinary personal computer. Knowing that prediction ability of the models via cross-validation are quite the same even with small number of common hybrids between two years, we may not need to perform further cross-validation analyses to select a best prediction model. Instead, we may focus on more accurate prediction of hybrid performance for each environment. Hence, it may be necessary to examine the fit of models with varying degrees of heterogeneity in genotype variance, correlation and error variance and we have a few choices such as AIC and BIC used in this study.

The Iowa Crop Performance Tests—Corn continued to observe reduction in number of hybrids each year across districts and number of common hybrids year to year. In the current paper, data combining was restricted within each district. In more recent years, we observed that there were more common hybrids between two adjacent districts than common hybrids between two consecutive years within a district. We could then exploit information available from neighboring district by combining data across districts. The correlation between environments in the mixed linear model would automatically provide for an appropriate weighting of the data among districts. District-specific predictor could be produced by setting up different inference spaces and averaging environment-specific BLUPs from each district while using all the data in the model.

Appendix

3-WAY interaction model with SAS

A univariate linear model that involves the 3-way interactions can be expressed as

\[ Y_{ijr|bg} = \mu + Y_{j} + L_{i} + Y_{r}L_{i} + R(YL)_{ryl} + B(RYL)_{bryl} + \gamma_{g} + \gamma_{Lg} + \gamma_{YLg} + \epsilon_{ijr|bg}, \]
where $Y_y$ is the $y^{th}$ year effect, $L_l$ is the $l^{th}$ location effect, $\gamma_{g_y}$ is the 2-way interaction effect between $g^g$th genotype and $y^y$th year, $\gamma_{g'_l}$ is the 2-way interaction effect between $g^g$th genotype and $l^l$th location, $\gamma_{yL_{g_l}}$ is the 3-way interaction of $g^g$th genotype in $y^y$th year at $l^l$ location and all the others are defined as before. Random effects of the genotype main effect, the interaction effects involving the genotype main effect and residual have a normal distribution with zero mean and variance of $\sigma^2_y$, $\sigma^2_{yY}$, $\sigma^2_{gL}$, $\sigma^2_{gL}$, $\sigma^2_{yl}$ and $\sigma^2$. All the others are considered fixed effects.

Phenotype covariance of yield of a hybrid measured in different replications can occur in four different ways:

1) In same year ($y, y$) at same location ($l, l$) as

$$\text{Cov}(Y_{yrbg}, Y_{ylrbg}) = \text{Cov}(\gamma_g + \gamma_{Y_{g'}}, \gamma_{gL} + \gamma_{Y_{gL}} + \gamma_{yL_{g'}}, \gamma_g + \gamma_{Y_{g'}}, \gamma_{gL} + \gamma_{Y_{gL}} + \gamma_{yL_{g'}})$$

$$= \text{Cov}(\gamma_g + \gamma_{Y_{g'}}, \gamma_{gL} + \gamma_{Y_{gL}} + \gamma_{yL_{g'}})$$

$$= \text{Cov}(\gamma_g, \gamma_{g'}) + \text{Cov}(\gamma_{Y_{g'}}, \gamma_{Y_{gL}})$$

$$= U \sigma^2_y + U \sigma^2_y + U \sigma^2_y + U \sigma^2_y \quad \ldots (1)$$

2) In same year ($y, y$) at different locations ($l, l'$) as

$$\text{Cov}(Y_{yrbg}, Y_{ylrb'}g) = \text{Cov}(\gamma_g + \gamma_{Y_{g'}}, \gamma_{gL} + \gamma_{Y_{gL}} + \gamma_{yL_{g'}})$$

$$= \text{Cov}(\gamma_g + \gamma_{Y_{g'}}, \gamma_{gL} + \gamma_{Y_{gL}})$$

$$= \text{Cov}(\gamma_g, \gamma_{g'})$$

$$= U \sigma^2_y \quad \ldots (2)$$

3) In different years ($y, y'$) at same location ($l, l$) as

$$\text{Cov}(Y_{yrbg}, Y_{ylrb'g}) = \text{Cov}(\gamma_g + \gamma_{Y_{g'}}, \gamma_{gL} + \gamma_{Y_{gL}} + \gamma_{yL_{g'}})$$

$$= \text{Cov}(\gamma_g + \gamma_{L_{g'}}, \gamma_{gL} + \gamma_{Y_{gL}})$$

$$= \text{Cov}(\gamma_g, \gamma_{L_{g'}})$$

$$= U \sigma^2_y \quad \ldots (3)$$

4) In different years ($y, y'$) at different location ($l, l'$) as

$$\text{Cov}(Y_{yrbg}, Y_{ylrb'g}) = \text{Cov}(\gamma_g + \gamma_{Y_{g'}}, \gamma_{gL} + \gamma_{Y_{gL}} + \gamma_{yL_{g'}})$$

$$= \text{Cov}(\gamma_g, \gamma_{gL})$$

$$= U \sigma^2_y \quad \ldots (4)$$
This can be further arranged in a matrix and the matrix is equivalent to the genotype variance covariance matrix ($\Sigma_{G_{tot}}$) in the multivariate linear mixed model when there are 2 years and 3 locations per year are combined:

\[
\begin{bmatrix}
    Y_1 & Y_2 & Y_1 & Y_2 \\
    L_1 & L_2 & L_3 & L_1 & L_2 & L_3 & L_1 & L_2 & L_3 \\
    (1) & (2) & (3) & (4) & (4) & (2) & (1) & (4) & (3) & (4) & (2) & (2) & (1) & (4) & (4) & (3) & (4) & (3) & (4) & (4) & (4) & (1) & (2) & (2) & (3) & (4) & (4) & (4) & (2) & (1) & (2) & (4) & (3) & (4) & (2) & (2) & (2) & (2) & (2) & (1) \\
\end{bmatrix}
\]

\[
\begin{bmatrix}
    \sigma^2_G & \sigma_{G_1} & \sigma_{G_2} & \sigma_{G_3} & \sigma_{G_4} \\
    \sigma_{G_1} & \sigma^2_G & \sigma_{G_2} & \sigma_{G_3} & \sigma_{G_4} \\
    \sigma_{G_2} & \sigma_{G_1} & \sigma^2_G & \sigma_{G_3} & \sigma_{G_4} \\
    \sigma_{G_3} & \sigma_{G_1} & \sigma_{G_2} & \sigma^2_G & \sigma_{G_4} \\
    \sigma_{G_4} & \sigma_{G_1} & \sigma_{G_2} & \sigma_{G_3} & \sigma^2_G \\
\end{bmatrix}
= \Sigma_{G_{tot}} (3-WAY)
\]

Hence, $\sigma^2_G$, $\sigma_{G_1}$, $\sigma_{G_2}$ and $\sigma_{G_3}$ in the multivariate 3-way interaction model are equal to $\sigma^2_Y + U \sigma^2_{Y_1} + U \sigma^2_{Y_2} + U \sigma^2_{Y_3}$, $\sigma^2_Y + U \sigma^2_{Y_1} + U \sigma^2_{Y_2} + U \sigma^2_{Y_3}$, $\sigma^2_Y + U \sigma^2_{Y_1} + U \sigma^2_{Y_2} + U \sigma^2_{Y_3}$ and $\sigma^2_Y$, respectively.

This structure is not pre-defined in MIXED procedure in SAS and must be constructed with TYPE=LIN($t$) for $t$ parameters in RANDOM statement followed by LDATA= option. The option LDATA= assigns a data table which contains information about the parameters and should be arranged in the following manner.
<table>
<thead>
<tr>
<th>ROW</th>
<th>COL</th>
<th>PARM</th>
<th>VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>5</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>6</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

The first row of a data table contains variable names, \( ROW \) and \( COL \) for location of a particular parameter in the matrix, \( PARM \) for the corresponding parameter defined as in the above matrix and \( VALUE \) for coefficients for a linear relationship among parameters. Variable names must be exactly as presented here so that SAS can recognize them for TYPE=LIN(6) option (here \( t=6 \)). When the elements in the upper and lower triangle of the matrix are equal as in the case of a variance covariance matrix, one of them can be omitted in the data table. If BY statement is to be used in PROC MIXED the data table for LDATA also must include the variable for BY statement and the above parameter information should be repeated for all level of BY variable.
References


Table 5. Parameters used for simulating unstructured variance covariance structure. Diagonal elements are environment-specific genotype variances. Genotype covariances and correlations between pairs of environments are above and below diagonal, respectively.

<table>
<thead>
<tr>
<th></th>
<th>Year 1</th>
<th></th>
<th>Year 2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Loc 1</td>
<td>Loc 2</td>
<td>Loc 3</td>
<td>Loc 1</td>
</tr>
<tr>
<td>Year 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Loc 1</td>
<td>0.4604</td>
<td>0.2494</td>
<td>0.2264</td>
<td>0.1135</td>
</tr>
<tr>
<td>Loc 2</td>
<td>0.61</td>
<td>0.3690</td>
<td>0.1382</td>
<td>0.0362</td>
</tr>
<tr>
<td>Loc 3</td>
<td>0.58</td>
<td>0.40</td>
<td>0.3301</td>
<td>0.1093</td>
</tr>
<tr>
<td>Loc 1</td>
<td>0.34</td>
<td>0.12</td>
<td>0.39</td>
<td>0.2392</td>
</tr>
<tr>
<td>Year 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Loc 2</td>
<td>0.52</td>
<td>0.65</td>
<td>0.37</td>
<td>0.59</td>
</tr>
<tr>
<td>Loc 3</td>
<td>0.65</td>
<td>0.28</td>
<td>0.80</td>
<td>0.60</td>
</tr>
</tbody>
</table>
Table 6. Number of data sets with each model as the best fitting model by cross-validation, AIC and BIC. The comparison was made in three ways: among all 24 models, 12 models under heterogeneous error variance assumption and the other 12 models under homogeneous error variance assumption within each data set.

<table>
<thead>
<tr>
<th>Model</th>
<th>No. of parameter†</th>
<th>Among all models</th>
<th>By error variance structure</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>CV</td>
<td>AIC</td>
</tr>
<tr>
<td><strong>Heterogeneous error variance models</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unstructured</td>
<td>15</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>FA₂(2)↓</td>
<td>14</td>
<td>0</td>
<td>12</td>
</tr>
<tr>
<td>FA₀(4)</td>
<td>14</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td>FA₁(3)</td>
<td>13</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>FA₀(3)</td>
<td>12</td>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td>FA₂(1)</td>
<td>10</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>FA₁(2)</td>
<td>10</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>FA₀(2)</td>
<td>9</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>FA₁(1)</td>
<td>6</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>FA₀(1)</td>
<td>5</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>3-Way interaction</td>
<td>4</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>Compound symmetry</td>
<td>2</td>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td><strong>Homogeneous error variance models</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unstructured</td>
<td>15</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>FA₂(2)</td>
<td>14</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>FA₀(4)</td>
<td>14</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>FA₁(3)</td>
<td>13</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>FA₀(3)</td>
<td>12</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>FA₂(1)</td>
<td>10</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>FA₁(2)</td>
<td>10</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>FA₀(2)</td>
<td>9</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>FA₁(1)</td>
<td>6</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>FA₀(1)</td>
<td>5</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>3-Way interaction</td>
<td>4</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>Compound symmetry</td>
<td>2</td>
<td>5</td>
<td>0</td>
</tr>
</tbody>
</table>

†Number of parameter in genotype variance covariance matrix. Models in each error variance assumption was sorted in descending order by the number of parameter.

↓FA: Factor Analytic model. Subscript identifies different D matrix structure described in Materials and Methods (0: no D, 1: homogeneous D, 2: Heterogeneous D). Number of factor (q) is indicated in parenthesis.
Figure Caption

Figure 2. Histogram of difference between maximum and minimum $RMSPD_{ALL}$ in a data set.

Figure 3. Correlation between difference in maximum and minimum $RMSPD_{ALL}$ among the 24 models in a data set and difference in a) environment-specific genotype variances, b) genotype correlations among pairs of environments and c) environment-specific error variances from the unstructured model with heterogeneous error variance assumption.

Figure 4. Correlation between the difference in the maximum and minimum $RMSPD_{ALL}$ among the 24 models in a data set from cross-validation analyses and the standard deviation of the differences in pooled GE interaction variance among the 24 models in a data set from full data set analyses. The differences in pooled GE interaction variance were computed as the difference between the GE interaction variance of compound symmetry structure with homogeneous error variance and the pooled GE interaction variance of the rest of the models.

Figure 5. Correlation between number of common hybrids and difference between maximum and minimum $RMSPD_{ALL}$ among models in a data set.

Figure 6. Boxplots for the deviation of pooled GE interaction variance of heterogeneous variance models from that of compound symmetry model based on 200 simulated data. Whiskers are at 10% and 90% percentiles.
Figure 2

The graph shows the frequency distribution of the difference between maximum and minimum \( \text{RMSPD}_{\text{ALL}} \). The x-axis represents the difference values, ranging from 0.00 to 0.14, while the y-axis indicates the frequency of occurrence.
Figure 3

(a) $y = 0.0484x + 0.0134$  
$R^2 = 0.2222$

(b) $y = 0.0432x - 0.0026$  
$R^2 = 0.2418$

(c) $y = 0.0092x + 0.028$  
$R^2 = 0.0259$
Figure 4

\[ y = 0.5495x - 0.0026 \]

\[ R^2 = 0.576 \]
Figure 5

The scatter plot shows the relationship between the difference between the maximum and minimum RMSPDALL and the number of common hybrids. The equation given is:

\[ y = -0.0004x + 0.0532 \]

with an \( R^2 \) value of 0.1078.
Figure 6

- **Compound symmetry**
- **Unstructured**

Deviation from GE interaction variance of compound symmetry model
CHAPTER 4. GENERAL CONCLUSIONS

Conclusions

Two research projects were carried out with the ultimate objective to develop analytical strategies for the Iowa Crop Performance Tests—Corn that make use of all available information for all comparisons rather than relying on sub setting the data into many individual analyses that ignore portions of available data.

The first project was undertaken in part to determine what levels of heterogeneity should be included in a more comprehensive study of analytical approaches to obtain the best predictions of hybrid performance. There was a conflicting result from two likelihood-based model selection criteria considered, in that AIC identified more complex models and BIC identified simpler models, making it hard to pick one criterion for model selection. For heterogeneity of genotypic variance and correlation, it was not possible from the analyses to conclude generally that heterogeneity of genotypic components would be important or that including heterogeneity in genotype variance or correlation would necessarily be required in all circumstances. There, however, was overwhelming evidence of the presence of heterogeneous error variances among locations in corn hybrid performance evaluation trials in Iowa based on both model selection criteria examined.

Having confirmed that there was an evidence of presence of heterogeneity in genotype variance, correlation or error variance, the second research was carried out to assess predictive ability of multivariate mixed models with varying degrees of heterogeneity of genotype variance, correlation and error variance via cross validation and to investigate the impact of small number of common hybrids on variance component estimates using simulation. Best predictive models by cross validation tended to have simpler genetic variance covariance structures, much like BIC tended to pick simpler models while AIC identified more complex models. Although cross validation and BIC had a tendency to select simpler models, they rarely agreed on the same model as the best in each data set. There seemed to be no apparent relationship between the model selection method via cross validation and either of likelihood-based information criteria.
considered. Nevertheless, there was no substantial difference in prediction ability among the 24 models examined via cross-validation. When there was a relatively larger difference between two models with the lowest and highest $RMSPD_{ALL}$, it appeared that the difference was due to poor estimation of variance components in models with large heterogeneity because of very small sample size caused by unbalancedness from combining two years of data sets. It was observed from the cross-validation and full data analyses that prediction ability difference among the 24 models was highly positively associated with the magnitude of standard deviation of the difference in pooled GE interaction variance. It appeared that understanding the cause of deviation of pooled GE interaction variance of heterogeneous variance models from that of compound symmetry structure would give us an indirect assessment of the difference in prediction ability among the models. Simulation study indicated that poor estimates of some of genotype variance covariance components in heterogeneous model by very small number of shared hybrids in unbalanced data case caused the pooled GE interaction variance to be much deviated from that of the homogeneous genotype variance covariance model. This indirectly supports our observation that there was a correlation between number of common hybrids and the difference between the best and worst prediction models by cross-validation.

The conclusion based on the two research project is that because the prediction ability based on an average BLUPs across environments are about the same for models with varying degrees of heterogeneity in genotype variance, correlation and error variance, time-consuming cross-validation analyses are not necessary to select a best prediction model in MET data analyses. Instead, we may need to pay more attention for more accurate estimator of environment-specific hybrid performance since the test report of the Iowa Crop Performance Tests—Corn would continue to provide information on the average hybrid performance across all three test locations which would be quite the same regardless of heterogeneity in the data and on the location-specific hybrid performance which would be quite different depending on the heterogeneity in genotype variance, correlation and error variance structure. An added benefit in using the likelihood-based mixed model is that two consecutive years of data within a district can
be analyzed together enabling to make comparisons among some pairs of hybrids that were not possible by sub setting of the data sets.

**Recommendations for Future Research**

1. The Iowa Crop Performance Tests—Corn continues to observe a reduction in number of hybrids entered by companies each year and in the number of common hybrids year to year. To be able to continue exploring all available data across years and across districts, a strategy of reducing bias in variance component estimates due to the small number of common hybrids across years or across districts may be needed. One way of doing this is to include more common check cultivars besides paid entries by companies. A simulation study is then suggested in determining how many common hybrids are needed and how the experimental design can be modified to incorporate the increased number of common hybrids.

2. The cross-validation study utilized combined data sets within each district and compared different models based on an average of environment-specific BLUPs across all environments included in modeling data. In more recent years, there are more common hybrids between some pairs of two adjacent districts than there are between two consecutive years within each district. Hence, multi-district analysis appears to have a potential to make use of more available data. In this regards, it would be suggested that the analysis be based on the entire district data sets as a whole and produce district-specific predictor by averaging location-specific BLUPs within each district. If there are common hybrids between two districts, the common hybrids would help estimate genetic correlation between the districts. When two districts are vastly different, very little information could come from them since the estimated correlation would be very low.

3. The district system of the Iowa Crop Performance Tests—Corn was not based on a biological research of the corn growing area. There is still a lack of sound research based clustering of homogeneous locations from entire sample population of locations from the
State of Iowa to minimize GE interaction variance. It could be suggested that state wide multi year yield trials with large number (e.g., more than hundreds) of balanced entries be performed for a more comprehensive understanding of the nature of GE interactions in the State of Iowa as one entire target environment and possibly identify clusters of homogeneous sub-environments. The highly unbalanced data used in this research project was not suitable for this type of study.

4. In suggestion #3, we can also attempt to separate GE interaction variance into repeatable and unrepeatable components by incorporating environmental covariates such as soil type, precipitation and temperature information and DNA marker information. This may aid more precise grouping of homogeneous districts. In addition, DNA marker incorporation enables us to utilize relationship matrix in mixed linear model for which present studies assumed no relationship among hybrids evaluated.