UHASSELT
KNOWLEDGE IN ACTION

Maastricht University

Faculty of Sciences
School for Information Technology

Master of Statistics

Master's thesis

Comparison of univariate versus multivariate models in large scale sugar beet field trials

Christine Jani

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics, specialization Biostatistics

SUPERVISOR :
Prof. dr. Dirk VALKENBORG

SUPERVISOR :
Mr. Juan VEGAS
Mr. Ibrahim ADETUNJI

2018
2019

Transnational University Limburg is a unique collaboration of two universities in two countries: the University of Hasselt and Maastricht University.

UHASSELT
KNOWLEDGE IN ACTION

www.uhasselt.be
Universiteit Hasselt
Campus Hasselt:
Martelarenlaan 42 | 3500 Hasselt
Campus Diepenbeek:
Agoralaan Gebouw D | 3590 Diepenbeek

Faculty of Sciences
School for Information Technology
Master of Statistics

Master's thesis

Comparison of univariate versus multivariate models in large scale sugar beet field trials

Christine Jani

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics, specialization Biostatistics

SUPERVISOR :

Prof. dr. Dirk VALKENBORG

SUPERVISOR :

Mr. Juan VEGAS

Mr. Ibrahim ADETUNJI

Acknowledgements

Many thanks to the Almighty God who gave me the gift of life, wisdom, and strength to face the challenges in this thesis.

Special thanks to SESVANDERHAVE for allowing me to do my thesis with them, it was an exciting learning opportunity in the application of statistics in crop sciences. Thank you again for providing the much-needed resources for this research. I submit my heartiest gratitude to my respected supervisors Professor Dirk Valkenborg (UHasselt) and Mr. Ibraheem Adetunji, Mr. Juan Vegas of SESVANDERHAVE for their sincere guidance and help throughout my thesis period.

I want to give my special thanks to my family, my mom, and dad, for their prayers that kept me going even in the most challenging times. My brothers and sisters Fungai, Lucia, Connie, Cordilia Farai, and Kuda for the unconditional support. It is my privilege to thank my husband Simbarashe Manyika for his constant encouragement throughout my studies and research.

I want to thank the VLIR-OUS for a study opportunity in Belgium, it was such an exciting exposure may you continue with your work and touch the lives of many. Not forgetting the University of Hasselt department of statistics for the knowledge I acquired throughout my study period.

I am deeply indebted to my colleagues from the Master of statistics class of 2019 for their valuable help in preparing this thesis, Adama, Mifflin and John who were always there to help, Nancy Teshome and Ocira who were there to boost my morale and confidence in facing up the challenges. I also express my cordial gratitude to my friends who became sisters and brothers Iris and Peter for the excellent food, Gogo Ayie for the long calls when I feel so low, Joe and Maxwell for making me feel at home away from home, Connie and Andrew to mention but a few. My special thanks also go to Faith worship center family, Pastor George and Seth for their prayers throughout my studies, indeed you are a home away from home.

2 Corinthians 3:5 Not that we are adequate in ourselves to consider anything as coming from ourselves, but our adequacy is from God

Contents

List of Figures

List of Tables

Abstract

Sugar beet breeding programs aim at creating stable and dependable varieties of sugar beets that give an optimized yield of white sugar per unit area as a function of the cost of production and meet the requirements of the environment, growers and sugar factories. This study aimed at comparing the predicted genetic breeding values of traits obtained from various univariate and multivariate methods. Data used in the analysis was obtained from a series of sugar beet field trials carried out in several countries around Europe. Exploratory showed that there is a high correlation amongst the derived traits. The principal analysis was done using the mixed models approach taking care of within-field trial variability as well as the genotype $\times$ environment interaction. Estimated breeding values were obtained from the models as Best Linear Unbiased Predictions. The analysis showed that similar genetic rankings were obtained from different univariate methods; hence, predictions of the breeding values are robust to model selections. A multivariate model was aimed at jointly modeling the traits and estimation of the breeding values and assess any potential gain in prediction. The results showed small differences in rankings and estimated values from the multivariate and the univariate procedures. The study established that one potential gain that the breeders gain from a multivariate model is the possibility of estimating the genetic and phenotypic correlations from the variance covariance matrix of random effects.

Keywords: *Field trials, Genotype$\times$Environment interaction, Multivariate,Series, Trait, Univariate.*

1 Introduction

1.1 Background

Hundreds of plant species are cultivated across the globe as food crops, among these are sugarcane and sugar beets which are the most two common sources of sugar (the common name for sucrose). Sugar beet is a specialized agricultural crop targeted at the refined sugar industry. The roots of the beets contain a high percentage of sucrose, the primary input for sugar processing; hence, sugar beet directly competes on a global scale with sugar cane (Dillen and Demont). Sugar beet is a common crop in more temperate and colder climatic regions around the world such as Europe and parts of North America where it is an essential agricultural crop (Draycott et al., 2006). Specific properties of sugar beet and its importance in industrialized countries with strong institutions and commercially oriented farmers makes it an interesting crop to be targeted by the biotechnology sector. The main characteristics of high quality beet are large concentrations of sugar (about 12%-21% of the sugar beet's total weight) and small concentrations of naturally- occurring constituents of the sugarbeet root, referred to as impurities i.e amino nitrogen (N), potassium (Na) and sodium (K) which obstruct sucrose extraction during routine factory operations (Campbell and Fugate, 2015). It is however believed genetic variance is essential in determining the variation in the relative levels of these sugar beet traits.

There is a need to maintain the competitiveness of the beet sugar industry hence a need for a continued and coordinated research efforts in all different areas related to sugar beet growing and processing. On the other hand, an increase in demand to adapt to less input-intensive and pesticide-dependent agriculture has led to the increased importance of sugar beet breeding for the past decades. The goals of sugar beet breeding programs are to create stable, dependable varieties that give the highest possible yield of white sugar per unit area in relation to cost of production and meet various other requirements of the environment, growers and sugar factories (Draycott et al., 2006). Sugar beet breeding has contributed mostly to improvements in the productivity of the crop, yield and chemical properties of the root hence an increase in the amount of white sugar extracted at the processing factories. Campbell (2002) asserts that there is a negative association between root yield and sucrose concentration, interactions among impurity components and between impurity components and sucrose concentration that complicates breeding efforts. These interactions are brought about by genetic variability in many of

the beet quality attributes, and considerable influence of environmental and agronomic factors, especially on the concentrations of amino nitrogen and sodium. In general, the quality of beet delivered to factories dramatically affects the efficiency and economics of the factory process. Several formulae and indices have been developed that weight the impurities in beet according to their influence on factory operations and the extractability of white, crystalline sugar (Harvey and Dutton, 1993).

In the process of sugar beet breeding, big data are generated, the most common type of data is the multi-trait and multi-environment data. Statistics offer methods to exploit these data in order to accelerate the breeding process as well as to understand the underlying biological mechanisms, i.e., the relationship between genotypes and phenotypic data and generate productive varieties improved for one variety (Balzarini, 2002).
Best linear unbiased predictions from the linear mixed models are commonly used for predictions and estimations of genetic merit of tested materials in plant breeding (Balzarini, 2002). Methods for analyzing plant breeding traits are grouped into two main categories based on the number of traits analyzed, univariate-trait and multivariate-trait models for one and at least two traits respectively (Montesinos-López et al., 2018). When there are many traits, breeders need Multivariate models in order to maximize the necessary information from the data. Univariate trait models involve single independent models that are trained separately for each trait; hence, they eliminate the possibility of modeling the complicated inter-dependencies between the traits. When genetic selection is based on many traits that are genetically correlated and analyzed separately, selection biases may arise (Volpato et al., 2019). This has resulted in the popularity of the Multivariate-Trait models which are concerned with simultaneously modeling of two or more traits based on a standard set of explanatory variables. The Multivariate trait models were designed to more efficiently capture the correlations between traits thus it is believed that these models give more accurate parameter estimates, better predictions and have more statistical power than the univariate trait models (Isik et al., 2017). In a comparative study by Guo et al. (2014), it was concluded that single trait models give inferior results in the presence of missing data compared to multi-trait models

Apart from the data being multi-trait these data are taken from field trials that are carried out at different agricultural areas across Europe in two years with different soil and meteorological

conditions, management choices and the incidence of abiotic and biotic stresses. New crop varieties are evaluated over many locations and years; hence, they are called Multi-Environmental trials (METs). Having experiments from different environments allows the investigation of sugar yield, yield stability, and other quality characteristics of the sugar beets to predict future genotype (seed variety) performance across different environments (Friesen et al., 2016). Several scholars have also studied the possibility of increasing these univariate and multivariate trait model predictive power by incorporating genotype by environment ($G \times E$) interaction to model relationships between environments (Ward et al., 2019). Early studies traditionally performed in plants typically used multi-environment ANOVA models to calculate adjusted means for genotypes. Critically, an ANOVA model fit in this context ignores $G \times E$ Interaction by assuming a uniform covariance between pairs of environments (Isik et al., 2017). This study adopts the mixed models approach that allow for heterogeneity of genetic variances and correlations across environments.

SESVANDERHAVE has been using univariate models to analyze trait data on sugar beet yields. However, it is believed that breeding decisions based on these univariate methods may not always be correct since they neglect the interrelationship among the sugar beet trait variables. Therefore it is hypothesized that more information may be found in the correlation pattern and not in the individual variables. Henceforth, the use of multivariate methods of analysis allows obtaining a complete and detailed analysis of the effect of seed varieties (genotypes) and environment on sugar yield. It is in this view that the study seeks to compare univariate and multi-trait models in estimating the genetic effect of different seed varieties of sugar beets using traits data.

This thesis has the following sections. In this section, we provide an introduction to the research problem. Description of the methods used for analysis is given in Section 2, while Section 3 give the results of the analysis are given. Section 4 elaborates the discussion and conclusion.

1.2 Objectives

The project aims to develop a multivariate statistical model for the calculation of white sugar yield and comparison of the proposed model with the currently used univariate approaches. This aim is achieved through the following objectives:

1. development of a multivariate statistical model for the calculation of white sugar yield,
2. validation of the proposed statistical model in a limited number of advanced series,
3. comparison of the proposed multivariate model with the currently used univariate models.

1.3 The Data

1.3.1 Experimental design and layout

The data analyzed in this research comes from designed sugar beet field trials. The primary treatment of interest factor is the genotype (represented by an object identifier) with different levels depending on the series of experiments under consideration. The experiments were implemented in several field trials across main sugar beet growing areas in 13 countries in Europe these include Belgium, Czech Republic, Denmark, France, German, Great Britain, Hungary, Italy, Netherlands, Poland, the Soviet Union, Spain and Sweden for a period of two years 2017 and 2018. In a field trial, the number of genetic lines tested are laid out using an Alpha design. Alpha designs described by Patterson and Williams (1976) consist of incomplete blocks nested within full blocks called replicates. The total number of plots for a field trial is the product of the number of lines to be tested and the number of replicates. A series represents a collection of lines which are to be tested in different locations. In other words, a typical series consists of multiple yield trials in different countries and includes a set of common checks. Field specialists monitor the different plots for emergence, pests and disease tolerance, bolting tolerance, and other plant characteristics throughout the growing season. At the end of the growing season, research staff harvests the field trials with specially-designed research harvesters to measure each variety for root yield (Net weight per plot (KG/PL)). Two samples are taken from each plot, and these samples are identified using a unique plot number. One of these samples is analyzed in the laboratory where four traits are measured these include the percentage of sugar (%S), Potassium (mM_k), Sodium (mM_Na) and Nitrogen (mM_N). The other sample is only analyzed in case of an unusual(outlying values) measurement values are observed (quality checks).

1.3.2 Sugar yield and yield quality traits

The main trait of interest is the recoverable sucrose percentage, which is referred to as total recoverable sugar percent denoted (%S). The other traits of interest are the impurities that are

Potassium (mM_k), Sodium (mM_Na), and Nitrogen (mM_N) measured in Moles. The other traits of interest are derived from the ones mentioned previously these include:

1. Netweight per Hactare (T_HA) which a Net weight per plot (KG/PL) corrected for the field trial and is calculated as: T_HA = KG/PL×correction of field trial×correction of plot. In most cases these correction factors are equal to 1 hence T_HA == KG/PL

2. Sugar yield (S_HA) which is the tonnage of sugar beets per hectare(T_HA) multiplied by the amount of sugar ($\frac{\%S}{100}$) calculated as: S_HA = $\frac{\text{T\_HA}\times\%S}{100}$

3. Percentage of white sugar (%WS) calculated as total recoverable sucrose percent (%S) corrected for the impurities. %WS = %S−[(0.14×(mM_k)+mM_Na)+(0.25×mM_N)+0.5]

4. white sugar yield (WSY) calculated is the percentage white sugar corrected for tonnage per hectare(T_HA) i.e. WSY = $\frac{(\text{T\_HA}\times\%WS)}{100}$

2 Methodology

2.1 Exploratory Data Analysis

Exploratory data analysis (EDA) is a crucial step of analysis in this research where we visualize, plot, and manipulate data without making any assumptions. The primary aim of the exploratory analysis is to examine the data for distribution and anomalies to assess the quality of the data and come up with a better strategy for model building (Komorowski et al., 2016). EDA techniques adopted for this paper include graphical techniques, box plots with a quantitative techniques, correlation analysis. Box plots of trait against field trials to show variation in phenotype data within field trials and between field trials in the same series. Measures of central tendency and dispersion were also calculated as part of the descriptive statistics for each trait. For bivariate relationships, correlation analysis was done to give an overview of the interdependences among the traits.

2.2 Correlation Analysis

Relationship strength between the traits was assessed in the correlation analysis. This enables us to establish if there is a possible association between variables (Chan, 2003). A Pearson correlation coefficient quantifies the correlation of two traits denoted as ρ for the population correlation and the sample correlation $r_{(y_i,y_j)}$ calculated as:

$$r_{(y_i,y_j)} = \frac{cov(y_i, y_j)}{\sqrt{Var(y_i).Var(y_j)}} \tag{2.1}$$

Where y_i and y_j are any two phenotypic variables and $r_{(y_i,y_j)}$ ranges from $[-1; 1]$, the direction of the relationship is indicated by the sign. Hypotheses tests for the significance of the linear relationship between traits i.e., $H_0 : \rho = 0$ vs. $H_1 : \rho \neq 0$ were also performed at 5% level of significance. Results from correlation analysis presented in correlation matrices.

2.3 Modelling in plant breeding experiments

Models adopted in plant breeding trials are usually ANOVA models, which takes into account the design of the field trial experiments. In this section, the two major types of models the fixed effects model and mixed effects model are described highlighting the significant differences between them. Furthermore, a description of the use of these two types of models in the analysis of the univariate and multivariate traits in sugar beet field trials.

2.3.1 Fixed effects model

In the fixed effects framework, all levels of a fixed factor are included in the model experiment, and inference is specific to these treatment levels. The interest of the researcher is testing the difference in means between treatments (Van Eeuwijk et al., 2011). Furthermore, the fixed effects model is based on the assumptions that the error terms are random, independent, and normally distributed with mean zero and have constant variances. The general form of a fixed effect model is given as:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{2.2}$$

where

$$\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \boldsymbol{\sigma}^2\mathbf{I})$$

$\mathbf{Y}$ is a matrix of response observations (trait), $\mathbf{X}$ is a design matrix i.e a matrix of the predictor variables, $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_p)'$ is a vector of fixed parameter estimate

In this study, a fixed effects model is the one that treats the genotype, and any other design effects as fixed effects. However, design factors impose restrictions on randomization that induce correlations for example, observations made within the same random block, where this block is one out of a population of blocks, are correlated (Van Eeuwijk et al., 2011). Such correlations should be included in the model to make valid inferences. Related genotypes and field heterogeneity also impose a correlation, and on top of these, there might be a correlation between environments (Gutierrez, 2012). Correlations within trial genetics and error variances are affected by the environment; therefore, different variances are also common in field experiments. The fixed effects linear model using the ordinary least squares estimation procedures is, therefore, too restrictive to perform adequate data analyses for data from most breeding programs because of the independence assumption. In the real world experiments, the error structure is generally a lot more complex than used in standard linear models for ordinary data analysis (Stroup, 1989). In this thesis, a fixed effect model is considered under the randomized complete block design analysis in the first stage of the two-stage analysis i.e., the genotype and the repetition effect are fixed.

2.3.2 Mixed effect models

In contrast to the fixed effect model mixed model analysis applies to research involving factors with a few levels that usually can be controlled by the researcher (fixed) as well as factors with levels that are beyond the researcher's control (random). Levels of the random factor are considered as a random sample from a population of factor levels and inference is about a population of factor levels. In the case of random treatment effects, the researcher's interest is in testing the population variance of a treatment (Van Eeuwijk et al., 2011). The general linear mixed model can easily accommodate covariances among observations i.e., correlated data by including random effects and estimating their respective variance components to model variability in addition to the residual error (Balzarini, 2002) and (Wolfinger and Tobias, 1998). In Multi-Environment Trials, mixed models facilitate the modeling of heterogeneity of genetic variances and correlations between environments on top of modeling the design features and spatial trends in individual trials (Van Eeuwijk et al., 2016). Furthermore, mixed model approaches are preferred to the ordinary ANOVA models because of the estimation procedures usually involved can overcome the troubles for handling unbalanced and incomplete data (Balzarini, 2002). The Mixed model formulation is given as:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\varepsilon} \tag{2.3}$$

where $\mathbf{Y}, \mathbf{X}, \boldsymbol{\beta}$ and $\boldsymbol{\varepsilon}$ are in that order the vector of continuous responses, the design matrix of predictors, the vector of fixed effects and the vector of residual error terms. Whereas $\mathbf{Z}$ and $\mathbf{b}$ are the matrix of covariates and corresponding vector of random effects.

$$\begin{bmatrix} \mathbf{b} \\ \boldsymbol{\varepsilon} \end{bmatrix} \sim N\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix} , \begin{bmatrix} \mathbf{G} & \mathbf{0} \\ \mathbf{0} & \mathbf{R} \end{bmatrix} \right)$$

$\mathbf{G}$ and $\mathbf{R}$ are the variance-covariance matrices of the random effects and the random residual error terms, respectively. The simplest form for $\mathbf{G}$ and $\mathbf{R}$ arise from independence and constant variances of the random effects and the error terms (the independent variance-covariance structure). Another form of these covariance matrices is the unstructured model in which all the elements of the matrices can be different. Intermediate structures of $\mathbf{G}$ and $\mathbf{R}$ (e.g., the compound symmetry) are mostly used since they allow modeling correlations with a smaller number of covariance parameters than the unstructured one. Generally, genetic correlations are introduced into the model trough $\mathbf{G}$ and the off-diagonal elements of $\mathbf{R}$ model experimental

correlations among observations. When data are indexed in space, the covariances in $\mathbf{R}$ may reflect correlations due to the spatial arrangement of the experimental units (Balzarini, 2002). In this thesis, all the other models except the first-stage RCBD model are mixed effects models.

The breeding programme uses estimated breeding values to select genotypes for crossing, and monitor genotype performances with the aim of producing genetically improved varieties. The mixed model offers an ideal platform for both estimating genetic variances as well as predicting the breeding values of individual genotypes (Littell et al., 2002). Plant breeding based on mixed models, use Best Linear Unbiased Predictors (BLUPs) to predict breeding values. BLUPs are obtained by averaging over levels of fixed effects $\boldsymbol{\beta}$ in the model, average over interactions between the random effect level of interest and all levels of other fixed effects, while ignoring other random effects, and include the intercept (μ) plus the effect prediction for a level of the random factor (Isik et al., 2017). A conditional BLUP, which is predicted for a specific set of random factor levels is a more useful type of BLUP in plant breeding. Thus the BLUP is a conditional expectation of the random effects given the data BLUP $= \mu + E(\mathbf{b}_i|y)$ (Littell et al., 2002).

Robinson et al. (1991) postulates that the BLUP-based selection method gives more accurate predictions of the genetic effects than the Best Linear Unbiased Estimates (BLUEs) -based methods. BLUEs are the predicted marginal mean value of a fixed effect is obtained as a least square mean. The least square mean for a level of a factor is an average over levels of other fixed factors and interaction effects involving the factor level in the model. The gain in accuracy by BLUPs is a result of the shrinkage property that is the individual genetic means are shrunken towards the overall mean according to their information (Piepho et al., 2008). The amount of shrinkage is also influenced by environmental variation. If BLUPs are used then there is shrinkage of genetic effects.

2.4 Within-Trial Variation

A trial location is seldom uniform across its whole area. Soil variation and the interactions of genotypes with that variation are primary factors that breeders need to accommodate (Kempton et al., 2012). Plant breeders try as much as possible to select homogeneous field sites for experiments through blocking, but some fields may show some heterogeneity due to differences in soil type, fertility, water retention capacity, among others. As the number of genotypes

becomes larger, the partitioning of the experimental fields into blocks of homogeneous experimental units turns out to be more difficult. The Incomplete block designs like the alpha design were developed to handle this situation. However, the Incomplete block designs introduce unbalance into the experimental design, which can be well handled by mixed models' analysis that simultaneously estimates the effects of random complete blocks and incomplete blocks along with the random or fixed treatment effects (Isik et al., 2017). The efficiency of incomplete block designs is high, given that the experimental units within an incomplete block are homogeneous, but there is no guarantee that this will be the case in most field experiments.

The incomplete block designs are usually less likely to capture the heterogeneity of the experimental fields. Row-column design allows for two-dimensional blocking hence simultaneously control two sources of variability (Van Eeuwijk et al., 2011). An additional source of variability in the fields is induced by the plot to plot interference i.e spatial variation (Kempton et al., 2012). In this study, the within-field trial variation was modeled in different forms, and a model that optimally model the non-genetic variance in the field experiments was chosen to give an improved estimation of the genetic means.

2.4.1 Randomized Complete Block and Row-Column Designs

The randomized complete block design approach helps in handling heterogeneity in the experimental units by fitting the main effects of blocks in the analysis. The degree to which plots in several repetitions(complete-blocks) are different is on average attributed to the complete-block effects, and the variation explained by the block effects is accounted for in the model itself, and so reducing the residual error variance compared to an analysis that ignores block effects. Incomplete block and row-column designs extend this concept further, absorbing additional variation due to the effects of incomplete blocks within repetitions or complete blocks, at an expense of more parameter estimates (Isik et al., 2017). Hence these incomplete block and row-column design effects were included in the analysis of field trials designed to evaluate breeding values or genetic values of the sugar beet. The models are explained below:

Models 2.4 is a randomized complete block design model; all components of this model are fixed. On the other hand, 2.5 is an Alpha (Incomplete block) design model with random incomplete blocks nested within replicates hence, it is a mixed effects model. These two models consider

one a dimensional blocking structure.

$$y_{ij} = \mu + G_i + r_j + \epsilon_{ij} \tag{2.4}$$

$$y_{ik(j)} = \mu + G_i + r_j + b_{k(j)} + \epsilon_{ijk} \tag{2.5}$$

where $y_{ik(j)}$ is the trait observation of the i^{th} object in k^{th} incomplete block placed within j^{th} repetition (Complete super-block) , μ is the intercept term, G_i is the i^{th} object effect of the genetic line, r_j is the effect of the j^{th} repetition, $b_{k(j)}$ is the k^{th} incomplete block effect within j^{th} replicate. The error terms are random and normally distributed i.e $\epsilon_{ij} \sim N(0, \sigma^2)$ The random block analysis is a generally preferred analysis it allows an additional use of inter block information(Van Eeuwijk et al., 2011). This imposes additional assumptions on the model, the extra assumptions are that the random block effects $b_{k(j)} \sim N(0, \sigma_b^2)$ and the random blocks effects are independent of the error terms $cov(b_j, \epsilon_{ij}) = 0$.

In addition to the models described above, we also consider correcting for the row-column design. Model 2.6 is developed to account for the likely two-dimensional (in row and column direction) variation on the field.

$$y_{irl} = \mu + G_i + R_r + C_l + \epsilon_{irl} \tag{2.6}$$

y_{irl} is the phenotypic observation of the i^{th} object placed within the r^{th} incomplete row of the l^{th} incomplete column. R_r is the r^{th} incomplete row effect, the C_l is the l^{th} incomplete column effect. These two factors C_l and R_r are assumed to be random. The error term (ϵ_{irl}) for Model 2.6, just like models 2.4 and 2.5, is assumed to be normally distributed $N(0, \sigma^2)$.

2.4.2 Spatial Variation

A model with random blocks implies that the plots in each block are uniformly correlated i.e., the yields are correlated in the same way irrespective of distance apart, this assumption is less likely to be practical (Negash et al., 2014). Agronomists have argued that in a field trial plots adjacent to each other share the same factors that generate micro-environments, therefore, yield and other traits of plots close together in rows or columns are likely to be more highly correlated than plots further apart hence the elements of the residual ($\boldsymbol{\varepsilon}$) are correlated (Rodríguez-Álvarez et al., 2016). This correlation is usually a function of the distance between

plots; hence, the model usually used in this case is an autoregressive correlation structure (AR) model (Kang, 2002). An AR(1) model requires that the yield of a plot is more affected by the yield of its neighboring plot, but not directly by the yield of plots further apart. The model takes into account the natural variation as the direct product of an autoregressive correlation structure for columns and an AR correlation structure for rows, denoted by $AR(1) \times AR(1)$. This structure is equivalent to a correlation r and c say for two plots side by side, r^d for two plots at a distance d apart along the row direction and c^f for plots at distance f apart along the column direction (Burgueño et al., 2000). An AR(2) model would imply that the direct influence is from the neighboring plot and the next neighbor. It is necessary to correct for this spatial variation when estimating genotypic effects hence we fit the spatial model.

$$y_{ikl} = \mu + G_i + \epsilon_{ikl} \tag{2.7}$$

This model assumes that plots adjacent to each other are correlated i.e an AR(1) model hence accounts for the spatial trends on the field. The error term in model 2.7 is spatially correlated which is assumed to follow a normal distribution with mean zero and variance covariance $= A\sigma^2$, where $A = \phi r^d \otimes \phi c^f$, ϕr and ϕc are the correlation coefficients along the row direction and column direction respectively, d= number of rows apart and f= number of columns apart.

2.4.3 Genotype Effects

Genotype effects can be fit as fixed or as random effects depending on the goal of the analysis. It is a fixed factor when the emphasis is on the comparison of tested genetic material for selection or recommendation. On the contrary, the genotype is random when the aim is to support decisions regarding elements of a breeding strategy by estimating the variance components, genetic parameters, and focus on predicting the potential breeding value of genotypes in future experiments using the BLUPs (Fischer et al., 2009). In the case of random genetic effects, the genotypes are taken as a representative sample of the relevant genetic base i.e., $G_i \sim N(0, \sigma^2_G)$ hence they is an additional source of variation. Furthermore Piepho et al. (2008) postulates that in most analysis of plant breeding trials, genotype effect should be random because the selection of varieties through rankings rather than comparisons is the main goal in both early breeding phases or advanced evaluation phases. In this thesis genotypes are taken as fixed when in the fist stage of the two-stage analysis and random otherwise.

2.5 Multi-Environment Testing ($G \times E$ Interaction)

The main aim of sugar beet breeding programs is to provide farmers with seed genotypes with an assured superior performance in terms of yield and other quality characteristics across a range of environments. The final composition of the sugar beets is the collective result of several underlying interactions between the genetic make-up of the plant and the conditions i.e., environment under which the beets are grown. Environments vary in the amount and quality of inputs and stimulus that they give to the plant, for example, the amount of water, nutrients, or sunlight. One major aim of plant breeding is matching genotypes and environments in such a way that improved quality beets are obtained. Some genotypes perform well across a wide range of conditions while others do better than others under restricted conditions these are called adapted genotypes. Adaptation of genotypes is related to the phenomenon of genetic by environment interaction ($G \times E$) (Malosetti et al., 2013).
Kempton et al. (2012) defines $G \times E$ interaction as the differential expression of genotypes across locations. This interaction reduces the relationship between phenotypic and genotypic values and may lead to the poor performance of selections from one environment when exposed to another environment. This forces breeders to examine genotypic adaptation. Multi-environment trials (METs) enable the assessment of the probable yield performance of several varieties across a range of environments and possibly over years or a combination of the two as in the case of this study. This allows the researchers to obtain information about whether a genotype performs well in all environments, or under which environment can it give better yield and quantify how much can a genotype gain from improving the environment (Gutierrez, 2012).

Analysis of data from METs can be done either by one stage or two stage modeling. Both types of analysis were done in this paper, and comparisons of the estimated breeding values and their respective rankings were done.

2.5.1 Two stage approach

First Stage

For the analysis of MET data using a two-stage approach explained by (Van Eeuwijk et al., 2016), in the first stage individual field trials are analyzed with models including terms of the design features i.e., the completely randomized design, the alpha design, the row-column design,

and spatial models described earlier. This stage aims to find one best model for each trial that best explains the variability within a field trial. Table 1 shows a summarized description of these for models. The best model is picked using the Akaike's Information Criterion, and adjusted genotypic means γit for trial t are estimated. For this thesis, this first stage done through the selection of the best model is referred to as the best-selection method. Different two-stage analyses were adopted to fit each of these four models to the field trial data separately and extracting the BLUEs for use in the second stage these forms of analysis were compared to the best-selection method.

Table 1: Models for within field trial variation

Model		**Fixed effects**	**Random effects**
2.4	Randomized complete block design	genotype + repetition	unit
2.5	Alpha design	genotype + repetition	blocks within repetitions + unit
2.6	Row column design	genotype	row +columns
2.7	Spatial	genotype	row(AR(1)) $\otimes$ column(AR(1))

Second Stage

In the second stage a weighted analysis is done. Here we fit the mixed model for the METs the model is defined as follows

$$\gamma_{it} = \mu + G_i + E_t + GE_{it} \tag{2.8}$$

Where γ_{it} is trait mean of the i^{th} genetic effect G and the t^{th} trial location E. The Environmental effect E_t is fixed while the genetic effect G_i is considered random i.e $G_i \sim N(0, \sigma_G^2)$. GE_{it} is the residual term since there is no replication of $G \times E$ effects within environment and it follows normal distribution with mean zero and variance σ_{GE}^2. This model produces the BLUPs of the genotypes i.e., shrunken means which are sensitive parameters for each of the genotypes.

(Möhring and Piepho, 2009) discussed different weighting methods that can be adopted in a two-stage analysis and further postulates that if a correct weighting is done for the two-stage analysis, the results will not be statistically different from the ones from a one-stage approach. In this thesis, we consider an unweighted two-stage analysis and compare the results to those obtained from each of the one stage analyses. Most scholars prefer the two-stage approach since a large number of field trials are usually anticipated in plant breeding data; hence, probable

mistakes are fewer compared to the one stage approach. SESVANDERHAVE use the best-selection two-stage approach for their analysis since they do not have to wait for all the data from all field trials to be available before they start to analyze.

2.5.2 One stage approach

In the one-stage approach field-plot data (information from the design of the experiment) and the $G \times E$ interaction are analyzed simultaneously. An extension of each of the four models described in the first stage of the two-stage approach was done to the one stage approach.

The primary purpose of MET analyses is to find an adequate model for the mean trait (phenotypic) responses as a function of genotypes and the environment. Reliable conclusions from the models are determined by a suitable variance-covariance structure for the GE_{it}. Assumptions on the variance of the $G \times E$ interaction entails the different number of model parameters (Ward et al., 2019). The assumption of independence of residuals between environments is highly unrealistic, i.e., this implies that covariance between every pair of different environment random terms is zero, or in other words, the environments provide independent information. Finding an appropriate structure for residuals that reflects the difference of genetic variances and correlations is a crucial step towards reliable conclusions on mean genetic effects (Balzarini, 2002). For compound symmetry variance-covariance structure, each environment has the same genetic variance that is confounded with $G \times E$ variance, and the genetic correlation is uniform between all pairs of environments. If we assume that each environment has its own genetic variance, and the genetic correlation can change between any pair of environments then it implies that the variance-covariance structure is unstructured. Ward et al. (2019) asserts that an unstructured variance-covariance matrix results in too many parameters i.e., $\frac{T(T+1)}{2}$ parameters were T is the total number of trials that need to be estimated which may lead to variance components estimate failing to converge especially when there is a large number of environments. Thus since the independence and unstructured suffer from the highlighted disadvantages. This study adopted a compound symmetry variance-covariance structure.

2.6 Univariate versus Multivariate Models

We give an overview of the two mixed model types that were fitted to the data i.e., univariate and multivariate trait models that are adopted in this study with the methods described earlier.

2.6.1 Univariate Trait Models

The univariate trait models independently model each of the eight traits with a separate model. Hence it assumes that the trait measurements within an individual plot are independent. We fit the model explained in equation 2.3

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\varepsilon} \tag{2.9}$$

where $\mathbf{Y}$ is the response vector of a trait for a given field trial(location), $\boldsymbol{\beta}$ and $\mathbf{b}$ are vectors of fixed effects and random effects respectively. $\mathbf{X}$ and $\mathbf{Z}$ are the design matrices for the fixed and random components respectively, and $\boldsymbol{\varepsilon}$ is a vector of error terms. The univariate trait models were done using both that two stage and the one stage methods and comparison of the predicted breeding value rankings from each of these methods were compared. The variations of models fitted in this thesis under the univariate modeling were the two-stage model using the best selection method, randomized block design model (RCBD), incomplete block design (Alpha design) model, spatial model and the row column model. One sage models of these types were also fitted for comparison purposes.

2.6.2 Multivariate Models

Multivariate analyses are necessary to obtain estimates of genetic associations between traits. Genetic correlations between traits are an indication that measurements of one trait contain information about other traits. On the other hand, observed measurements of the traits from the same plot are often correlated, environmental factors and genetic effects contribute to observed correlations among traits (Isik et al., 2017). The univariate analyses described before assume that all correlations between traits are zero. Therefore the univariate models exclude any opportunity of learning from the potential associations among traits because a single, independent model is fitted for each trait separately. If the assumption of zero correlation is not correct, then genetic selection performed and the estimated breeding values based on the single trait models may be biased. The critical assumption of the multivariate extension is that trait measurement taken from the same plot are correlated; hence, joint analyses of correlated

traits utilize information on the association between traits gives more accurate predictions (Meyer, 1991). We fitted a model of the form:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Zb} + \boldsymbol{\varepsilon} \tag{2.10}$$

The dimension of the matrices are as follows

$$\mathbf{Y}_{n\times q} = \mathbf{X}_{n\times(p+1)}\boldsymbol{\beta}_{(p+1)\times q} + \mathbf{Z}_{n\times k}\mathbf{b}_{k\times q} + \boldsymbol{\varepsilon}_{n\times q}$$

n is the number of individual experimental plots, and q is the number of dependent variables i.e., traits. $\mathbf{X}$ is the design matrix with dimensions $n \times (p+1)$, where p is the number of fixed predictors for one trait and an additional column is added for the intercept. $\boldsymbol{\beta}$ is the matrix of fixed effects to be estimated with dimensions $(p+1) \times q$. The rows of $\boldsymbol{\beta}$ correspond to number predictor variables and the columns to the number of response variables. The design matrix $\mathbf{Z}$ has dimensions $n \times k$ where k is the number of random effects per trait, and $\mathbf{b}$ is an $k \times q$ matrix of random effects. Other components of the model are explained under equation 2.3.

Selecting the covariance structure

The structure of variance-covariance matrices of the random effects $\mathbf{G}$ and the residuals $\mathbf{R}$ in a multivariate model have different specifications. Many factors were considered in selecting the structures which include the number of parameters estimated in line with the principle of parsimony as well as providing answers to some research questions of interest (Kincaid, 2005). In this thesis we, therefore, start with the most straightforward extensions of univariate models to the multivariate approach. Of interest also in this study is the genetic variance-covariance structure estimated by the matrix $\mathbf{G}$. $\mathbf{G}$ allows to study and summarize a portion of the variation in the traits that is due to the genetic variability i.e., pairwise genetic covariances between characteristics of interest. Pigliucci (2006) asserts that the $\mathbf{G}$ matrix describes the degree to which the genetic architecture determines how a population respond to natural selection. Positive genetic correlations between traits imply that other factors being constant if selection favors an increase in one trait the other trait will be indirectly lifted upwards. On the other hand, a negative genetic correlation any selection favoring an increase in one trait will decrease the mean response of the other trait. A weak correlation implies the traits respond independently to selections of each other. On the other side, the $\mathbf{R}$ matrix of residuals takes into account the correlation between the residual effect (Pigliucci, 2006). To correct for correlation between

the residuals, an unstructured **R** matrix was chosen i.e., the correlation between traits measurements within each plot is unique for every pair. For a simple method, a simpler covariance structure for the random effects was chosen. However, to gain information about the genetic correlations between traits, an unstructured variance structure for the random genotype effects is needed. On the other hand, to estimate the phenotypic correlations between traits, we need unstructured covariances of the genotype, $G \times E$ and the residual effect between trait pairs (Isik et al., 2017). genetic correlation between two traits y_1 and y_2 is calculated as

$$r_{G(y_1,y_2)} = \frac{Cov_G(y_1, y_2)}{\sqrt{Var_G(y_1).Var_G(y_2)}} \tag{2.11}$$

While the phenotypic correlation is given as

$$\begin{aligned} r_{P(y_1,y_2)} &= \frac{Cov_P(y_1, y_2)}{\sqrt{Var_P(y_1).Var_P(y_2)}} \\ &= \frac{\sigma^2_G(y_1, y_2) + \sigma^2_{GE}(y_1, y_2) + \sigma^2_\varepsilon(y_1, y_2)}{\sqrt{(\sigma^2_G(y_1) + \sigma^2_{GE}(y_1) + \sigma^2_\varepsilon(y_1)) \times (\sigma^2_G(y_2) + \sigma^2_{GE}(y_2) + \sigma^2_\varepsilon(y_2))}} \end{aligned} \tag{2.12}$$

Where $\sigma^2_G(y_i)$, $\sigma^2_{GE}(y_i)$ $\sigma^2_\varepsilon(y_i)$ are the genotype, $G \times E$ interaction and residual covariance parameters corresponding to trait i.

2.7 Model selection

The analyses of field trial data are aimed at selecting a parsimonious model that accurately and characterizes the variation effects in the field, and consequently provides the most accurate and precise estimates of genotype effects. The relevance of component terms in the model is assessed by procedures such as the likelihood ratio test, Akaike's information criterion (AIC) or Schwarz's Bayesian Information Criteria (BIC) (Piepho and Williams, 2010). In this thesis, the AIC criterion was used for model selection. AIC is computed as minus twice the restricted loglikelihood (the so-called deviance), plus a penalty term defined as twice the number of fitted parameters i.e.,

$$AIC = -2\text{loglikelihood} + 2K$$

where K is the number of parameters in the model. The penalty term is designed to strike the right balance between model practicality and parsimony. Burnham and Anderson (1998) postulates that the philosophy behind the use of AIC is the view that there is no correct model and the best we can hope for is to find a good working model that is close enough to the correct underlying model. Minimization of AIC is a strategy to search for an approximating model

that shows the smallest discrepancy to the correct model and is parsimonious enough to be supported by the data (Piepho and Williams, 2010).

2.8 Criteria for comparing models

Different models were used to estimate the breeding values i.e., two-stage models, one stage models, and the multivariate models described previously. The Spearman's rank correlation of the predicted breeding values was used as a performance measure (Zambrano et al., 2015). The Best-selection two-stage model was used as the benchmark for the assessment. Möhring and Piepho (2009) postulates that since the main focus of such an analysis is on comparison of genotypes the Mean Square Prediction Difference (MSPD) of the estimated difference between estimated breeding values from the benchmark method compared to the other methods.

2.8.1 Spearman's rank correlation coefficient

This measure was chosen as an attempt to measure the degree of similarity between the rankings of the genotypes by their breeding values. It is calculated using the equation

$$r_s = 1 - \frac{6\sum_1^n(d^2)}{n^3 - n} \tag{2.13}$$

Where d is the difference in rankings of the genetic values of $(y_{ij} - y_{ik})$ order. y_{ij} is the i^{th} genetic value of method j and y_{ik} is the i^{th} genetic value of method k, n is the number of (y_{ij}, y_{ik}) pairs which is the same as the number of genotypes. A high correlation implies that ranking is nearly the same regardless of the methods used (Newcom et al., 2005).

2.8.2 Mean Squared Prediction Difference

The MSPD measures the average squared distance between the predicted breeding values of the benchmark model and the other model being compared to. It helps to measure the goodness of the estimated breeding values predicted by different approaches to be used as the reasonable alternative to the currently used best-selection two stage method. MSPD is calculated as:

$$MSPD = \frac{1}{n}\sum_{i=1}^{n}(y_{ik} - y_{i0})^2$$

where y_{ik} and y_{i0} are the estimated breeding values of genotype i from method k and the benchmark model n is number of genotypes. The smaller the MSPD the more reasonable the alternative is.

2.9 Implementation of data analysis

A database of several datasets was available for analysis and validation of results. Results from two analyses are shown and the other datasets were used for validation is case of conflicting results from the two analyses. Analysis was done in SAS version 9.2 using the Proc Mixed procedure.

3 Results

The methods explained in section 2 were implemented in providing answers to the research questions of this study. This section of the thesis is organized as follows the exploratory analysis of traits is presented first, comparison of various univariate methods of estimating the breeding values. Lastly, a comparison of the multivariate model with the best-selection two-stage analysis and the best one-stage model. Since a big database of different series of experiments was provided, two datasets were used to run parallel analysis. The parallel analysis allows for consistency checks and robustness of findings to changes in datasets i.e., the genotypes and field trials under consideration.

3.1 Exploratory Analysis

Two series of experiments considered for the main analysis are series 1111, and series 1131. In the case of conflicting results, other datasets were taken for analysis and validation of results. The number of field trials, plots, and genotypes used in each series of experiments is shown in Table 2

Table 2: Distribution of field trials and genotypes

Series	Field trials	Plots	Genotypes
1111	33	4941	72
1131	9	972	54

3.1.1 Descriptive Statistics

The mean T_HA recorded was 89.77t/ha and 79.35t/ha for series 1111 and 1131 respectively the range of values for T_HA were wider for series 1111 (16.3 -137.35t/ha) than series 1131 (33..51 -118.58t/ha). The range of values shows that there are some plots with extremely below average T_HA and some with extremely high T_HA. On the other side, the average %S extracted from the beets is almost the same in the two sets of data i.e., approximately 17.8% and the variation in the percentage of sugar is small. These are only insights into the distribution of the data in the two series of experiments. In addition to these box-plots of the traits are shown to give a pictorial view of the variability within and between field trials.

Table 3: Summary Statistics

	Series 1111				Series 1131			
Variable	**Mean**	**Std Dev**	**Min**	**Max**	**Mean**	**Std Dev**	**Min**	**Max**
T_HA	89.77	19.21	16.3	137.35	79.35	16.24	33.51	118.58
S_HA	15.97	3.52	3.21	24.04	14.00	2.36	6.17	19.99
%WS	16.42	1.26	11.99	21.21	16.21	1.64	12.20	20.35
WSY	14.74	3.28	3.03	22.09	12.69	2.08	5.65	18.37
Mm_K	3.58	0.75	2.13	7.89	3.77	0.58	2.46	5.39
Mm_Na	0.42	0.24	0.10	1.93	0.56	0.45	0.10	2.75
Mm_N	1.25	0.74	0.26	4.40	2.03	1.24	0.55	8.12
%S	17.79	1.19	13.86	22.82	17.83	1.49	13.94	21.70

3.1.2 Box-plots of Traits

Figure 1 displays the boxplots for each trait separately for several field trials included in the study series 1111. For T_HA, the median T_HA for various field trials fluctuates around the grand median value of 97 t/ha. Variation in T_HA for plots within the same field trial and between field trials is small. Field trial 734 has the lowest recorded T_HA. On the other hand, the boxplots shows some variation in the amount of %S among different field trials. Furthermore, the variability in %S between plots within the same field trial differs for different field trials as shown by the different lengths of the box and whiskers. The boxplots show that T_HA, S_HA and WSY are similar as well as that of %S extracted and %WS this suggest that there may be high inter-dependencies amongst some of these traits. These relationships were further explored using the Pearson product moment correlation analysis of traits. There seems to be considerable variability in Mm_K content between field trials with some field trials with Mm_K content below the grand meanwhile others are above average. Similarities in variability patterns for the other impurities (Mm_Na and Mm_N) i.e., field trials with smaller variation in sodium also have smaller variation is nitrogen. All in all the boxplots show some outlying plots that have either low measurements or high measurements of the trait. This variability can be either due to different environmental and management conditions at different trial locations or most importantly, due to the effect of the genotypes used.

Figure 2 shows boxplots from series 1131; however, it still shows the inter-dependencies amongst T_HA, %S, WSY, %WS and S_HA. Variability in Mm_N content within most field trials is minimal while for Mm_K the variation is quite considerable. The difference in these two series of experiments is since a different set of experimental genotypes are used.

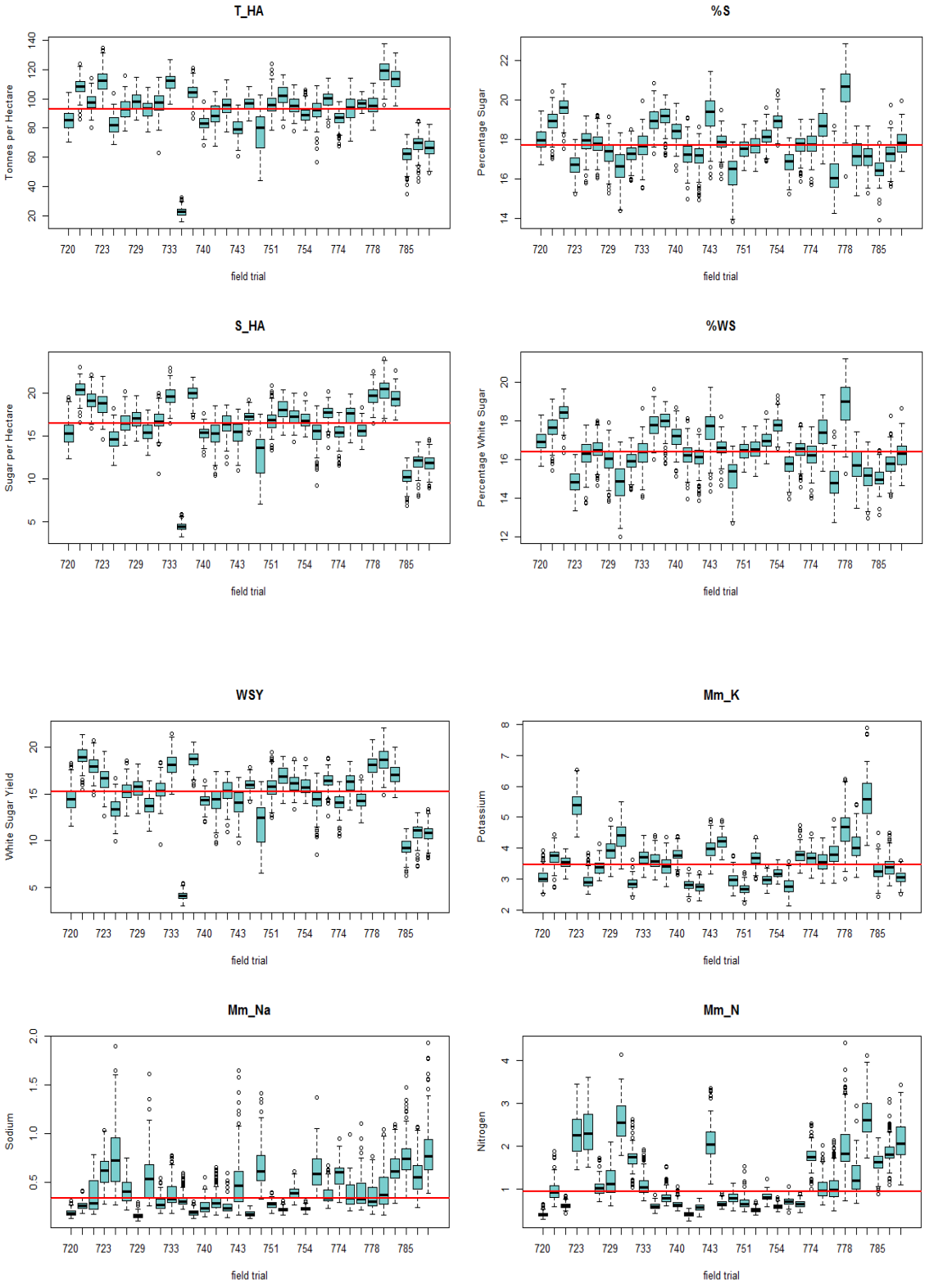

Figure 1: Boxplots of the traits series 1111

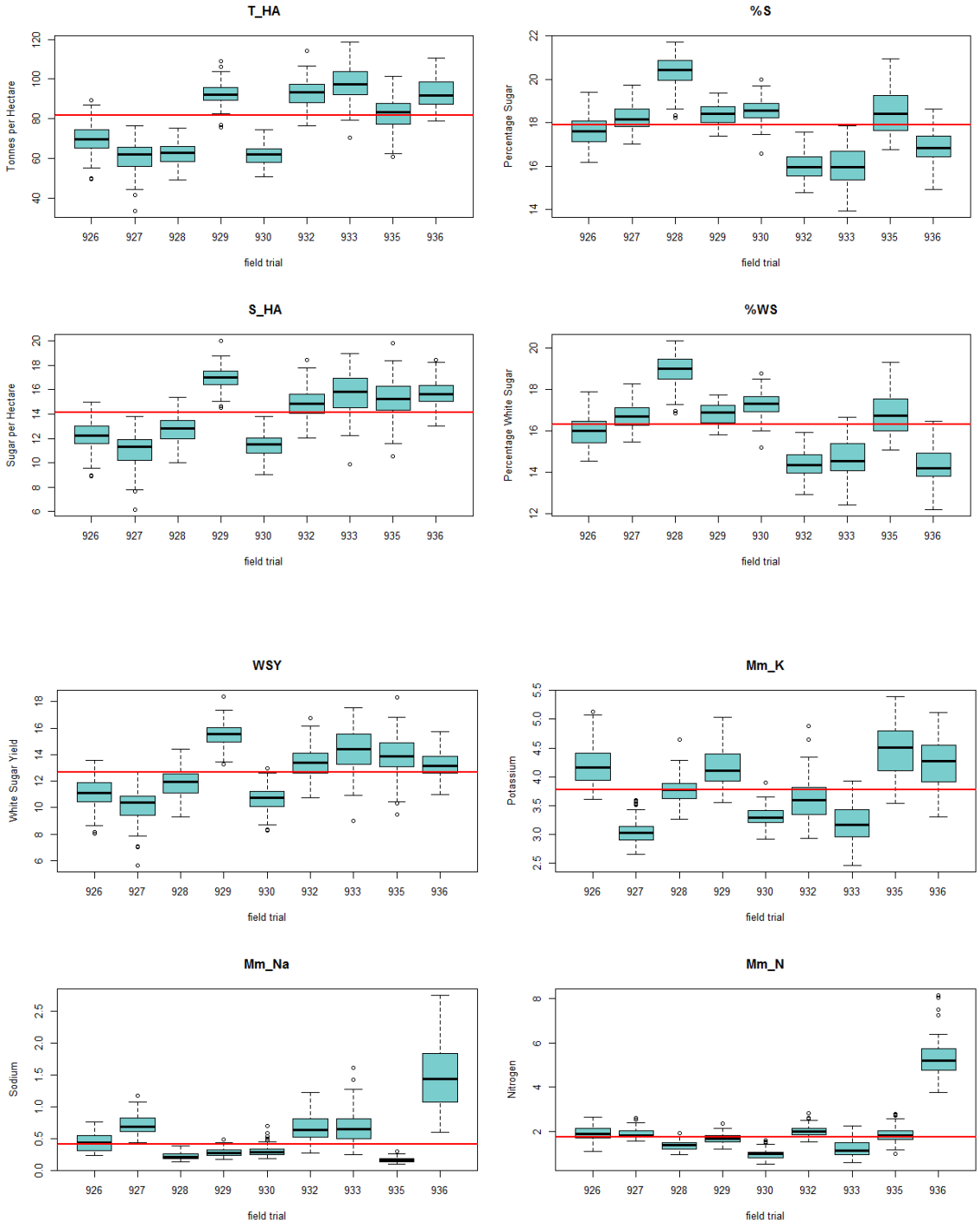

Figure 2: Boxplots of the traits series 1131

3.1.3 The Pearson Correlation Coefficient between traits

Tables 4 and 5 show the Pearson product-moment correlations between the traits and the respective test for significance at a 5% level.

Table 4 shows the correlation of traits for series 1111. The correlation matrix shows a small, weak negative correlation T_HA and %S. The correlation between the %S and S_HA is ($r = 0.218, p < 0.0001$). There is a significant moderate linear relationship between T_HA and Mm_K content in the sugar beet roots ($r = 0.322$), i.e., an increase in Mm_K results in a relative increase in T_HA . Amount of Mm_K in the beets have a non-significant linear relationship with %S extracted from the beets ($r = 0.019, p = 0.6$) but have a positive linear relationship with T_HA ($r = 0.322, p < 0.001$). Hence as the amount of Mm_K in the roots increases the T_HA increases relatively, but no significant linear relationship with %S. Mm_Na has an inverse relationship with T_HA ($r = -0.223, p < 0.0001$) that is as the amount of Mm_Na in the beets increases the weight of the sugar beet roots decrease proportionately.

On the other hand, Mm_N does not have a significant linear relationship with T_HA ($p = -0.006, p = 0.4$) however it has a negative association with %S extracted ($r = -0.177, p < 0.001$). As suggested by the box plots there are high positive correlation between %S and %WS ($r = 0.977$), S_HA and WSY ($r = 0.996$), T_HA and S_HA ($r = 0.953$) and also between WSY and T_HA ($r = 0.936$). These correlations suggest an almost perfect linear relationship amongst these traits. There is negative association between %WS and each of the impurities.

Table 4: Matrix of correlations series 1111

Variables	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1. **T_HA**	1							
2. **%S**	-0.081*	1						
3. **S_HA**	0.953*	0.218*	1					
4. **%WS**	-0.096*	0.977*	0.198*	1				
5. **WSY**	0.936*	0.255*	0.996*	0.251*	1			
6. **Mm_K**	0.322*	0.019	0.303*	-0.138*	0.237*	1		
7. **Mm_Na**	-0.223*	-0.385*	-0.324*	-0.480*	-0.372*	0.083*	1	
8. **Mm_N**	-0.006	-0.177*	-0.061*	-0.370*	-0.138*	0.479*	0.566*	1

[1] * *P value* < 0.05 *i.e., correlations significant at 5% level of significance*

Table 5 show the correlation matrix for series 1131. Results form series 1131 suggest a moderate negative correlation between T_HA and %S ($r = -0.6094$, $p < 0.0001$). The correlation between T_HA and S_HA ($r = 0.9136, p < 0.0001$) as well as with WSY ($0.8662, p < 0.0001$) is high as seen for series 1111 though with differences in magnitude. On the other hand, the correlations between T_HA and Mm_Na is moderate ($0.3314, p < 0.0001$), which implies that an increase in Mm_Na results in a proportionate increase in the T_HA. Positive correlations between T_HA and all the impurities, implying an increase in the amount of Mm_N, Mm_Na and Mm_K in the beets results in a relative increase in T_HA.

Table 5: Matrix of correlations series 1131

Variables	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1. **T_HA**	1							
2. **%S**	-0.609*	1						
3. **S_HA**	0.914*	-0.242*	1					
4. **%WS**	-0.632*	0.973*	-0.285*	1				
5. **WSY**	0.866*	-0.160*	0.98364*	-0.169*	1			
6. **Mm_K**	0.202*	0.122*	0.334*	-0.005	0.277	1		
7. **Mm_Na**	0.331*	-0.541*	0.121*	-0.663*	-0.026	-0.079*	1	
8. **Mm_N**	0.29445*	-0.27019*	0.23227*	-0.4793*	0.06154	0.3719*	0.7256*	1

[1] ** P value < 0.05 correlations significant at 5% level of significance*

The results from the correlation analysis suggest for a critical look in the genetic correlations amongst the traits. As the difference in the direction of the correlations may be due to the different sets of the genotypes used in the two series.

3.2 Univariate Analysis

Various models were fitted to the data to check similarities in the genotype rankings and estimate of the breeding values. The Best-selection method was used as the benchmark; hence, all comparisons were made against this model. The aim is to check how sensitive the rankings and the estimated breeding values are to model selections.

3.2.1 Two stage analysis

Figure 3 shows how frequent each of the four models explained earlier is picked as the best model using the Akaike's information criterion. The Figure shows that the spatial model was often picked as the best model in explaining the within-field variability for traits like %S, S_HA, WSY, Mm_Na and Mm_N. On the other hand, row-column models were also picked off the models explaining the variability in the Mm_K content of the beets in various field trials. The alpha design models were frequently picked in the modeling of the T_HA compared to the other models.

Also, separate two-stage models were fitted to the data and compared to the best-selection two-stage model comparisons are shown in Tables 6 to Table 9

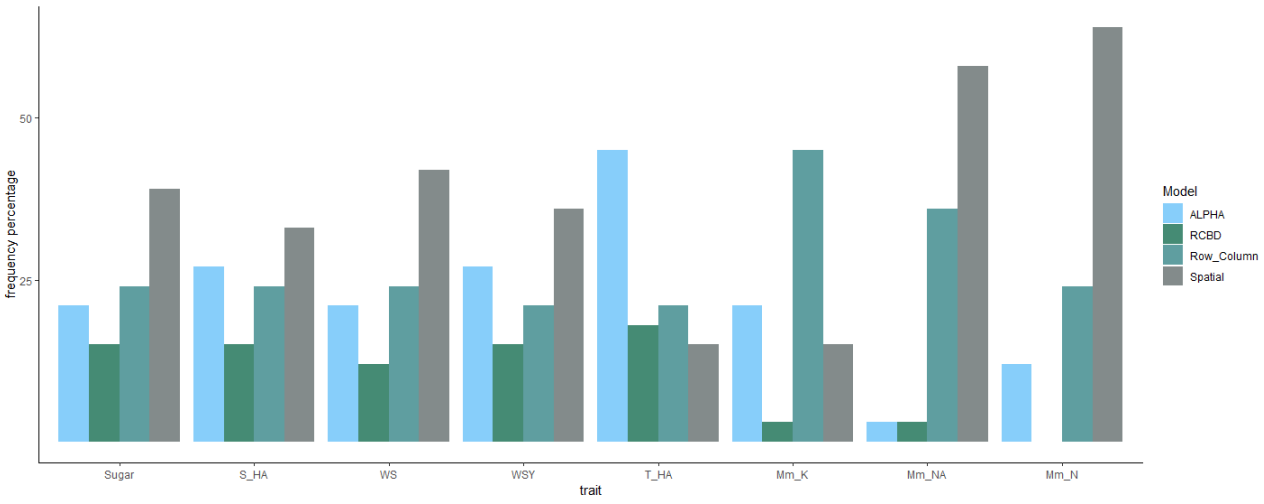

Figure 3: Frequency of selection of different models

3.2.2 One stage Models

Three models were fitted under the one stage model approach that is the RCBD model, the alpha design model, and the row-column design model. The extension of the spatial model was fruitless due to convergence problems. For each trait, the AICs of the models were recorded, and an aggregated count of which model was mostly picked as the one with the smallest AIC was recorded. Out of the 24 models fitted separately i.e., three models per trait. The row-column model was frequently picked as the best.

3.2.3 Comparison of univariate trait models and the best selection model

The main aim of the analysis was to determine if the rankings of the genotypes as well as the estimated breeding values from different models are similar. Hence the Spearman correlation coefficients and the mean square error were used.

Rank Correlations

Table 6 shows the correlation of the rankings for series 1111. For T_HA the correlation between the best-selection two-stage model and the one stage models for the RCBD, Alpha,and the row-column models were exactly the same($r = 0.987$) while, on the other hand the rankings between the best-selection models and the two-stage models for the alpha design and the row-column design are almost perfect ($r = 0.99$). For the ranking of the breeding values for percentage white sugar all the six models have an almost perfect correlation ($r = 0.99$) this implies that there is an almost complete agreement in the order of the ranks of the genotypes' breeding values. The correlations of the rankings also depend on the trait under consideration but, this correlation is larger than 0.95.

Table 6: Spearman rank correlation coefficients for comparison of different models to the best-selection two stage model series 1111.

Comparison		**T_HA**	**S_HA**	**%S**	**WSY**	**%WS**	**Mm_K**	**Mm_Na**	**Mm_N**
Best-Selection Two-Stage	**Versus**								
	One stage RCBD	0.9875	0.9769	0.9943	0.9753	0.9902	0.9832	0.9863	0.9607
	One stage Alpha	0.9875	0.9747	0.9941	0.9746	0.9904	0.9837	0.9863	0.9600
	One stage R&C	0.9872	0.9757	0.9938	0.9761	0.9902	0.9836	0.9865	0.9605
	Two stage RCBD	0.9872	0.9800	0.9947	0.9774	0.9909	0.9866	0.9881	0.9628
	Two stage Alpha	0.9944	0.9872	0.9951	0.9854	0.9975	0.9977	0.9927	0.9800
	Two stage R&C	0.9930	0.9816	0.9987	0.9760	0.9943	0.9956	0.9969	0.9863
	Two stage Spatial	0.9940	0.9935	0.9981	0.9912	0.9949	0.9929	0.9967	0.9944

[1] *all correlations are significant at* $\alpha = 0.05$ *level of significance.*

Results in Table 7 show a small change in the magnitude of the correlations. The smallest rank correlation of 0.924 was observed for the best-selection two-stage model, and one stage row column model. Like in the analysis for series 1111, the strength of the rank correlation depends on the trait under consideration.

Table 7: Spearman rank correlation coefficients for comparison of different models to the best-selection two stage model series 1131.

Comparison		T_HA	S_HA	%S	WSY	%WS	Mm_K	Mm_Na	Mm_N
Best-Selection Two-Stage	**Versus**								
	One stage RCBD	0.9842	0.9509	0.9789	0.9473	0.9789	0.9826	0.9921	0.9569
	One stage Alpha	0.9833	0.9268	0.9784	0.9309	0.9789	0.9818	0.9925	0.9530
	One stage R&C	0.9831	0.9345	0.9777	0.9241	0.9784	0.9824	0.9925	0.9539
	Two stage RCBD	0.9842	0.9525	0.9797	0.9480	0.9786	0.9826	0.9926	0.9555
	Two stage Alpha	0.9847	0.9860	0.9876	0.9842	0.9941	0.9904	0.9965	0.9871
	Two stage R&C	0.9878	0.9470	0.9911	0.9559	0.9816	0.9883	0.9970	0.9744
	Two stage Spatial	0.9806	0.9592	0.9929	0.9489	0.9896	0.9948	0.9951	0.9845

[1] *all correlations are significant at* $\alpha = 0.05$ *level of significance.*

Which shows that the robustness of the analysis depends on the data at hand i.e., for bigger datasets like series 1111 the difference in ranking for these models is very small and in most cases, the rankings are robust to the model selection. While for smaller datasets some changes are anticipated. On the other note, the rank correlations between the benchmark model and the one- stage models are lower than those between the benchmark model and the two-stage model. This difference is as a result of the anticipated loss of efficiency when modeling is done using the two-stage model. This loss in efficiency is more substantial in smaller datasets than in bigger datasets.

Overall the correlations of the rankings of the genotypes obtained from the different models are very high and positive, which shows that there are a few re-rankings of the genotypes that may result from the different models.

3.2.4 Mean Square Prediction Difference

The MSPDs values show small differences in the magnitude of the breeding values estimated from the best-selection model and those from other models shown in Tables 8 and 9. The magnitude of the MSPDs differs for each trait as well as the dataset at hand. It is observed from the tables that the MSPD values for models of the impurities are almost zero for the two datasets, which shows that the estimations of genetic values are robust to any model selection. On the other hand, MSPD values for T_HA depends on the dataset used and the models being compared. Smaller values were observed for series 1111 compared to values recorded for series

1131. The same reasoning as in the correlation is in the loss of efficiency in two-stage models, the MSPD values between the Best-selection two stage model and the one stage models are bigger that comparisons to two stage models.

Table 8: MSPD comparison of the Univariate Models to the two stage best-selection model series 1111.

Comparison		T_HA	S_HA	%S	WSY	%WS	Mm_K	Mm_Na	Mm_N
Best-Selection Two-Stage	**Versus**								
	One stage RCBD	1.2320	0.0448	0.0014	0.0393	0.0021	0.0006	0.0002	0.0014
	One stage Alpha	1.2432	0.0466	0.0015	0.0408	0.0022	0.0006	0.0002	0.0014
	One stage R&C	1.2656	0.0463	0.0015	0.0405	0.0021	0.0006	0.0002	0.0014
	Two stage RCBD	0.0870	0.0027	0.0009	0.0025	0.0012	0.0003	0.0000	0.0004
	Two stage Alpha	0.0406	0.0015	0.0031	0.0003	0.0000	0.0000	0.0000	0.0002
	Two stage R&C	0.0631	0.0025	0.0007	0.0023	0.0007	0.0000	0.0000	0.0002
	Two stage Spatial	0.0612	0.0013	0.0003	0.0011	0.0004	0.0001	0.0000	0.0000

Table 9: MSPD comparison of the Univariate Models to the two stage best-selection model series 1131.

Comparison		T_HA	S_HA	%S	WSY	%WS	Mm_K	Mm_Na	Mm_N
Best-Selection Two-Stage	**Versus**								
	One stage RCBD	0.1937	0.0078	0.0049	0.0071	0.0051	0.0007	0.0004	0.0013
	One stage Alpha	0.2482	0.0165	0.0037	0.0154	0.0058	0.0008	0.0004	0.0019
	One stage R&C	0.2292	0.0131	0.0059	0.0120	0.0059	0.0007	0.0004	0.0016
	Two stage RCBD	0.1874	0.0047	0.0046	0.0044	0.0045	0.0006	0.0000	0.0014
	Two stage Alpha	0.1376	0.0026	0.0020	0.0024	0.0012	0.0003	0.0000	0.0003
	Two stage R&C	0.1938	0.0058	0.0033	0.0054	0.0039	0.0000	0.0000	0.0008
	Two stage Spatial	0.1656	0.0075	0.0013	0.0065	0.0011	0.0002	0.0000	0.0004

3.3 Multivariate Analysis

3.3.1 Model building

Aim of multivariate analysis is to jointly model traits accounting for the correlation between them. Extension to the multivariate analysis was done by starting with the simplest possible model since the univariate modeling showed that all the models are good enough in estimating the genetic values. Furthermore, since the one stage model is preferred for efficiency, extensions were done in a single stage approach. Like before modeling is done for the two series to get a

fair assessment of the multivariate approach.

The correlation matrices in the exploratory data analysis show that the correlation values between traits are unique for each pair of traits hence suggest the use of an unstructured matrix of residuals **R**. Correlations for the derived traits %WS, WSY and S_HA and T_HA as well as %S are strong i.e., $r \geq 0.75$ showing that much of the variability in these traits is explained already by the measured traits hence joint analysis was done on the measured traits. On the covariance structures of the random effects **G** we started with the simplest and built upon the model.

Model for series 1111 failed to converge even with the simplest form of **G**. While on the other hand convergence was met for series 1131 multivariate extension of the randomized complete block design RCBD. Results for series 1131 are shown in table 10. There are relatively high correlations between the rankings from all the models. Higher correlations were observed between the genotype rankings from multivariate analysis and the one stage and two stage RCBD models $r = 0.9803$ for T_HA and $r = 0.99$ for all the other traits. The two-stage best selection model is still good enough as there still a high correlation for all the traits the lowest being for rankings on nitrogen with $r = 0.96$

Table 10: Spearman rank comparison of the multivariate RCBD Model to the Univariate RCBD and Best selection model series 1131.

Comparison		**T_HA**	**%S**	**Mm_K**	**Mm_Na**	**Mm_N**
Multivariate Analysis RCBD	**Versus**					
	One stage RCBD	0.9803	0.9916	0.9971	0.9988	0.9957
	Two stage RCBD	0.9803	0.9918	0.9967	0.9985	0.9952
	Two stage Best-selection	0.9710	0.9816	0.9829	0.9922	0.9574

[1] *all correlations are significant at* $\alpha = 0.05$ *level of significance.*

Table 11 shows the MSPDs of the breeding values obtained from the multivariate model and those obtained from the univariate RCBD and two-stage best selection model. The results of the comparison are consistent with the conclusion from the rank correlations in table 10 small MSPD between breeding values from the one stage RCBD model than with the two-stage best-selection model.

Table 11: MSPD for comparison of the multivariate RCBD Model to the Univariate RCBD and Best selection model.

Comparison		T_HA	%S	Mm_K	Mm_Na	Mm_N
Multivariate Analysis RCBD	**Versus**					
	One stage RCBD	0.2250	0.0023	0.0000	0.0000	0.0001
	Two stage RCBD	0.2494	0.0026	0.0002	0.0003	0.0005
	Two stage Best-selection	0.3495	0.0052	0.0007	0.0004	0.0013

3.3.2 Other considerations on multivariate analysis

Due to problems of convergence of the multivariate model when the number of genotypes and field trials get big. We looked for other reduced models that may be important to the genetic selections in sugar beet breeding. Two traits of significant concern are T_HA and %S, hence, a bivariate model of these two traits was done. The RCBD extension of the bivariate analysis was done to allow for consistency in method comparison. Results are shown in Tables 12 and 13. The rank correlations Table 12, show that there are much re-rankings that may be introduced by the joint modeling of these two traits. For series 1111 since the exploratory analysis showed a very small observed correlation $r = -0.081$ no much gain was obtained from the multivariate analysis since the variable T_HA contain very little information about %S. However a reduction in the rank correlation for series 1131 shows that the joint modelling brought some changes to the rankings since for this dataset some of the variability in T_HA is explained by %S $r = -0.6093$.

Table 13 shows that there are very small differences in the predicted breeding values for T_HA and %S as evidenced by small values of the MSPDs.

Table 12: Spearman rank comparison of the bivariate model to the univariate RCBD and the Best selection models.

Comparison		T_HA	%S
Bivariate Analysis RCBD	**Versus:**		
Series 1111	**One stage RCBD**	0.9987	0.9997
	Two stage RCBD	0.9966	0.9992
	Two stage Best-selection	0.9863	0.9942
Series 1131	**One stage RCBD**	0.9842	0.9952
	Two stage RCBD	0.9842	0.9951
	Two stage Best-selection	0.9731	0.9812

Table 13: MSPD comparison of the bivariate models to the univariate RCBD and the Best selection models.

Comparison		T_HA	%S
Bivariate Analysis RCBD	**Versus:**		
Series 1111	**One stage RCBD**	0.0158	0.0001
	Two stage RCBD	1.2301	0.0008
	Two stage Best-selection	1.2401	0.0015
Series 1131	**One stage RCBD**	0.1832	0.0007
	Two stage RCBD	0.2087	0.0011
	Two stage Best-selection	0.3338	0.0049

Comparison of the multivariate and bivariate models

After reducing the analysis from a multivariate analysis with all the measured traits to a bivariate analysis we look at the difference in the genetic predicted values and their respective rankings for the two analyses. The results in Table 14 show that there are small differences in rankings from the multivariate and bivariate analyses $r = 0.99$. The difference is due to some information on T_HA and %S that contained in the impurities, Mm_K, Mm_N, Mm_Na.

Table 14: Spearman correlation and MSPD comparison of the bivariate model to the multivariate RCBD for series 1131.

Comparison		T_HA	%S
Bivariate Analysis RCBD	**Versus:**		
Spearman Correlation	**Multivariate RCBD**	0.9938	0.9941
MSPD	**Multivariate RCBD**	0.0399	0.0016

3.3.3 Genetic and Phenotypic correlations

An unstructured **G** matrix allows the calculations of the genetic and phenotypic correlation between traits shown in Table 15. These correlations are relevant to the breeders to know the indirect impact of selections of one trait to another trait. The genetic correlation between T_HA and %S is moderate negative $r_G = 0.7152$ for series 1111 and $r_G = 0.7906$ for series 1131 this means that an increase in T_HA is combined with a decrease in %S extracted. Furthermore, the phenotypic correlations between tonnes per hectare and percentage sugar are weak negative,

i.e. $r_P = -0.2451$ for series 1111 and $r_P = -0.2113$ for series 1131 this shows that there is a significant relationship between T_HA and %S induced by the combined effect of genetics and the environment.

Table 15: Genetic and phenotypic correlations between tonnes per hectare and percentage sugar

	Series 1111	**Series 1131**
Genetic Correlation	-0.7152	-0.7906
Phenotypic Correlation	-0.2451	-0.2113

4 Discussion and Conclusions

4.1 Discussion

The study aimed at comparing various univariate and multivariate models used in the estimation of breeding values for sugar beet field trials. The estimated breeding values are the predicted means, BLUPs obtained from linear mixed models approach. The comparison was made using the Spearman rank correlation coefficients and the Mean square prediction difference (MSPD). Results form the comparisons showed similarities in genetic rankings and the estimated breeding values obtained from different methods.

Exploratory data analysis, suggested that there is variability in trait measurements for a plot within field trials as well as between field trials. Furthermore, this within-plot variability may be due to different aspects of the experimental design while the between field-trial variability is due to different environmental and management factors. The magnitude of the correlations was different for tonnes per hectare and percentage sugar in the two data-sets a weak negative correlation was observed in series 1111, and a moderate negative correlation in series 1131. This difference in correlation suggested a need to look at the genetic correlation between these two traits as well as the phenotypic correlations to be able to give a meaningful conclusion on the direction and magnitude of the relationship between the two traits using a multivariate approach.

In the first stage of the best-selection two-stage approach four models were fitted for each trial and one best model was chosen using the Akaike's Information criterion to estimate the least square adjusted genetic means, BLUEs used as input in the second stage of the analysis. The spatial model and the row-column models were the most frequently picked as the best models for explaining the within-field variation. Many researchers also reached this conclusion and argued that the models that take into account the two-dimensional variability within the field trials give more accurate and precise estimates of the genotype effect than the complete or the incomplete block analysis (Cullis et al., 1998). In the second stage, the variability in adjusted genetic means was modeled as a function of the genotype, environment and the $G \times E$ Interaction.

Additional variations of the two-stage models were also adopted for the study. A separate two-stage analysis was done for each of the four models to check if the estimated breeding values and

their rankings are robust to model selections in the first stage of the two-stage analysis. Results showed that the predictions from these methods are similar to the best-selection approach. The Spearman rank correlation coefficients showed that the rankings of these models are similar in at least 95% of the cases regardless of the data. The results were found to be consistent in the data-sets used thus the estimates of the breeding values obtained from the analyses are robust to the model selections done in the best-selection two-stage approach.

Though most researchers favor the two-stage methods proposed to reduce the complexity and increase the computational speed during the modeling process. It also have an advantage in time gain since the researchers do not have to wait until all the data from all field trials are available, there has debate throughout literature about the use of the two-stage approach adopted as some scholars argue that there is loss of efficiency (Monneveux et al., 2014). This thesis also adopted a one stage modeling approach to compare the estimated breeding values as well as their respective ranking to the estimates from the two-stage best selection approach. Of the three one stage approaches done the row-column design, as well as the alpha design models, were picked frequently picked as the best models in the one stage models for most traits. Comparison of results from the one stage analyses and the best selection analysis showed that the predictions from this analysis are very close and the similar rankings were observed to each other as evidenced by small values of the MSPD and larger values of the rank correlation coefficients. Many researchers like Piepho and Möhring (2011) also established this conclusion in their study on comparison of two stages and one stage models also concluded that the two-stage approach weighted produce acceptable results when compared to the one stage approach in four data sets.

A multivariate analysis was done on the measured traits only since the exploratory data analysis suggested that almost all the variability in the derived traits explained by tonnes per hectare, and the percentage of extracted sugar. The multivariate analysis allows for the joint modeling of traits hence makes use of the correlations amongst the traits. A multivariate analysis requires correctly specified variance-covariance structures for both the residuals and the random effects. For computational efficiency, extensions to the multivariate models started with the simplest model possible, which is the randomized complete block design to build it up further. An unstructured covariance matrix of residuals as suggested by the correlation matrix and a simple diagonal matrix for variances and covariances of random effects was selected.

However, convergence attained for series 1131 even with complex unstructured covariance structures and convergence problems were encountered for series 1111 even with the simple form of the variance-covariance structure of random effects. Piepho and Möhring (2011) argues that convergence problems are attributed to the data-set under consideration which explains the difference in convergence issues when using different data-sets since series 1111 had 72 genotypes and 33 field trials there are many parameters to be estimated compared to 54 genotypes and 9 environments in series 1131. (Ward et al., 2019) alluded that difficulties in the convergence of multi-trait, multi-environment trials for several traits highlight practical limitations to the technique of using mixed linear models with data collected across environments and traits since as the number of environments increases, so is the number of parameters to be estimated during model fitting.

On the same note, Stringer (1996) argues that many equations need to be solved in the multivariate analysis resulting in increased computational demand as compared to the univariate procedures hence there is need to weigh the values of the gain in accuracy against the cost of increased computational demand. Therefore this thesis took into account the computation time and resources as well as the relative importance of the traits in the selection and decided on a bivariate analysis of the two most essential traits in the selection of sugar beet genotypes. As asserted by Biancardi et al. (2010) gross sugar yield is the most essential trait for sugar beet growers and it depends on tonnes per hectare and percentage hence a bivariate analysis was done on these two traits. Results from the bivariate analysis showed that the estimated breeding values from the joint analysis were not significantly different from the univariate analysis as shown by large values of rank correlations and small mean square prediction difference. Since the multivariate takes in additional information from the other trait in the computation of predictions, the standard errors of predictions are lower than those of the univariate trait models (Isik et al., 2017). Stringer (1996) asserts that if the traits under consideration are correlated a correct specification of the genetic and environmental covariance structure amongst traits also decrease the error variances of the estimated breeding values.

One potential gain obtained from the multivariate analysis allows for a model with unstructured matrix of random effects $\mathbf{G}$ is it allows for the calculation of genetic correlation between any two given traits (Isik et al., 2017). A negative moderate genetic correlation ($r \approx -0.7$) between

tonnes per hectare and percentage of extracted sugar was observed for both series which suggest that a genotype selection that favors an increase in weight of the sugar beet roots results in a proportionate decrease in the amount of extracted sugar. This is consistent with what was suggested in literature by (Campbell (2002),Biancardi et al. (2010)) who also established that there is a high correlation between sugar yield and root yield, and selections that increase root weight tend to lower the sugar content and vice-versa. On the other hand, a smaller phenotypic correlation between these tonnes per hectare and percentage sugar($r \approx -0.2$), relative to the genetic correlation shows that there is a relatively more substantial impact of the environment on the relationship of these two traits. Tenkouano et al. (2002) postulates that if the genetic correlation and phenotypic correlations are very different in magnitude, it suggests there is a genotype $\times$ environment interaction effect on the phenotypic relationship between the two traits. Thus in this view it can be concluded that the inheritance of sugar yield is quantifiable and strongly influenced by the environment (Biancardi et al., 2010). Hence it can be seen that the negative correlations observed for these two traits in the exploratory analysis are also due to genetic,environment and residual variations.

Extensions of the spatial model to the one stage analysis as well as the multivariate analysis became more computationally demanding in SAS hence efforts to extend run such models were fruitless. It is recommended that extensions of such models are tried out in more computationally efficient software packages like Asreml and Genstat that are specifically made to fit spatial models in plant breeding (Piepho and Möhring, 2011).

4.2 Conclusion

This thesis was aimed at a comparative study of various univariate models and extensions to a multivariate approach. The results suggested that there the predicted breeding values from univariate and multivariate approach are similar and produce the same genetic rankings. One advantage that comes with the multivariate models is the ability to estimate the genetic and phenotypic correlations but this comes with a lot of computational cost especially when the number of parameters to be estimated is larger relative to the data available. Hence overally weighing the gain from the multivariate analysis and the computational demand as number of genotypes and environments increase it is recommended to continue with best selection two stage analysis.

References

Balzarini, M. (2002). 23 applications of mixed models in plant breeding. *Quantitative genetics, genomics, and plant breeding*, page 353.

Biancardi, E., McGrath, J. M., Panella, L. W., Lewellen, R. T., and Stevanato, P. (2010). Sugar beet. In *Root and tuber crops*, pages 173–219. Springer.

Burgueño, J., Cadena, A., Crossa, J., Banziger, M., Gilmour, A., and Cullis, B. (2000). *User's guide for spatial analysis of field variety trials using ASREML*. Cimmyt.

Burnham, K. P. and Anderson, D. R. (1998). Practical use of the information-theoretic approach. In *Model Selection and Inference*, pages 75–117. Springer.

Campbell, L. and Fugate, K. (2015). Relationships among impurity components, sucrose, and sugarbeet processing quality. *Journal of Sugar Beet Research*, 52(1/2):2.

Campbell, L. G. (2002). Sugar beet quality improvement. *Journal of crop production*, 5(1-2):395–413.

Chan, Y. (2003). Biostatistics 104: correlational analysis. *Singapore Med J*, 44(12):614–9.

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

Dillen, K. and Demont, M. 41 sugar beet.

Draycott, A. P. et al. (2006). *Sugar beet*, volume 474. Wiley Online Library.

Fischer, S., Möhring, J., Maurer, H., Piepho, H.-P., Thiemt, E.-M., Schön, C., Melchinger, A., and Reif, J. (2009). Impact of genetic divergence on the ratio of variance due to specific vs. general combining ability in winter triticale. *Crop science*, 49(6):2119–2122.

Friesen, L. F., Brûlé-Babel, A. L., Crow, G. H., and Rothenburger, P. A. (2016). Mixed model and stability analysis of spring wheat genotype yield evaluation data from manitoba, canada. *Canadian journal of plant science*, 96(2):305–320.

Guo, G., Zhao, F., Wang, Y., Zhang, Y., Du, L., and Su, G. (2014). Comparison of single-trait and multiple-trait genomic prediction models. *BMC genetics*, 15(1):30.

Gutierrez, L. (2012). *Lecture 9 GxE Mixed models.* Tucson Winter Institute.

Harvey, C. and Dutton, J. (1993). Root quality and processing. In *The sugar beet crop*, pages 571–617. Springer.

Isik, F., Holland, J., and Maltecca, C. (2017). *Genetic data analysis for plant and animal breeding.* Springer.

Kang, M. S. (2002). *Quantitative genetics, genomics, and plant breeding.* CABI.

Kempton, R. A., Fox, P. N., and Cerezo, M. (2012). *Statistical methods for plant variety evaluation.* Springer Science & Business Media.

Kincaid, C. (2005). Guidelines for selecting the covariance structure in mixed model analysis, paper 198-30 in proceedings of the thirtieth annual sas users group conference. *Inc., Cary, North Carolina.*

Komorowski, M., Marshall, D. C., Salciccioli, J. D., and Crutain, Y. (2016). Exploratory data analysis. In *Secondary Analysis of Electronic Health Records*, pages 185–203. Springer.

Littell, R. C., Stroup, W. W., and Freund, R. J. (2002). *SAS for linear models.* SAS institute.

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

Meyer, K. (1991). Estimating variances and covariances for multivariate animal models by restricted maximum likelihood. *Genetics Selection Evolution*, 23(1):67.

Möhring, J. and Piepho, H.-P. (2009). Comparison of weighting in two-stage analysis of plant breeding trials. *Crop Science*, 49(6):1977–1988.

Monneveux, P., Ribaut, J.-M., and Okono, A. (2014). *Drought phenotyping in crops: from theory to practice.* Frontiers E-books.

Montesinos-López, O. A., Montesinos-López, A., Crossa, J., Gianola, D., Hernández-Suárez, C. M., and Martín-Vallejo, J. (2018). Multi-trait, multi-environment deep learning modeling for genomic-enabled prediction of plant traits. *G3: Genes, Genomes, Genetics*, 8(12):3829–3840.

Negash, A. W., Mwambi, H., Zewotir, T., and Aweke, G. (2014). Mixed model with spatial variance covariance structure for accommodating of local stationary trend and its influence on multi-environmental crop variety trial assessment. *Spanish journal of agricultural research*, (1):195–205.

Newcom, D., Baas, T., Stalder, K., and Schwab, C. (2005). Comparison of three models to estimate breeding values for percentage of loin intramuscular fat in duroc swine. *Journal of animal science*, 83(4):750–756.

Patterson, H. and Williams, E. (1976). A new class of resolvable incomplete block designs. *Biometrika*, 63(1):83–92.

Piepho, H., Möhring, J., Melchinger, A., and Büchse, A. (2008). Blup for phenotypic selection in plant breeding and variety testing. *Euphytica*, 161(1-2):209–228.

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

Piepho, H.-P. and Möhring, J. (2011). On estimation of genotypic correlations and their standard errors by multivariate reml using the mixed procedure of the sas system. *Crop Science*, 51(6):2449–2454.

Pigliucci, M. (2006). Genetic variance–covariance matrices: a critique of the evolutionary quantitative genetics research program. *Biology and Philosophy*, 21(1):1–23.

Robinson, G. K. et al. (1991). That blup is a good thing: the estimation of random effects. *Statistical science*, 6(1):15–32.

Rodríguez-Álvarez, M. X., Boer, M. P., van Eeuwijk, F. A., and Eilers, P. H. (2016). Spatial models for field trials. *arXiv preprint arXiv:1607.08255*.

Stringer, J. (1996). Evaluation of methods of estimating breeding value of sugarcane parental clones: Srdc final project report bs75s.

Stroup, W. (1989). Why mixed models in applications of mixed models in agriculture and related disciplines. *South Coop Ser Bull*, 343:1–8.

Tenkouano, A., Ortiz, R., and Baiyeri, K. (2002). Phenotypic and genetic correlations in musa populations in nigeria. *African Crop Science Journal*, 10(2):121–132.

Van Eeuwijk, F. A., Bustos-Korts, D. V., and Malosetti, M. (2016). What should students in plant breeding know about the statistical aspects of genotype×environment interactions? *Crop Science*, 56(5):2119–2140.

Van Eeuwijk, F. A., Malosetti, M., Kruijer, W., and Boer, M. (2011). *Design of experiments and Analysis by mixed models*. UCL.

Volpato, L., Alves, R. S., Teodoro, P. E., de Resende, M. D. V., Nascimento, M., Nascimento, A. C. C., Ludke, W. H., da Silva, F. L., and Borém, A. (2019). Multi-trait multi-environment models in the genetic selection of segregating soybean progeny. *PloS one*, 14(4):e0215315.

Ward, B. P., Brown-Guedira, G., Tyagi, P., Kolb, F. L., Van Sanford, D. A., Sneller, C. H., and Griffey, C. A. (2019). Multienvironment and multitrait genomic selection models in unbalanced early-generation wheat yield trials. *Crop Science*.

Wolfinger, R. D. and Tobias, R. D. (1998). Joint estimation of location, dispersion, and random effects in robust design. *Technometrics*, 40(1):62–71.

Zambrano, A., Rincón, F., López, H., Echeverri, Z., et al. (2015). Estimation and comparison of conventional and genomic breeding values in holstein cattle of antioquia, colombia. *Revista MVZ Córdoba*, 20(3):4739–4753.

5 Appendix

5.1 Series 1111 top ten rakings

The tables show examples of rankings and respective estimated breeding values produced by the different univariate models in series 1111

Table 16: Top 10 Genotype Rankings by different Models selecting for Tonnes per Hactare

Best-Selection Two stage		RCBD One Stage		Alpha One Stage		One Stage Row-Column		Two Stage RCBD		Two Stage Alpha		Two Stage Row-Column	
Rankings	**BLUPS**	**Rankings**	**BLUPS**	**Rankings**	**BLUPS**	**Rankings**	**BLUPS**	**Rankings**	**BLUPS**	**Rankings**	**BLUPS**	**Rankings**	**BLUPS**
44	97.00	44	95.94	44	96.017	44	96.03	44	96.74	44	96.77	44	96.71
46	95.58	49	94.86	49	94.911	49	94.95	49	95.89	49	95.96	46	95.82
49	95.53	46	94.38	46	94.439	46	94.44	46	95.51	46	95.38	49	95.58
48	95.33	65	94.09	65	94.141	65	94.17	65	95.13	65	95.26	48	94.94
65	95.32	52	94.02	52	94.077	52	94.08	52	95.08	48	95.20	61	94.92
52	95.05	61	93.89	61	93.887	61	93.91	61	94.87	52	95.09	52	94.90
61	94.94	48	93.73	48	93.766	48	93.78	48	94.70	61	95.09	65	94.89
62	94.08	62	93.58	62	93.545	62	93.61	62	94.33	11	94.23	59	94.36
59	93.98	53	93.17	53	93.209	53	93.22	53	94.19	62	94.17	11	94.28
12	93.93	11	93.09	11	93.124	11	93.14	11	94.16	12	94.03	62	93.84

Table 17: Top 10 Genotype Rankings by different Models selecting for Sugar per hectare

Best-Selection Two stage		RCBD One Stage		Alpha One Stage		One Stage Row-Column		Two Stage RCBD		Two Stage Alpha		Two Stage Row-Column	
Rankings	**BLUPS**	**Rankings**	**BLUPS**	**Rankings**	**BLUPS**	**Rankings**	**BLUPS**	**Rankings**	**BLUPS**	**Rankings**	**BLUPS**	**Rankings**	**BLUPS**
12	16.7229	62	16.6038	62	16.6009	62	16.6167	62	16.7487	62	16.7287	14	16.7096
14	16.7036	65	16.5079	65	16.5279	65	16.5374	65	16.695	65	16.726	12	16.6774
65	16.693	61	16.4805	63	16.4976	63	16.493	61	16.6572	61	16.7015	65	16.6737
62	16.6846	63	16.4732	61	16.4786	61	16.4871	12	16.6545	12	16.7002	62	16.6698
61	16.657	6	16.4552	6	16.472	6	16.4693	6	16.6515	14	16.697	63	16.6516
52	16.6553	12	16.4535	52	16.466	52	16.4654	14	16.6512	2	16.6691	11	16.6493
6	16.6453	52	16.4524	11	16.4612	11	16.4638	52	16.6494	52	16.6455	6	16.6455
2	16.6042	11	16.4505	12	16.453	12	16.4533	11	16.6478	11	16.6408	52	16.6286
63	16.5871	14	16.4284	14	16.4508	14	16.4318	63	16.6466	6	16.612	61	16.6251
11	16.5531	2	16.4191	53	16.4277	53	16.4308	2	16.6194	53	16.587	2	16.581

Table 18: Top 10 Genotype Rankings by different Models selecting for Percentage White Sugar

Best-Selection Two stage		RCBD One Stage		Alpha One Stage		One Stage Row-Column		Two Stage RCBD		Two Stage Alpha		Two Stage Row-Column	
Rankings	**BLUPS**	**Rankings**	**BLUPS**	**Rankings**	**BLUPS**	**Rankings**	**BLUPS**	**Rankings**	**BLUPS**	**Rankings**	**BLUPS**	**Rankings**	**BLUPS**
5	17.4708	5	17.4187	5	17.4218	5	17.4217	1	17.4349	5	17.4597	5	17.4477
1	17.4262	1	17.4069	1	17.4116	1	17.4115	5	17.4163	1	17.4347	1	17.4303
22	17.3258	22	17.3492	22	17.3522	22	17.3525	22	17.3685	22	17.3577	22	17.3288
31	17.163	19	17.1048	19	17.113	19	17.1117	31	17.1318	31	17.1849	31	17.1469
70	17.1258	31	17.0996	31	17.1027	31	17.1028	70	17.1233	70	17.1449	70	17.143
19	17.1177	70	17.0756	70	17.0807	70	17.0805	19	17.122	19	17.107	19	17.1298
25	17.0788	25	17.0209	25	17.0155	25	17.022	25	17.0566	50	17.1069	50	17.1025
50	17.0778	50	17.0057	50	17.008	50	17.0081	50	17.0258	25	17.0463	25	17.0758
43	16.97	43	16.9577	43	16.9602	43	16.9604	43	16.9946	43	17.0051	43	16.9704
27	16.9147	71	16.9187	71	16.9212	71	16.9211	71	16.9482	71	16.9344	71	16.9679

Table 19: Top 10 Genotype Rankings by different Models selecting for White Sugar Yield

Best-Selection Two stage		RCBD One Stage		Alpha One Stage		One Stage Row-Column		Two Stage RCBD		Two Stage Alpha		Two Stage Row-Column	
Rankings	**BLUPS**	**Rankings**	**BLUPS**	**Rankings**	**BLUPS**	**Rankings**	**BLUPS**	**Rankings**	**BLUPS**	**Rankings**	**BLUPS**	**Rankings**	**BLUPS**
14	15.4405	62	15.3269	62	15.3243	62	15.3388	62	15.468	62	15.4519	14	15.442
65	15.4134	65	15.238	65	15.2568	65	15.2664	65	15.4168	65	15.4489	62	15.4007
62	15.4081	63	15.2191	63	15.2417	63	15.2377	6	15.3973	14	15.4388	63	15.3984
6	15.3791	6	15.21	6	15.2267	6	15.2245	63	15.3853	61	15.4199	65	15.3924
61	15.377	61	15.209	61	15.2062	61	15.2148	14	15.3839	2	15.4028	6	15.3895
12	15.374	14	15.1725	14	15.1946	14	15.1766	61	15.3764	6	15.3566	61	15.3459
2	15.3416	52	15.164	52	15.1765	52	15.176	52	15.3528	12	15.349	52	15.334
21	15.3332	2	15.15	21	15.155	21	15.1542	2	15.3429	52	15.3476	12	15.3239
52	15.3331	21	15.1351	2	15.1531	2	15.1431	11	15.3073	63	15.3258	21	15.3218
63	15.3304	53	15.1217	53	15.1298	53	15.1328	12	15.3069	21	15.3073	2	15.316

Table 20: Top 10 Genotype Rankings by different Models selecting for percentage Sugar

Best-Selection Two stage		RCBD One Stage		Alpha One Stage		One Stage Row-Column		Two Stage RCBD		Two Stage Alpha		Two Stage Row-Column	
Rankings	**BLUPS**	**Rankings**	**BLUPS**	**Rankings**	**BLUPS**	**Rankings**	**BLUPS**	**Rankings**	**BLUPS**	**Rankings**	**BLUPS**	**Rankings**	**BLUPS**
5	18.7546	5	18.7039	5	18.7079	5	18.7079	1	18.7222	5	18.7471	5	18.7303
1	18.7043	1	18.7011	1	18.7053	1	18.7052	5	18.7115	1	18.7207	1	18.7248
22	18.674	22	18.6884	22	18.6925	22	18.6926	22	18.7085	22	18.7026	22	18.681
70	18.5646	70	18.5164	70	18.5199	70	18.5196	70	18.5589	70	18.5764	70	18.5829
31	18.4883	31	18.4603	31	18.4633	31	18.4634	31	18.4815	31	18.5086	31	18.4767
19	18.4116	19	18.4189	19	18.4264	19	18.4255	19	18.433	50	18.4248	19	18.4436
50	18.3935	50	18.3381	43	18.3405	43	18.3408	43	18.3671	19	18.4159	50	18.4287
25	18.3718	43	18.338	50	18.3402	50	18.3403	25	18.3536	43	18.3796	25	18.3842
43	18.3598	25	18.3259	25	18.3221	25	18.3273	50	18.3504	25	18.3521	71	18.3483
71	18.3044	71	18.3166	71	18.3191	71	18.3191	71	18.3412	71	18.3169	43	18.3472

Table 21: Top 10 Genotype Rankings by different Models selecting for Nitrogen

Best-Selection Two stage		RCBD One Stage		Alpha One Stage		One Stage Row-Column		Two Stage RCBD		Two Stage Alpha		Two Stage Row-Column	
Rankings	**BLUPS**	**Rankings**	**BLUPS**	**Rankings**	**BLUPS**	**Rankings**	**BLUPS**	**Rankings**	**BLUPS**	**Rankings**	**BLUPS**	**Rankings**	**BLUPS**
63	1.03536	65	1.07123	65	1.06435	65	1.06603	65	1.05113	65	1.03555	63	1.05306
65	1.0503	9	1.08778	9	1.08204	9	1.08211	63	1.06744	4	1.0657	65	1.05641
61	1.068	63	1.09266	4	1.08834	63	1.08819	4	1.07215	61	1.06728	4	1.06513
4	1.07072	4	1.09342	63	1.08847	4	1.08844	61	1.07553	8	1.06861	61	1.07727
19	1.07521	61	1.1012	8	1.09669	8	1.0963	9	1.07649	63	1.07037	62	1.08235
8	1.0753	8	1.10146	61	1.09928	61	1.09768	8	1.07904	62	1.07248	9	1.08695
62	1.08746	62	1.12024	62	1.11675	62	1.11682	62	1.09604	9	1.08216	19	1.09001
9	1.08994	19	1.13265	19	1.12924	19	1.13008	25	1.10952	19	1.10317	8	1.09055
64	1.12821	25	1.1408	25	1.14171	25	1.13892	19	1.11611	64	1.12799	3	1.12898
25	1.13058	33	1.15506	33	1.15215	33	1.15156	33	1.12377	25	1.12915	67	1.13225

Table 22: Top 10 Genotype Rankings by different Models selecting for Sodium

Best-Selection Two stage		**RCBD One Stage**		**Alpha One Stage**		**One Stage Row-Column**		**Two Stage RCBD**		**Two Stage Alpha**		**Two Stage Row-Column**	
Rankings	**BLUPS**	**Rankings**	**BLUPS**	**Rankings**	**BLUPS**	**Rankings**	**BLUPS**	**Rankings**	**BLUPS**	**Rankings**	**BLUPS**	**Rankings**	**BLUPS**
50	0.27098	50	0.28257	50	0.28129	50	0.2813	50	0.2812	50	0.27277	50	0.26531
31	0.29456	31	0.3025	31	0.30151	31	0.30144	31	0.2983	31	0.29055	31	0.29278
30	0.30513	30	0.30985	30	0.30875	30	0.30871	30	0.3046	30	0.302	30	0.2999
60	0.32914	16	0.32507	16	0.32401	16	0.32437	16	0.31992	16	0.32253	60	0.32745
51	0.32955	60	0.33051	60	0.33013	60	0.32957	60	0.32569	51	0.323	51	0.32906
16	0.32991	51	0.33298	51	0.33224	51	0.33213	51	0.32766	17	0.32386	16	0.33066
17	0.3314	17	0.33582	17	0.33508	17	0.33495	17	0.32938	60	0.32592	17	0.33209
33	0.33405	39	0.34064	39	0.3398	39	0.34003	39	0.33348	39	0.33839	18	0.33568
18	0.33636	33	0.34982	33	0.34958	33	0.35048	33	0.33951	33	0.34127	39	0.33675
32	0.33753	18	0.35222	18	0.35149	18	0.3516	18	0.34323	32	0.3415	33	0.33692

Table 23: Top 10 Genotype Rankings by different Models selecting for Pottasium

Best-Selection Two stage		**RCBD One Stage**		**Alpha One Stage**		**One Stage Row-Column**		**Two Stage RCBD**		**Two Stage Alpha**		**Two Stage Row-Column**	
Rankings	**BLUPS**	**Rankings**	**BLUPS**	**Rankings**	**BLUPS**	**Rankings**	**BLUPS**	**Rankings**	**BLUPS**	**Rankings**	**BLUPS**	**Rankings**	**BLUPS**
0	3.23045	60	3.21557	60	3.20943	60	3.20984	60	3.22808	60	3.21754	60	3.22439
5	3.27739	1	3.27086	1	3.26606	1	3.26678	1	3.28838	5	3.27225	5	3.27839
1	3.28562	5	3.27784	5	3.27331	5	3.27397	5	3.28886	1	3.28334	1	3.29622
23	3.33149	9	3.30912	9	3.30488	9	3.30548	9	3.32987	23	3.33367	9	3.33538
9	3.33179	23	3.32837	23	3.32467	23	3.32518	23	3.34808	9	3.33484	23	3.33794
58	3.37365	58	3.33126	58	3.32715	58	3.32769	58	3.35458	58	3.37151	58	3.37043
50	3.37948	38	3.35764	38	3.35422	38	3.3546	38	3.3812	50	3.38078	40	3.376
38	3.38616	40	3.36518	40	3.36164	40	3.36236	40	3.38259	40	3.39347	38	3.38287
40	3.39784	25	3.37918	25	3.37523	25	3.37408	50	3.3855	38	3.39433	50	3.38597
14	3.42061	50	3.37954	50	3.37643	50	3.37679	25	3.38839	14	3.41204	25	3.41548

5.2 SAS CODES

```
PROC IMPORT DATAFILE="E:\Thesis_files\datas\wideformat.csv"
  OUT=datawide
  DBMS=csv
  REPLACE;
Delimiter=',';
  GETNAMES=Yes;
RUN;
libname proj 'E:\Thesis_files\series 1131';
/*******SUBSETTING THE DATA FOR ONE SERIES*********/
/* SERIES 1111 2017*/
data proj.series1111;
set datawide;
where series_id=1111 & Year=2017;
run;
/* UNIVARIATE ANALYSIS CODE EXAMPLE FOR ONE TRAIT*/
/**** BEST SELECTION APPROACH*****/
/* RUN MODEL 1 RCBD and Extract AICs*/
/* MODEL 1 */
ods graphics off;
ods exclude all;
PROC MIXED data=proj.out;
BY field_trial_id;
class repetition_id object_id ;
model sugar=object_id repetition_id;
ods output FitStatistics=proj.aic1_sugar (RENAME=(Value=Value1)); *Output fit statistics (Al
quit;
ods exclude none;
/* MODEL 2 INCOMPLETE BLOCK DESIGN */
ods graphics off;
ods exclude all;
```

```
PROC MIXED data=proj.out;
BY field_trial_id;
class repetition_id object_id Block_id;
model sugar=object_id repetition_id;
random  block_id(repetition_id);
ods output FitStatistics=proj.aic2_sugar (RENAME=(Value=Value2)); *Output fit statistics (Al
quit;
ods exclude none;
/* MODEL ROW COLUMN DESIGN */
ods graphics off;
ods exclude all;
PROC MIXED data=proj.out;
BY field_trial_id;
class X_coordinate_field Y_coordinate_field object_id ;
model sugar=object_id;
random X_coordinate_field;
random Y_coordinate_field;
ods output FitStatistics=proj.aic3_sugar (RENAME=(Value=Value3)); *Output fit statistics (Al
quit;
ods exclude none;
/* MODEL 4  SPATIAL MODEL*/
ods graphics off;
ods exclude all;
PROC MIXED data=proj.out;
BY field_trial_id;
class  X_coordinate_field Y_coordinate_field object_id;
model sugar=object_id;
repeated X_coordinate_field*Y_coordinate_field  /subject=int type=sp(powa)(X_coordinate_fiel
ods output FitStatistics=proj.aic4_sugar (RENAME=(Value=Value4)); *Output fit statistics (Al
quit;
ods exclude none;
 /****** EXTRACTING THE AICS FOR SELECTING BEST MODEL FOR EACH FIELD TRIAL*****/
```

```
data proj.aic_sugar;
MERGE proj.aic1_sugar  proj.aic2_sugar  proj.aic3_sugar proj.aic4_sugar;
BY field_trial_id;
WHERE Descr = 'AIC (Smaller is Better)'; *Select only AIC values;
array values Value1-Value4;
index = whichn(min(of values[*]), of values[*]); *Get model number with lowest AIC value;
run;
DATA proj.out2_sugar (DROP = Descr Value1-Value4);
MERGE proj.out proj.aic_sugar; *Merge orig data with the model number with lowest AIC value
BY field_trial_id;
RUN;
/********* FITTING THE BEST MODELS FOR EACH FIELD TRIAL*******/
/******** AND EXTRACT THE ADJUSTED MEANS THE BLUES*********/
/* MODEL1, adj means */
ods graphics off;
ods exclude all;
Proc Mixed data=proj.out2_sugar;
BY field_trial_id;
class repetition_id object_id ;
model sugar=object_id repetition_id /solution;
lsmeans object_id;
WHERE index=1;
ods output LSMeans=proj.adjmean1_sugar;
quit;
ods exclude none;
/* MODEL2, adj means */
ods graphics off;
ods exclude all;
Proc Mixed data=proj.out2_sugar;
BY field_trial_id;
class repetition_id object_id Block_id;
model sugar=object_id repetition_id /solution;
```

```
lsmeans object_id;
random  block_id(repetition_id);
WHERE index=2;
ods output LSMeans=proj.adjmean2_sugar;
quit;
ods exclude none;
/* MODEL3, adj means */
ods graphics off;
ods exclude all;
Proc Mixed data=proj.out2_sugar;
BY field_trial_id;
class X_coordinate_field(ref=first) Y_coordinate_field(ref=first) object_id ;
model sugar=object_id /solution;
lsmeans object_id;
random X_coordinate_field;
random Y_coordinate_field;
WHERE index=3;
ods output LSMeans=proj.adjmean3_sugar;
quit;
ods exclude none;
/* MODEL4, adj means */
ods graphics off;
ods exclude all;
Proc Mixed data=proj.out2_sugar;
BY field_trial_id;
class  X_coordinate_field Y_coordinate_field object_id;
model sugar=object_id;
lsmeans object_id;
repeated X_coordinate_field*Y_coordinate_field  /subject=int type=sp(powa)(X_coordinate_fiel
WHERE index=4;
ods output LSMeans=proj.adjmean4_sugar;
quit;
```

```
ods exclude none;
/*****************  CReating a dataset for the outputs(Adjusted Means From the first stage c
DATA proj.results_sugar1 (KEEP = FIELD_TRIAL_ID OBJECT_ID Estimate StdErr); *Merge adjusted
SET  proj.adjmean2_sugar proj.adjmean3_sugar proj.adjmean4_sugar;
BY field_trial_id;Run;
/**** Check which model was frequently picked****/
proc freq data=proj.aic_sugar;tables index/nopercent;
run;

/********************************** SECOND STAGE  Calculating  Estimated breeding vales ***
/******** the Input are the adjusted mean from each field trial**********/
proc mixed data=proj.results_sugar1;
 class field_trial_id object_id;
 model estimate=field_trial_id/outp=proj.pred_sugar;
 random intercept/ subject=object_id; run;
 proc print data=proj.pred_sugar; run;
  proc means data=proj.pred_sugar MEAN ;
 var pred;
 class object_id;
 output out=proj.EBVs_sugar ; run;
PROC SORT DATA=proj.EBVS_sugar;
by OBJECT_ID;RUN;
PROC TRANSPOSE DATA =proj.EBVS_sugar OUT=proj.EBVS1_sugar;
BY  OBJECT_ID  ;
ID  _STAT_;
VAR PRED;RUN;
proc print data=proj.EBVS1_sugar;run;
DATA proj.EBVS2_sugar(Keep=object_id Mean std);
SET proj.EBVS1_sugar;
WHERE OBJECT_ID NE .; RUN;
 /*****Note that this code can be manipulated for other types of univariate models  by using
/*****One Stage model EXAMPLES are given for only one trait and the code can be customised f
```

```
/**** Model 1 ********************** BALANCED COMPLETE BLOCK DESIGN**********************
 /****************** T_HA*************************************************************
proc mixed data=proj.series1111;
   class field_trial_id object_id repetition_id ;
   model t_ha=field_trial_id repetition_id(Field_trial_id)/outp=proj.one_t_ha ;
  random object_id;
   random object_id*field_trial_id;
  run;
 proc means data=proj.one_t_ha MEAN ;
 var pred;
 class object_id;
 output out=proj.EBVsone_t_ha ;
 run;
PROC SORT DATA=proj.EBVSone_t_ha;
by OBJECT_ID;
RUN;
PROC TRANSPOSE DATA =proj.EBVSone_t_ha OUT=proj.EBVS1one_t_ha;
BY  OBJECT_ID  ;
ID  _STAT_;
VAR PRED;RUN;
PROC PRINT DATA=proj.EBVS1one_t_ha;
RUN;
DATA proj.EBVS2one_t_ha (Keep=object_id mean std);
SET proj.EBVS1one_t_ha ;
WHERE OBJECT_ID NE .;
run;

/********************************** T_HA***** MODEL 2*** INCOMPLETE BLOCK DESIGN*********
proc mixed data=proj.series1111;
   class field_trial_id object_id repetition_id block_id ;
  model t_ha=field_trial_id repetition_id(Field_trial_id)/outp=proj.two_t_ha  ;
```

```
 random object_id;
random object_id*field_trial_id ;
random block_id(repetition_id*Field_trial_id);   run;
proc means data=proj.two_t_ha MEAN ;
 var pred;
 class object_id;
 output out=proj.EBVstwo_t_ha ;
 run;
PROC SORT DATA=proj.EBVStwo_t_ha;
by OBJECT_ID;RUN;
PROC TRANSPOSE DATA =proj.EBVStwo_t_ha OUT=proj.EBVS1two_t_ha;
BY  OBJECT_ID  ;
ID  _STAT_;
VAR PRED;RUN;
DATA proj.EBVS2two_t_ha (Keep=object_id mean std);
SET proj.EBVS1two_t_ha ;
WHERE OBJECT_ID NE .;run;
/**** Model 3 *** ROW COLUMN DESIGN*/
/********************************** T_HA*************************************************
  proc mixed data=proj.series1111;
  class field_trial_id object_id repetition_id block_id x_coordinate_field y_coordinate_fiel
  model t_ha=field_trial_id/outp=proj.three_t_ha;
  random object_id;
  random object_id*Field_trial_id;
 random  x_coordinate_field(field_trial_id);
 random y_coordinate_field(field_trial_id);  run;
   proc means data=proj.three_t_ha MEAN ;
 var pred;
 class object_id;
 output out=proj.EBVsthree_t_ha ; run;
PROC SORT DATA=proj.EBVSthree_t_ha;
by OBJECT_ID;RUN;
```

```
PROC TRANSPOSE DATA =proj.EBVSthree_t_ha OUT=proj.EBVS1three_t_ha;
BY  OBJECT_ID  ;
ID  _STAT_;
VAR PRED;RUN;
DATA proj.EBVS2three_t_ha (keep=object_id Mean std);
SET proj.EBVS1three_t_ha;
WHERE OBJECT_ID NE .;run;
/******* These codes can be customized to different types of univariate analysis that were c
/ * MULTIVARIATE DATA ANALYSIS ***** USED DATA IN TALL FORMAT*******/
PROC IMPORT DATAFILE="f:\project\tall_fomart.csv"
  OUT=datawide
  DBMS=csv
  REPLACE;
Delimiter=',';
  GETNAMES=Yes;RUN;
libname new1 'f:\Project\Output_META-R';
data new1.multivariate;
set datawide;run;
/****** SUBSETTING DATA *******/ SERIES 1131 2018*******/
data new1.series1131;
set new1.multivariate;
where series_id=1131 & Year=2018;
run;
 data new1.series1131_twotraits;
 set new1.series1131;
 where characteristic_id in (5,11,17,19,21); run;
/*** MODEL THE MEASURED TRAITS*****/
/* Three Traits Impurities*/
/*** ALL SERIES 1131 *********/
proc mixed data=new.series1111_twotraits;
class object_id field_trial_id block_id repetition_id plot_id characteristic_id;
model absolute= characteristic_id field_trial_id*characteristic_id repetition_id*Field_trial
```

```
repeated characteristic_id /type=un subject=plot_id*object_id;
random characteristic_id/subject=object_id type=un;
random characteristic_id / subject= object_id*field_trial_id type=un;
run;
quit;
/******* THE BREEDING VALUES*********/
proc sort data=new1.kirie;
by characteristic_id;
run;
 proc means data=new1.kirie  ;
 var pred;
 class object_id;
 by characteristic_id;
 output out=new1.kirie1 ; run;
/***This code can be extended o any multivariate analysis my different modifications.
Different datasets were obtaine as output from different modifiactions of the code for diffe
the spearman rank correlations where calculateted for ranks of the genotypes for different r
 /************** Correlation Analysis  Of EBLUPS spearman*******************/
/*************** TONES PER HACTARE*********************/
proc corr data=proj.combined_BLUPS spearman;
var T_HA1 T_HA2 T_HA3 T_HA4 T_HA5 T_HA6 T_HA7 T_HA8;
run;
/*************** SUGAR PER HACTARE*********************/
proc corr data=proj.combined_BLUPS spearman;
var S_HA1 S_HA2 S_HA3 S_HA4 S_HA5 S_HA6 S_HA7 S_HA8;
run;
/*************** SUGAR*********************/
proc corr data=proj.combined_BLUPS spearman;
var SUGAR1 SUGAR2 SUGAR3 SUGAR4 SUGAR5 SUGAR7 SUGAR6 SUGAR8;
run;
/*************** WS *********************/
proc corr data=proj.combined_BLUPS spearman;
```

```
var WS1 WS2 WS3 WS4 WS5  WS6  WS7 WS8;
run;
/*************** WSY ********************/
proc corr data=proj.combined_BLUPS spearman;
var WSY1 WSY2 WSY3 WSY4 WSY5 WSY6 WSY7 WSY8;
run;
/*************** mM_K********************/
proc corr data=proj.combined_BLUPS spearman;
var mM_K1 mM_K2 mM_K3 mM_K4 mM_K5 mM_K6 mM_K7 mM_k8;
run;
/*************** mM_Na ********************/
proc corr data=proj.combined_BLUPS spearman;
var mM_Na1 mM_Na2 mM_Na3 mM_Na4 mM_Na5 mM_Na6 mM_Na7 Mm_Na8;
run;
/*************** mM_N ********************/
proc corr data=proj.combined_BLUPS spearman;
var mM_N1 mM_N2 mM_N3 mM_N4 mM_N5 mM_N6 mM_N7 mM_N8;
run;
####### Mean square Error was calculated in R
####### WE customize the code to get all MSE values needed
EBVS=read.csv(file.choose(),sep=',')
head(EBVS)
library(MLmetrics)

MSE(EBVS$WSY1,EBVS$WSY2)
MSE(EBVS$WSY1,EBVS$WSY3)
MSE(EBVS$WSY1,EBVS$WSY4)
MSE(EBVS$WSY1,EBVS$WSY5)
MSE(EBVS$WSY1,EBVS$WSY6)
MSE(EBVS$WSY1,EBVS$WSY7)
MSE(EBVS$WSY1,EBVS$WSY8)
```