**University of São Paulo
“Luiz de Queiroz” College of Agriculture**

**The statistical paradigm: probabilistic and multivariate analysis applied
through computational simulation in the interaction between genotype
x environment**

Danilo Augusto Sarti

Thesis presented to obtain the degree of Doctor in Science. Area: Statistics and Agricultural Experimentation

**Piracicaba
2019**

Danilo Augusto Sarti
Agronomyst

**The statistical paradigm: probabilistic and multivariate analysis applied
through computational simulation in the interaction between genotype
x environment**

versão revisada de acordo com a resolução CoPGr 6018 de 2011

Advisor:

Prof. Dr. **CARLOS TADEU DOS SANTOS DIAS**

Thesis presented to obtain the degree of Doctor in Science. Area: Statistics and Agricultural Experimentation

Piracicaba
2019

**Dados Internacionais de Catalogação na Publicação
DIVISÃO DE BIBLIOTECA - DIBD/ESALQ/USP**

Sarti, Danilo Augusto

The statistical paradigm: probabilistic and multivariate analysis applied through computational simulation in the interaction between genotype x environment / Danilo Augusto Sarti. -- versão revisada de acordo com a resolução CoPGr 6018 de 2011. -- Piracicaba, 2019 .

73 p.

Tese (Doutorado) -- USP / Escola Superior de Agricultura "Luiz de Queiroz".

1. Ajuste probabilístico 2. Análise multivariada 3. Simulação computacional 4. Interação genotipoambiental 5. Raspagem de dados 6. Software R I. Título.

DEDICATION

To my parents and wife for their unbiased support.

ACKNOWLEDGEMENTS

To my parents, Carmen and Carlos, for all their loving support.

To my brother Jr and sister Ana (*in memoriam*). To all my family, specially my dear nephews, Maria e Otávio.

To my grandma, Paula Bergamin (*in memoriam*), for all the valuable lessons she taught me.

To Mayra, my beloved wife, who helped me not give up and persevered with me in hard moments.

To my friends from all walks of life.

To my advisor Carlos Tadeu dos Santos Dias for the great opportunity of doing this work.

To the researcher and friend Davi Inada for all the support with regular expressions programming without whom this work would not be possible.

To the CAPES for material support.

To my dear friend Nancy Edit Bertone for proof-reading and English corrections.

To professors Sônia Piedade, Lúcio Borges, Roberto Frietsche, Silvio Govone, Caio Azevedo, Juliana Petrini, Idemauro de Lara for the opportunity to evaluate this work. To professor César Gonçalves de Lima for the support in the qualification exam.

To all LCE/ESALQ/USP professors and to professor Clarisse Borges Demétrio by the courses on experimental statistics and mixed models

To all LCE/ESALQ/USP workers and colleagues.

To all LES/ESALQ/USP teachers and colleagues for the courses taken on applied economics.

To all LGN/ESALQ/USP teachers and colleagues for the courses taken on genetics and plant breeding

To João Rosolem, maize breeder who took part in the development of the specialty corn chain in Brazil, for valuable observations and insights without which the work wouldn't be possible.

To the Institute of Mathematics Statistics and Scientific Computing IMECC-Unicamp for the courses in undergraduate and graduate levels.

SUMMARY

Resumo	7
Abstract	8
1 Introduction	9
References	10
2 Multi environment maize trials	11
2.1 Introduction	11
2.2 Methodologies	11
2.2.1 Embrapa national maize trials	11
2.2.2 Data and methodology of scrapping of Embrapa MET datasets	12
2.3 Discussion	14
2.3.1 Results	14
2.4 Conclusions	15
References	15
3 Fitting Statistical Distribution to Data	17
3.1 Introduction	17
3.2 Methodology	17
3.2.1 Fitting of univariate distributions to data	17
3.2.2 Fitting mixture of distribution to data	21
3.3 Results	23
3.3.1 Probabilistic models for gw1, gwc and gws data sets	23
3.3.2 Probabilistic models for gwn and data set	30
3.4 Conclusions	31
References	35
4 Simulation of genotype x environment interaction data in a block experimental design	37
4.1 Introduction	37
4.2 Methodology	37
4.2.1 Table to generate the genotypes x environments interaction	37
4.2.2 Results	39
4.3 Conclusions	41
References	42
5 Overview on Gene-Environment Interaction	43
5.1 Introduction	43
5.2 The biological universality of genotype and environment interactions	43
5.3 Classification of genotype x environment interaction	43
5.4 Historical perspectives on gene x environment modelling via AMMI models	45
5.5 Conclusion	46
References	47
6 AMMI, weighted AMMI and GGE Biplot Methods	49
6.1 Introduction	49
6.2 Methodology	49
6.2.1 Analysis of Variance	49
6.2.2 AMMI Model	50
6.2.3 Weighted AMMI model	53
6.2.4 Genotype main effects + genotype environment interaction - GGE methods	53

6.2.5	Information criteria	57
6.3	Results	57
6.3.1	AMMI Analysis of simulated dataset	57
6.3.2	Comparison between AMMI, w-AMMI and GGE biplots	66
6.4	Conclusions	67
	References	70
	Appendices	71
	Code	73

RESUMO

O paradigma estatístico: análise probabilística e multivariada aplicadas via simulação computacional no contexto da interação genótipo ambiente

A estatística fundamenta-se em um paradigma elementar, baseado na relação entre a inferência indutiva probabilística, geração e validação de modelos e o uso de tais informações como subsídio em decisões em um domínio específico de conhecimento. Aliado a isso, técnicas podem ser utilizadas para delinear tipos específicos de experimentos, como os ensaios multi ambientais MET para estudos de interação entre genótipos e ambientes. O ajuste de distribuição de probabilidades a dados provenientes de fenômenos permite o conhecimento do comportamento de variáveis aleatórias e posterior uso de tais modelos em simulação computacional. Tal procedimento foi realizado no ajuste de modelos para peso de grãos de genótipos de milho em ensaios multi ambientais, através de diversos métodos de ajuste de distribuição e misturas de distribuições normais pelo algoritmo EM. Os dados foram obtidos através do uso de scrapping via software R. Por sua vez, os modelos ajustados foram utilizados para simular, através de métodos computacionais implementados em linguagem R, dados com comportamento conhecido em termos paramétricos, através de uma tabela que simula a interação entre os fatores genótipo e ambiente. Tais dados simulados foram utilizados para verificar, e comparar os modelos baseados em análise multivariada de dados, a saber AMMI, AMMI ponderado e GGE, para o estudo da interação genótipo ambiente (GxE). Os resultados demonstraram a grande efetividade dos modelos em captar as propriedades dos dados simulados, contextualizando-os como ferramentas informacionais no desenvolvimento de novos produtos.

Palavras-chave: Ajuste probabilístico; Análise multivariada; Simulação computacional; Interação genotipoambiental; Raspagem de dados; Software R

ABSTRACT

The statistical paradigm: probabilistic and multivariate analysis applied through computational simulation in the interaction between genotype x environment

Statistical analysis is based on an elementary paradigm and the relationship between probabilistic inductive inference, generation and validation of models, and the use of such information in decisions within a specific domain of knowledge. Additionally, techniques can be used to design specific experiments, such as the multi-environmental trials MET, to study the interaction between genotypes and environments. The fitting of probability distributions to data from phenomena allows the knowledge of the behavior of random variables and the later usage of such models in computational simulation. This procedure was carried out in the adjustment of models for maize grains weight, obtained via multi environmental trials. Several methods of adjustment of distribution and mixtures of normal distributions by the EM algorithm were used. The data were obtained through the use of scrapping with software R. Adjusted models were used to simulate, through computational methods implemented in language R, data with behavior known in parametric terms, generating a table that simulates the interaction between genotype and environment factors. Such simulated data were used to verify and compare models based on multivariate analysis, namely AMMI, weighted AMMI and GGE for the study of genotype environment interaction GxE. The results demonstrated the great effectiveness of the models in capturing the properties of the simulated data, contextualizing them as informational tools in the development of new products.

Keywords: Probabilistic fitting; Multivariate analysis; Computational simulation; GxE interaction; scrapping; R programming

1 INTRODUCTION

Statistics is based on a paradigm whose foundations are inferences based on probability theory, the production and validation of models and the use of information provided by them in decisions from a specific domain of knowledge AZEVEDO (2010b). A comprehensive historical perspective of the evolution of statistical paradigm throughout time is presented in STIGLER (1986) and in HALD (2007).

The assessment of uncertainty related to a random variable can be made by the fitting of a probability distribution. Such procedure is easily carried out with the use of computational techniques. Examples of the implementation of such methods are shown in DELIGNETTE-MULLER and DUTANG (2015) and KARIAN and DUDEWICZ (2011).

Simulation is an important technique implemented nowadays in software. Most part of the techniques are based on Markov Chain Monte Carlo methods.

The use of computational simulation has been proved an essential tool for science and its use is disseminated in several fields.

On the other hand, one key question for humanity is food production by agriculture techniques. Plant breeding is crucial for agriculture and the development of new products, being one of the first steps for chains of food production ALLARD (1999) and SARTI (2013).

An important problem in plant breeding is genotype between environment interactions. Such phenomenon has profound impact on the decisions about the development of new products, regulation and recommendation, delineating the process of allocation of products in the market.

To verify the behaviour of different genotypes over a set of environments a specific kind of experiment can be designed and implemented. Such experiments are named multi environmental trials MET and are described in the following chapters of this thesis.

Data acquired by METs can be used to feed models that claim to determine the patterns of genotype x environments interaction such as AMMI, weighted AMMI and GGE HONGYU *et al.* (2014).

The title was given due to the use of probabilistic techniques, multivariate analysis and simulation. The term statistical paradigm was by the fact that the work covered the sequence of probabilistic analysis, computational simulation and inference, as well as modelling, obeying the classical statistical paradigm.

This work aims, in a interdisciplinary context, to show how data from a public source of MET in Brazil can be explored and how such data can be used to fit probability distributions to data.

Then the fitted distributions are used to generate a simulation table that produces data on genotype x environment interaction.

These data has specific known parameters and can be used to compare AMMI, W-AMMI and GGE models in its properties to model genotype x environment interactions.

The structure of this thesis is based on six chapters. The second chapter, describes Embrapa maize trials and how its data was used in this work.

Chapter three describes the statistical and computational methods used to fit proper probability distributions to datasets.

By its turn, chapter four shows how a table can be used to generate simulations and data with specific parametric properties on genotype x environment interactions.

Chapter five puts the genotype x environment interactions in context and AMMI analysis history on plant breeding studies.

Finally, chapter six describes AMMI, w-AMMI and GGE models and their behaviour facing simulated datasets. This simulation based context is one important contribution of this text to the overall literature. Also its results show how such models can be used to analyze data independently of the underlying distribution.

A concise meaning of the thesis can be checked at appendices and R coding implementation in R code section.

References are displaced in each correspondent chapters.

The written final text was a challenge posed by the interdisciplinary characteristic of the work and by the multiple kinds of potential readers. The focus is to maintain a balance between technical mathematical rigor and comprehension by broader audiences.

References

- ALLARD, R., 1999 *Principles of plant breeding*. John Wiley, New York.
- AZEVEDO, A. J. B. V., FILHO, 2010b *Introdução à Estatística matemática aplicada - Fundamentos*. Create Space, Palo Alto.
- DELIGNETTE-MULLER, M. and C. DUTANG, 2015 Fitdistrplus: an r package for fitting distributions. *Journal of Statistical Software*, Articles **64**: 1–34.
- HALD, A., 2007 *A History of Parametric Statistical Inference from Bernoulli to Fisher, 1713-1935*. Springer, New York.
- HONGYU, K., M. GARCIA-PENA, L. ARAÚJO, and C. DIAS, 2014 Statistical analysis of yield trials by ammi analysis of genotype x environment interaction. *Biometrical Letters* pp. 89–102.
- KARIAN, A. and J. DUDEWICZ, 2011 *Handbook of fitting statistical distributions with R*. CRC Press, New York.
- SARTI, D., 2013 Gerenciamento de incertezas por análise de decisões:aplicações otimização da produção e demandas incertas. Universidade de São Paulo .
- STIGLER, S., 1986 *The history of statistics*. Harvard University Press, Massachusetts.

2 MULTI ENVIRONMENT MAIZE TRIALS

2.1 Introduction

One of the ground foundations of agriculture is plant breeding which the main objective is develop new cultivars. This process is continuous over time because of the fact that the relationship between plants, environmental issues, diseases and crop requirements is a very dynamic system as shown in ALLARD (1999).

One key issue in cultivar development is yield stability over environments as shown in SARTI (2013), YAN and KANG (2003) and DUARTE and VENCovsky (1999).

To assess the stability of a given genotype over a set of different environments experimental procedures can be conducted. Such experiments are named Multi Environment Trials.

Multi environmental trials are and will be necessary in the future to improve the knowledge on yield stability and specific genotype adaptation to different environments supplying the increasing demand for food as shown in FAO (2009).

A perspective on the use of multi environmental trials in agriculture and forestry activities can be seen in COE (2012).

Multi environmental evaluations can also play an important part in the decisions made in plant breeding as shown in SARTI (2013). They also provide guidelines for optimizing breeding programs and in countries where agriculture has an economic importance, these trials can be seen as a key factor for socio-economic development.

In 1962 in Brazil the Embrapa National Maize Trials was created to assist enterprises to comply with regulatory affairs EMBRAPA (2018).

These trials were an important source of data about maize genotypes in Brazil and also supplied data for studying the behaviour of cultivars across environments. Also, these MET were one of the most important public initiatives regarding maize breeding in Brazil.

Breeders, researchers and personnel from public and private sector were connected by a structured network of maize related professional during the trials period.

The aim of this article is to show the methodology applied by Embrapas's MET and use the data from the harvest of 2010/11 as source for the next chapters.

2.2 Methodologies

2.2.1 Embrapa national maize trials

Objectives

The main objectives of Embrapa Maize Trials are EMBRAPA (2018):

- network evaluation of maize genotypes.
- supply information of maize genotypes.
- study cultivars (genotypes) performance across the main environments of maize production in Brazil.
- make the cultivars recommendations regional.
- construct a data set about maize in Brazil.

Design of maize METs

The type of experiment used by the National Embrapa trials are block designs, described in EMBRAPA (2018) and as requirement for Usage Value Determination (VCU's) MAPA (2018).

The model considered for this trial in this work is show in Equation 6.2.1

$$y_{ijh} = \mu + g_i + e_j + (ge)_{ij} + b_{h(j)} + \epsilon_{ijh} \quad (2.1)$$

Where:

- y_{ijr} is the response of i-th genotype in j-th environment for block r and $i = 1...g, j = 1...e$ and $r = 1...b$
- μ is constant, frequently the general mean
- g_i is the effect of the genotype i
- e_j is the effect of the environment j
- $(ge)_{ij}$ is the interaction factor for genotype i in the environment j
- $b_{r(j)}$ is the effect of block h inside environment j.
- ϵ_{ijr} is the experimental error associated to genotype i, inside environment j in the block r, assumed $N(\mu, \sigma^2)$

After the trials are conducted, the data is collected and made available on Embrapa's website.

In the harvest of 2010/11 trials were conducted in central, north and south regions of Brazil.

The main considered variable was grain weight.

40 genotypes were tested and named g1 to g40 to preserve the identity of their commercial names.

Stratification in embrapa maize trials

The main criteria of classification in Embrapa maize trials:

- genotype cycle: early and super early
- region: center, north and south

2.2.2 Data and methodology of scrapping of Embrapa MET datasets

The data collected in trials comprises the variables described in Table 2.1. The main random variable considered was grains weight.

The process of data acquisition and related dynamics are shown in Figure 2.1.

Data collected in the trials are available at Embrapa website EMBRAPA (2018). In the case of the harvest considered in this article the data is available in PDF.

These files were downloaded and pre-processed by a computer program which than extracted the data. The download was made manually because there are no available tools for automatic downloading from the website EMBRAPA (2018)

The program developed for pre-processing was based on the use of regular expressions for scrapping

From all the available data in downloaded PDF files from the website not all were usable. Despite of this, the number of files which contained regular expressions was enough to assure the dynamics shown in this paper.

Table 2.1: Embrapa's trials variables

characteristics	description
grains weight (kg/ha)	grain weight kilograms per hectare obtained in trials
ears weight (kg/ha)	ears weight in kilograms per hectare obtained in trials.
broken plants and bedded plants (%)	percentage of plants that were diagnosed as broken or bedded in experimental units of the trials
flowering (days)	number of days until flowering
plants height (cm)	plants height in centimeters measured at the final stage of the trials
ears height (cm)	ears height in centimeters measured at the final stage of the trials
stand (x1000 plants)	number of plants in one hectare based at the final stand
number of ears (x1000 plants)	number of ears in one hectare based at the final stand
sick ears (%)	percentage of sick ears at the stand
grain moisture (%)	percentage of grains moisture

Figure 2.1: Overall thesis diagram

The processing of data via regular expressions produced four datasets named gw1, gwc, gwn and gws.

The relationship between the four datasets is as follow: After the scrapping through regular expressions of all the data available in embrapa's website we constructed dataset gw1. This data set contains observations from all the genotypes, all environments and all cycles, containing 826 observations on the variables shown in Table 2.2. In this paper we will consider just grains weight.

After this gw1 dataset was split in gwc, gwn and gws described in the previous table.

Table 2.2: Description of the overall datasets. gw1,gwc,gwn and gws are the datasets obtained in the process shown in Figure 2.1

dataset	characteristics	number of observations
gw1	overall dataset resulting from scrapping of data available at the website	826
gwc	early genotypes evaluated in the center region	156
gwn	early genotypes evaluated in the north region	18
gws	early genotypes evaluated in the south region	18

2.3 Discussion

2.3.1 Results

Datasets from Embrapa maize trials

The main results are the datasets described in Table 2.2.

The first dataset gw1 has its histograms shown in Figure 2.2. The summary statistics are shown in Table 2.3. This histogram will be used in the following articles in order to fix a probability model to the data sets. The model then will be used for computational simulations. The same pattern of results for the data sets gwc, gwn and gws are shown in Figures 2.2 and 2.3.

Figure 2.2: Histogram of gw1 and gwc

Figure 2.3: histograms of gwn and gws

Table 2.3: Datasets statistics. Unity is in Kg/ha. gws, gw1,gwn and gws are datasets described in Table 2.2

statistics	gw1	gwc	gwn	gws
minimum	1614	4832	2271	7705
1st quartile	7302	8111	6166	10776
median	8526	8774	7893	11356
mean	8304	8716	7507	11524
3rd quartile	9452	9429	9016	12752
maximum	15037	12033	11436	15037

2.4 Conclusions

This article has shown how Embrapa's maize MET can be used as an important source for genotypes performance across environments. The methodologies of the MET were described as long as well how the resulting data were scrapped in order to produce the data sets gw1, gwc, gws and gwn. Such data sets will be used in other portions of this thesis.

References

- ALLARD, R., 1999 *Principles of plant breeding*. John Wiley, New York.
- COE, R., 2012 Multi environment trials: an overview. World Agroforestry .
- DUARTE, J. and R. VENCovsky, 1999 *Interação Genótipos Ambiente: uma introdução à modelagem AMMI*. SBC, Ribeirão Preto.
- EMBRAPA, ., 2018 Ensaio nacional. Technical Report .
- FAO, 2009 How to feed the world. Technical Report pp. 1–33.
- MAPA, ., 2018 Rnc. Technical Report .
- SARTI, D., 2013 Gerenciamento de incertezas por análise de decisões:aplicações otimização da produção e demandas incertas. Universidade de São Paulo .
- YAN, W. and M. KANG, 2003 *GGE Biplot Analysis: A Graphical Tool for Breeders, Geneticists, and Agronomists*. Boca Raton.

3 FITTING STATISTICAL DISTRIBUTION TO DATA

3.1 Introduction

Phenomena modelling is a key issue in science and engineering. The development of a good representation for a problem or a situation brings insights and new perspectives about it.¹

Probability theory is the main set of tools developed to deal with uncertainty. Such reasoning system based axiomatically on initial Kolmogorov ideas evolved on time and led to the theory of probability distributions which, in association to advances of computational methods, led to methods of fitting probabilistic distributions to models of data.

In fact its possible to say that if a probability distribution underlying an uncertain phenomenon is determined, all the information about it can be considered completely described.

Another huge advantage of probabilistic models is the versatility they present, allowing us to cover situations in various domains. Proof of such statement is that the normal distribution itself and its applications to model data from social-economic to biological sectors. This property of such models brings an interdisciplinary perspective to statistical modelling.

This paper uses interdisciplinary techniques such as the ones described in CULLEN and FREY (1999) that are also used in risk assessment to model biological the quantitative dataset obtained in experiments described in Chapter 1.

The parameters of a probability distribution that models a phenomenon can be determined by analytically or iterative methods.

The aim of this paper is to fit probability distributions to data sets gw1, gwn, gwc and gws described in Chapter 2.

After fitting a parametric model to a dataset, the fitted model can be used to generate data which can feed models that claim to make inferences about them. This will be the case in the chapters about AMMI models.

3.2 Methodology

3.2.1 Fitting of univariate distributions to data

According to DELIGNETTE-MULLER and DUTANG (2015) before fitting distribution to a data set it's necessary to choose candidates from a pre defined set of distributions. In this article we will use the Cullen and Frey graphics method CULLEN and FREY (1999) which is based in third and fourth moments of probabilistic distributions.

Figure 3.1 shows examples of Cullen and Frey graphics. These graphics consist of plotting the square of skewness in x-axis and the kurtosis in y-axis. As described in CASELLA and BERGER (2002) and AZEVEDO (2010b) a non zero skewness reveals a lack of symmetry of the empirical distribution. Kurtosis measures the deviance of distribution tails when compared to a normal distribution (kurtosis equals 3).

The Cullen and Frey graphic is provided by the function `descdist` of the package `fitdistrplus` (DELIGNETTE-MULLER and DUTANG, 2015). For normal, logistic, uniform and exponential distributions there is only one value possible for skewness and kurtosis. For gamma and log normal distributions areas of possible values are represented consisting in lines or larger areas, the last is the case of beta distribution DELIGNETTE-MULLER and DUTANG (2015).

In the graphics of Cullen and Frey provided by the package `fitdistrplus` the values of skewness and kurtosis for the main distributions are displayed in such a way that the user can choose the distribution more closely to the target data set.

¹One way one to check if a subject understands a situation is to ask him to draw a schema of the problem

Figure 3.1: Cullen and Frey graphic example from DELIGNETTE-MULLER and DUTANG (2015)

The kurtosis and skewness and their corresponding unbiased estimators are shown in Equations 3.1, 3.2, 3.3, 3.4 and 3.5 where m_k denotes the k -th moment defined as shown in Equation 3.4 and x_i the observations CASELLA and BERGER (2002).

$$sk(X) = \frac{E[(X - E[X])^3]}{VAR(X)^{3/2}} \quad (3.1)$$

$$\hat{sk} = \frac{\sqrt{n(n-1)}}{n-2} \times \frac{m_3}{m_2^{3/2}} \quad (3.2)$$

$$krt(X) = \frac{E[(X - E[X])^4]}{VAR(X)^2} \quad (3.3)$$

$$\hat{kr} = \frac{n-1}{(n-2)(n-3)} ((n+1) \times \frac{m_4}{m_2^2} - 3(n-1)) + 3 \quad (3.4)$$

$$m_k = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^k \quad (3.5)$$

DELIGNETTE-MULLER and DUTANG (2015) states the fact that skewness and kurtosis are not robust measures and suggests the method of bootstrapping EFRON and TIBSHIRANI (1994) to evaluate uncertainties associated to the estimation of such parameters. In the Cullen and Frey graph available at fitdistrplus package its possible to choose the argument for boot which will produce bootstrapped values for skewness and kurtosis displayed in the Cullen and Frey graphics.

As discussed in DELIGNETTE-MULLER and DUTANG (2015) higher moments of distributions have greater variances, a problem that bootstrap cannot solve completely. So the graphic should be contrasted with properties of the random variable such as expected value and range.

Methods for fitting the parameter of a distribution

After choosing the candidates distributions to fit the data set, the next step is to choose a proper method for estimating the parameters. Here we review methods available at fitdistrplus R package DELIGNETTE-MULLER and DUTANG (2015).

Tabela 3.1: Goodness of fit statistics from DELIGNETTE-MULLER and DUTANG (2015) based in LUCEÑO (2006) where $F_i =^{\Delta} F(x_i)$

Statistic	General formula	Computational formula
Kolmogorov Sminov(KS)	$\sup F_n(x) - F(x) $	$\max(D_+, D_-)$ with $D_+ = \max_{i=1 \dots n} (\frac{i}{n} - F_i)$
Cramer von Mises(CvM)	$n \int_{-\infty}^{\infty} (F_n(x) - F(x))^2 dx$	$\frac{1}{12n} + \sum_{i=1}^n (F_i - \frac{2i-1}{2n})^2$
Anderson Darling(AD)	$n \int_{-\infty}^{\infty} \frac{(F_n(x) - F(x))^2}{F(x)(1-F(x))} dx$	$-n - \frac{1}{n} + \sum_{i=1}^n (2i-1) \log(F_i(1-F_{n+1-i}))$

Tabela 3.2: Anderson Darling variations for goodness of fit statistics from DELIGNETTE-MULLER and DUTANG (2015) based in LUCEÑO (2006) where $F_i =^{\Delta} F(x_i)$

Statistic	General formula	Computational formula
Right tail (ADR)	$\int_{-\infty}^{\infty} \frac{(F_n(x) - F(x))^2}{1 - F_x} dx$	$\frac{n}{2} - 2 \sum_{i=1}^n F_i - \frac{1}{n} 2 \sum_{i=1}^n (2i-1) \ln(\bar{F}_{n+1-i})$
Left tail (ADL)	$\int_{-\infty}^{\infty} \frac{(F_n(x) - F(x))^2}{F_x} dx$	$-\frac{3n}{2} + 2 \sum_{i=1}^n F_i - \frac{1}{n} 2 \sum_{i=1}^n (2i-1) \ln(\bar{F}_i)$
Right tailed 2nd order (AD2R)	$ad2r = \int_{-\infty}^{\infty} \frac{(F_n(x) - F(x))^2}{(1 - F_x)^2} dx$	$ad2r = 2 \sum_{i=1}^n \log \bar{F}_i + \frac{1}{n} \sum_{i=1}^n \frac{2i-1}{F_{n+1-i}}$
Left tailed 2nd order (AD2L)	$ad2l = \int_{-\infty}^{\infty} \frac{(F_n(x) - F(x))^2}{(F_x)^2} dx$	$ad2l = 2 \sum_{i=1}^n \ln \bar{F}_i + \frac{1}{n} \sum_{i=1}^n \frac{2i-1}{F_i}$
AD 2nd order (AD2L)	$ad2l + ad2r$	$ad2l + ad2r$

Maximum Likelihood:

Considering a parametric distribution $f(\cdot|\theta)$ with the parameter θ being an element of R^d , the **Maximum Likelihood** method is based, under i.i.d assumptions, in the likelihood function defined in Equation 3.5. The estimatives of the parameters given by the method are such that maximizes the likelihood function CASELLA and BERGER (2002), AZEVEDO (2010a) and ROUSSAS (2003).

$$L(\theta) = \prod_{i=1}^n f(x_i|\theta) \quad (3.6)$$

Maximum goodness of fit:

These methods are based in maximizing the goodness of fit, also named minimum distance estimation as described in D'AGOSTINO and STEPHENS (1986) and DELIGNETTE-MULLER and DUTANG (2015).

The package fitdistrplus offers eight different distances, three are stated as classics and are described in the Table 3.2 from DELIGNETTE-MULLER and DUTANG (2015) and the others are variations of the Anderson Darling distance as proposed in LUCEÑO (2006) and discussed in DELIGNETTE-MULLER and DUTANG (2015). The name of the Anderson Darling variations are related to the tail for which more weight will be given. Variations can be obtained when using second order AD statistics.

According to DELIGNETTE-MULLER and DUTANG (2015) and to D'AGOSTINO and STEPHENS (1986) the objective of the goodness of fitting statistics is to measure the difference between the fitted cumulative distribution F and the empirical F_n .

The gof criteria to be used can be chosen as arguments of the functions from the package fitdistrplus via method=mge. The same set of functions determines more information such as the estimated values for parameters, goodness of fit values for statistics related to each model and Akaike information criteria, as well as Bayesian information criteria.

As a concluding remark DELIGNETTE-MULLER and DUTANG (2015) citing CULLEN and FREY (1999) states that, given the structure of the underlying statistic, Anderson Darling methods captures the centrality and the tails of the distributions. The same authors, however, point out that AD from several statistical distributions fitted are difficult to compare among themselves. The same reasoning is applied for the number of parameters. When such a number is different between models the Cramer

von Mises and Kolmogorov Smirnov statistics lead to the decision for the model with greater number or parameters.

Method of the moments

Another method commonly used for estimation is the method of moments MME CASELLA and BERGER (2002) and MOOD *et al.* (1974). The dynamics of the method is to find a value for the parameter that equalizes the first raw moments of the parametric model to the corresponding empirical raw moments as can be viewed in Equation 3.6, where $k=1\dots d$, d is the number of parameters to be estimated and x_i the n observations from variable X , as stated in DELIGNETTE-MULLER and DUTANG (2015).

$$E(X^k|\theta) = \frac{1}{n} \sum_{i=1}^n x_i^k \quad (3.7)$$

For moments greater than 2 DELIGNETTE-MULLER and DUTANG (2015) it's relevant to make the procedure match centered moments given in Equation 3.7 and 3.8 where m_k is the empirical centered moment.

$$E(X|\theta) = \bar{x} \quad (3.8)$$

$$E((X - E(X|\theta))^k) = m_k \quad (3.9)$$

According to DELIGNETTE-MULLER and DUTANG (2015) the solution of the method of moments for the following distributions has a closed formula: normal, lognormal, exponential, Poisson, gamma, logistic, negative binomial, geometric, beta and uniform. The same authors states that for geometric, Poisson and exponential the parameter can be found simply by matching theoretical and observed means and for two variances matching is made for mean and variance DELIGNETTE-MULLER and DUTANG (2015) citing VOSE (2010). To carry out the method of moments for other distributions we can use numerical techniques, all of them available via optim methods in functions of package *fitdistrplus*. Alongside with maximum likelihood method of moments is one of the most used techniques to estimate parameters of a distribution CULLEN and FREY (1999) and DELIGNETTE-MULLER and DUTANG (2015).

As an important remark CULLEN and FREY (1999) affirm that the method of moments is more sensitive to outliers and the final decision for a particular method should consider a set of characteristics of the data sets.

Quantile Matching Estimation:

Theoretical quantile matching method consists in matching theoretical quantiles of parametric against the empirical quantiles as shown in DELIGNETTE-MULLER and DUTANG (2015) and TSE (2009). The Equation 3.9, described in DELIGNETTE-MULLER and DUTANG (2015) where $k = 1, \dots, d$ d being the numbers of parameters to estimate (dimension of θ if there is no fixed parameter) Q_{n,p_k} the empirical quantiles of the data for a given probability p_k . This method is available via package *fitdistrplus* DELIGNETTE-MULLER and DUTANG (2015). The quantiles matching is of special interest when we need to focus in a particular quantile of a distribution, this fact is common in risk assessment analysis.

$$F^{-1}(p_k|\theta) = Q_{n,p_k} \quad (3.10)$$

3.2.2 Fitting mixture of distribution to data

BENAGLIA *et al.* (2009) shows that in some cases populations of individuals may be divided in subgroups, implying that each subgroup is distributed according to a particular parametric model.

Formally speaking, a mixture of models can be stated as follows, mainly to show that such mixture adheres to all the probability functions properties requirements described in ROUSSAS (2003). The reasoning will be the same shown in ASSIS (2017).

Be K in $1, \dots, k$ distribution functions $F_1(\cdot), \dots, F_k(\cdot)$ and S another random variable that follows the multinomial distribution $(1 : p_1, \dots, p_k)$ for some positive integer, so we can write $P(S = k) = p_k$ for $1 \leq k \leq K$ and $\sum_{k=1}^K p_k = 1$, in this context we define a mixture of distributions as described in Equation 3.11 which has the following properties:

$$F(x) = \sum_{k=1}^K p_k F_k(x) \quad (3.11)$$

Properties:

- superior limit: $\lim_{x \rightarrow \infty} F(x) = \lim_{x \rightarrow \infty} \sum_{k=1}^K p_k F_k(x) = \sum_{k=1}^K p_k \lim_{x \rightarrow \infty} F_k(x) = \sum_{k=1}^K p_k = 1$
- inferior limit: $\lim_{x \rightarrow -\infty} F(x) = \lim_{x \rightarrow -\infty} \sum_{k=1}^K p_k F_k(x) = \sum_{k=1}^K p_k \lim_{x \rightarrow -\infty} F_k(x) = \sum_{k=1}^K p_k * 0 = 0$
- non decreasing function. Given $x \leq y$ so we have that $p_k F_k(x) \leq p_k F_k(y) \forall k$ therefore $F(x) = \sum_{k=1}^K p_k F_k(x) \leq \sum_{k=1}^K p_k F_k(y) = F(y)$
- as a mixture is a linear combination (convex combination) of right functions continuous it's also right continuous. Mathematically, if we have $x_n \downarrow x \in \mathbb{R}$ so: $\lim_{x_n \downarrow x} F(x_n) = \lim_{x_n \downarrow x} \sum_{k=1}^K p_k F_k(x_n) = \sum_{k=1}^K p_k \lim_{x_n \downarrow x} F_k(x_n) = \sum_{k=1}^K p_k F_k(x) = F(x)$

More details about probability functions definitions and properties can be found in ROUSSAS (2003) and JAMES (1981).

Regarding the random variable defined by the mixture S is a random variable with realizations in a k index, defined in $1, 2, \dots, K$ with its probability distribution given by $P(S = k) = p_k$

Defining the event $X_k = (X|S = k)$ ASSIS (2017) states that for a general borelian set B

$$P(X \in B) = \sum_{k=1}^K P(X \in B|S = k)P(S = k) = \sum_{k=1}^K P(X \in B)P(S = k) = \sum_{k=1}^K p_k P(X_k \in B) = F(x)$$

making a restriction $B = (-\infty, x]$ so we can write $F_X(x) = \sum_{k=1}^K p_k F_{X_k}(x)$

For more details on σ -Algebras and Borelian sets we suggest the reading of ROUSSAS (2003) and JAMES (1981).

ASSIS (2017) details the properties of the multinomial model behind the mixture model and states that a natural interpretation of a mixture model is a two stage process in which first we select a component distribution from a multinomial with parameter p and then we have a realization of the selected distribution.

In case we have a set of random vectors, each of the elements with the same dimensions, multivariate probabilities distributions are characterized.

Details about mixture models and its range of applications can be found in DAY (1969), McLA-CHLAN *et al.* (2006), BRAME *et al.* (2006), LINDSAY and STEWART (2010), PFANZAGL (1990), BAXTER

(2017), HOCHSTETLER *et al.* (2016), XU and JORDAN (1996), SCHLATTMANN (2009), G. J. McLACHLAN (1987) and UNKNOWN (2016)

Fitting mixture of normal distributions

When the family of distributions in 3.11 is from normal family we have a normal mixture of distributions. For a specific mixture of two normal distributions we would have the model described in Equation 3.12 in which p_i is the proportion of contribution from each component to the overall model.

Figure 3.2: Example of mixEM plot return from mixtools package BENAGLIA *et al.* (2009)

$$f(x|\mu, \sigma^2, p) = (p_1 \times \frac{1}{\sigma_1 \sqrt{2\pi}} e^{-(x-\mu_1)^2/2\sigma_1^2}) + (p_2 \times \frac{1}{\sigma_2 \sqrt{2\pi}} e^{-(x-\mu_2)^2/2\sigma_2^2}) \quad (3.12)$$

After determining that the underlying parametric model behind a phenomenon, a mixed model parameters need to be fitted. In this paper we will use the Expectation-maximization [EM] algorithm proposed in DEMPSTER *et al.* (1977) and discussed in CEPPELINI *et al.* (1955) in genetic counting methods context, by SUNDBERG (1974) and SUNDBERG (1976) in exponential families context.

BENAGLIA *et al.* (2009) shows how EM class algorithms are implemented in R language via mixtools package.

According to the approach presented in BENAGLIA *et al.* (2009) mixtures of models can be viewed as a special case of maximum likelihood estimation, where observations take the form of incomplete data, situation that implies in considering two sample spaces: one with the incomplete information and one with complete data. Details about the mathematical formulation of this problem can be found in BENAGLIA *et al.* (2009).

In general an **EM** algorithm consists in maximizing the operator $\theta^{t+1} = Q[\theta|\theta^t]$ and details are provided in BENAGLIA *et al.* (2009).

Be $\mathbf{X}$ a set of data observed, $\mathbf{Z}$ a set of latent data or some missing values, and a vector θ of parameters not known. We can define $L(\theta, X, Z) = p(X, Z|\theta)$ and the maximum likelihood estimator should be found maximizing the likelihood of the observed data, an expression that is, in general, not analytically treatable, so the EM method should be applied.

The procedure used by **EM** algorithms implies fundamentally in two stages:

- Expectation step [e-step]: defines an operator $Q[\theta|\theta^t]$ as expected value of log likelihood function of θ considering the conditional of Z given X and the current estimates of parameters of θ^t . The operator can be written as $Q[\theta|\theta^t] = E_{Z|X, \theta^t}[\log(\theta; X, Z)]$
- Maximization step [M-step]: determination of the parameters that maximize $\theta^{t+1} = Q[\theta|\theta^t]$

Details about convergence of EM algorithm can be found in WU (1983) and about its implementation in R language in BENAGLIA *et al.* (2009). The main method used here is the function `normalmixEM` from such package, details can be obtained via `?normalmixEM` command.

The application of `normalmixEM` functions return a "mixEM" object which delivers the plots shown in Figure 3.2. In this example, obtained from BENAGLIA *et al.* (2009) we can see the sequence $t \rightarrow L_X(\theta^t)$ of log likelihood observed values and the histogram with fitted models.

3.3 Results

3.3.1 Probabilistic models for gw1, gwc and gws data sets

The histograms for gw1 and gwc suggest that the models underlying them are uni variate models. Therefore for such datasets we will apply the methods proposed by CULLEN and FREY (1999), D'AGOSTINO and STEPHENS (1986), KARIAN and DUDEWICZ (2011) and DELIGNETTE-MULLER and DUTANG (2015). The results are obtained via functions of `fitdistrplus` R package DELIGNETTE-MULLER and DUTANG (2015). The complete results of histograms for gw1 data set are shown in Figures 3.10, 3.11 and 3.12, which describe also the behaviour of the different goodness of fitting statistics behaviour, described in the methodology section. The complete values for such goodness of fitting statistics are shown in Tables 3.6 and 3.8. Both the graphics place gw1 dataset at logistic distribution region. For general information about logistic distribution we suggest the reading of BALAKRISHNAN (1992).

Figure 3.3: Candidate distribution to gws dataset

Figure 3.4: Candidate distribution to gws dataset with bootstrapped values

Figure 3.5: gws dataset maximum likelihood, method of moments and quantiles

Once guided by Cullen and Frey methodology, we proceed with the fitting of logistic distribution parameters via the following methods: maximum likelihood, method of moments, quantiles matching methods and maximum goodness of fitting estimation. For the last one we use the maximization of Cramer-von-Mises, Kolmogorov- Smirnov and Anderson Darling in the AD, ADR, ADL, AD2R, AD2L

Figure 3.6: gws Anderson Darling (AD), Kolmogorov Smirnov (KS) and Cramer von Mises fitting measures

Figure 3.7: gws dataset different Anderson Darling fitting

and AD2 versions, described Tables 3.1 and 3.2. The results are shown in Tables 3.6 and 3.8

The fitting for gw1 data set shown in Tables 3.6, 3.7 and 3.8 as well as in Figures 3.6, 3.7 and 3.8 shows different estimates for the logistic distribution fitted.

Table 3.3: Fitting characteristics of models for gws data-set, mgeadl-maximum goodness of fit considering anderson darling left, mgeadr-maximum goodness of fit considering anderson darling right, mgecvm - maximum goodness of fit considering cramer von mises

parameter and standard deviation and goft criteria	mgeadl	mgeadr	mgecvm
sd (Kg/ha)	1707.624	1526.131	1572.987
mean (Kg/ha)	11593.602	11593.438	11591.002
Kolmogorov-Smirnov	0.0962	0.0808	0.0857
Cramer von Mises	0.0488	0.04218	0.0411
Anderson Darling	0.2923	0.3035	0.2721
Akaike's information criteria	2650.097	2650.555	2650.253
Bayesian information criteria	729.5325	729.5000	729.2969

Table 3.4: Fitting characteristics of models for gws data-set, mgead maximum goodness of fit anderson darling, mgead2 maximum anderson darling version 2, mgead2l maximum anderson darling version 2 left

parameter and standard deviation and goft criteria	mgead	mgead2	mgead2l
sd (Kg/ha)	1620.763	1676.081	1836.287
mean (Kg/ha)	11562.492	11452.696	11645.247
Kolmogorov-Smirnov	0.0964	0.1257	0.0947
Cramer von Mises	0.0432	0.07125	0.0735
Anderson Darling	0.2564	0.3518	0.4575
Akaike's information criteria	2649.689	2650.053	2651.214
Bayesian information criteria	729.2096	729.3926	730.6814

Figure 3.8: Cullen and Frey graph for gw1 data set

In graphic terms, we can say that the fitting via maximum likelihood, method of moments and

Table 3.5: Fitting characteristics of models for gws mgead2r maximum goodness fit considering Anderson Darling version 2 right tail, mgeks maximum goodness of fit considering smirnov kolmogorov, mme method of moments, qme - quantiles estimation method, mle - maximum likelihood estimation

parameter and stan- dard devia- tion and goft criteria	mgead2r	mgeks	mme	qme	mle
sd (Kg/ha)	1512.881	1389.39	1608.327	1192.898	1608.327
mean (Kg/ha)	11595.482	11569.35	11524.268	11705.495	11524.268
Kolmogorov- Smirnov	0.079	0.0732	0.1044	0.1274	0.1044
Cramer von Mises	0.0428	0.0597	0.0480	0.1458	0.0480
Anderson Darling	0.3173	0.5371	0.2684	1.6951	0.2684
Akaike´s information criteria	726.1655	727.7384	725.7548	735.7279	725.7548
Bayesian information criteria	729.5927	731.1656	729.1820	739.1550	729.1820

Figure 3.9: Cullen and Frey graph for gw1 data set with bootstrapped values

quantiles are closer, as shown in Figures 3.10, 3.11 and 3.12. On the other hand, as Cramer von Mises method returned a better fitting for central moments area and, Komogorov-Smirnov / Anderson Darling suggested similar results. Concluding, the fitting maximizing Anderson Darling and its variations AD2, AD2R and AD2L showed the expected behaviour. AD original measure resulted in a good fitting all along the data sets range.

Once we are going to use the fitted model to simulate data for posterior studies in stability across environments, we should choose the fitting that considers both tails and main body of distributions, in order to consider the whole significance of genotypes levels. Given the particular interest by optimizing

Figure 3.10: gw1 data set. Maximum likelihood, method of moments, method of quantiles.

Figure 3.11: gw1 data set. Cramer von Misès, Kolmogorov Smirnov, Anderson Darling.

the fitting previously discussed, we are going to choose the method of goodness of fitting that maximizes AD statistic.

As discussed in D'AGOSTINO and STEPHENS (1986) and DELIGNETTE-MULLER and DUTANG (2015) such criteria are better guaranteed by the Anderson Darling statistic, therefore, for gw1 data we will consider a logistic distribution with parameters of scale=1106.83 Kg and location=8392.22 Kg, or $gw1 \text{ Log}(sc = 1106.83, loc = 8392.22)$. Similarly, the same results for gwc datasets are presented in Tables 3.9, 3.10, 3.11, Figures 3.15, 3.16, and 3.17 as well as in Figures . This dataset resembles a little bit deviance for normality proper of logistic models. The Cullen and Frey CULLEN and FREY (1999) graphic shown in Figures 3.13 and 3.14 for dataset gwc also suggests the logistic model.

By the same decision criteria exposed to dataset gw1, we choose the logistic distribution for this gwc data with parameters $gwc (sc = 644.118Kg, loc = 8742.884Kg)$. On its turn the dataset gws, which represents the data from south, has its Cullen and Frey graphics shown in Figures 3.5, 3.6 and 3.7, its fitting results in Tables 3.3, 3.4 and 3.5. The visual graphics of such fitting are shown in Figures 3.5, 3.6 and 3.7. The results for its data set suggest the decision for a normal distribution with parameters $mean = 11562.492Kg$ and $sd = 1620.763$.

Figure 3.12: gw1 data set. Anderson Darling and its variations.

Table 3.6: Fitting characteristics of models for gw1. mgeadl - maximum goodness of fitting anderson darling left, mgeadr - maximum goodness of fitting anderson darling right, mgecvm - maximum goodness of fitting cramer von mise

parameter and standard deviation and goft criteria	mgeadl	mgeadr	mgecvm
Scale (Kg/ha)	1187.35	1021.68	1044.99
Location (Kg/ha)	8432.11	8425.97	8427.52
Kolmogorov-Smirnov	0,0537	0,0485	0.0462
Cramer von Mises	0,5468	0,2443	0.2343
Anderson Darling	4,0458	4,2423	3.7338
Akaike's information criteria	14937.11	14945.08	14940.43

Table 3.7: Fitting characteristics of models for gw1. mgead - maximum goodness of fit anderson darling, mgead2 - maximum goodness of fit anderson darling version 2, mgeadl2 - maximum goodness of fit anderson darling version 2

parameter and standard deviation and goft criteria	mgead	mgead2	mgeadl
Scale (Kg/ha)	1106.83	1191.374	1335.37
Location (Kg/ha)	8392.22	8243.014	8504.92
Kolmogorov-Smirnov	0.0403	0.0713	0.0890
Cramer von Mises	0.3179	1.1970	1.5453
Anderson Darling	3.0625	5.9978	9.6567
Akaike's information criteria	14933.91	14941.27	14967.53
Bayesian information criteria	14943.34	14950.70	14976.96

Table 3.8: Fitting characteristics of models for gw1 mgead2r maximum goodness fit considering Anderson Darling version 2 right tail, mgeks maximum goodness of fit considering smirnov kolnmogorov, mme method of moments, qme - quantiles estimation method, mle - maximum likelihood estimation

parameter and standard deviation and goft criteria	mgead2r	mgeks	mme	qme	mle
Scale (Kg/ha)	1044.91	1120.70	1143.592	1106.83	1126.230 (33.23)
Location (Kg/ha)	8406,1587	8405.06	8303.653	8446.997	8390.447 (67.30)
Kolmogorov-Smirnov	0.045	0.0374	0.0595	0.0450	0.0406
Cramer von Mises	0.2456	0.3324	0.6901	0.2586	0.3608
Anderson Darling	3.6274	3.1087	4.0327	4.6587	3.111
Akaike's information criteria	14940.22	14933.62	14935.42	14947.29	14933.55
Bayesian information criteria	14949.66	14943.06	14944.85	14956.73	14942.99

Table 3.9: Fitting characteristics of models for gwc. mgeadl - maximum goodness of fitting anderson darling left, mgeadr - maximum goodness of fitting anderson darling right, mgecvn - maximum goodness of fitting cramer von mise, mgead - maximum goodness of fit anderson darling, mgead2 - maximum goodness of fit anderson darling version 2, mgeadl2 - maximum goodness of fit anderson darling version 2

parameter and standard deviation and goft criteria	mgeadl	mgeadr	mgecvn
Scale (Kg/ha)	676.1327	609.5607	617.1611
Location (Kg/ha)	8759.5619	8756.6903	8758.8572
Kolmogorov-Smirnov	0.0547	0.0499	0.0479
Cramer von Mises	0.0785	0.0497	0.0490
Anderson Darling	0.4881	0.5047	0.4694
Akaike's information criteria	2650.097	2650.555	2650.253
Bayesian information criteria	2656.196	2656.655	2656.353

3.3.2 Probabilistic models for gwn and data set

The histograms of dataset gwn presented in Chapter 2 suggest a mixture of normal distributions. The density plots are shown in Figure 3.12 which describes the histogram of the model and the speed of convergence of loglikelihood function in EM algorithm.

These figures show the histogram of fitted model via EM algorithm and the evolution of the the loglikelihood function over iterations.

The parameters for gwn dataset are shown in Figure 3.13. The results suggests a binormal mixture in which the "left"side model is a $N(\mu = 4640.6826Kg, \sigma = 1239.7254Kg)$ contributing with 0.242768 for the whole mixture. The second component of the model is a $N(\mu = 8425.8423, \sigma = 1239.7254)$ and the contribution is 0.757232. The model is shown graphically in Figure

Table 3.10: Fitting characteristics of models for gwc. mgead - maximum goodness of fit anderson darling, mgead2 - maximum goodness of fit anderson darling version 2, mgeadl2 - maximum goodness of fit anderson darling version 2

parameter and standard deviation and goft criteria	mgead	mgead2	mgeadl2
Scale (Kg/ha)	644.118	667.9257	704.1054
Location (Kg/ha)	8742.884	8702.6782	8768.8723
Kolmogorov-Smirnov	0.0549	0.0734	0.0564
Cramer von Mises	0.0580	0.1053	0.1131
Anderson Darling	0.3960	0.5281	0.68205070
Akaike's information criteria	2649.689	2650.053	2651.214
Bayesian information criteria	2655.789	2656.152	2657.314

Table 3.11: Fitting characteristics of models for gwc. mgead2r - maximum goodness of fitting anderson darling version 2 right tail, mgeks - maximum goodness of fit kolmogorov smirnov, mme - method of moments, qme - method of quantiles, mlelogis - maximum likelihood

parameter and standard deviation and goft criteria	mgead2r	mgeks	mme	qme	mlelogis
Scale (Kg/ha)	644.9887	637.5088	651.6482	508.7827	648.33(43.79)
Location (Kg/ha)	8725.0267	8771.2055	8716.2436	8779.3331	8742.9183(89.46)
Kolmogorov-Smirnov	0.06158113	0.0433	0.0659	0.0881	0.0556
Cramer von Mises	0.0690	0.0553	0.0801	0.1798	0.0599
Anderson Darling	0.4161	0.4497	0.4431	2.4292	0.3973
Akaike's information criteria	2649.726	2649.844	2649.771	2663.935	2649.680
Bayesian information criteria	2655.826	2655.943	2655.871	2670.034	2655.779

A summary of the models chosen to be used in simulations described in Chapter is presented in Table 3.14

3.4 Conclusions

This article described the use of Cullen and Frey graphic methods CULLEN and FREY (1999) to select the appropriate distributions to gw1, gwc, gws and gwn datasets. The probabilistic models were adjusted to the data according to several goodness of fit criteria and estimation methods available at fitdisptplus R package. The results showed that the graphic method and the measures of gof led to the

Figure 3.13: Cullen and Frey graphic for gwc dataset

Figure 3.14: Cullen and Frey graphic for gwc dataset with bootstrapped values

Table 3.12: Mixture of distributions to dataset gwn

	model 1	model 2
proportion	0.242768	0.757232
mu	4640.6826	8425.8423
sigma	1239.7254	1239.7254
loglik	-2445.739	

Table 3.13: Mixture for gwn dataset

Figure 3.15: Fitting gwc data set. Maximum likelihood, method of moments and method of quantiles

Figure 3.16: Fitting gwc data set. Cramer von Misès, Anderson Darling and Kolmogorov Smirnov

final decision of considering the models shown in Results, regarding data sets gw1, gwc and gws. On the other hand, gwn dataset was best fitted by a binormal mixture distribution, using the EM methods available at package mixtools. The parameters of this model are shown in Table 3.13. The models fitted for the data sets gw1, gwc, gws and gwn can be used to simulate observations with known patterns and its results can be used to study models in which such data are used as input.

Figure 3.17: Fitting gwc data set. Anderson Darling and its variations

Table 3.14: Summary of the fitted model chosen

dataset	probabilistic model / parameters
gwc	logistic scale = 1106.83 kg location = 8392.22 kg
gwn	normal mixture mean1 = 4640.68 kg mean2 = 8425.84 kg standard deviation1=1239.72 kg standard deviation2=1239.72 kg
gws	normal mean = 11562.492 kg standard deviation = 1620.763 kg

References

- ASSIS, R. C., 2017 *Inferência em modelos de mistura via algoritmo EM estocástico*. Ph.D. thesis, Universidade Federal do São Carlos.
- AZEVEDO, A. J. B. V., FILHO, 2010a *Introdução à Estatística matemática aplicada - Distribuições e Simulação*. Create Space, Palo Alto.
- AZEVEDO, A. J. B. V., FILHO, 2010b *Introdução à Estatística matemática aplicada - Fundamentos*. Create Space, Palo Alto.
- BALAKRISHNAN, N., 1992 *Handbook of the Logistic Distribution*. Marcel Dekker, New York.
- BAXTER, R. A., 2017 pp. 841–844 in *Mixture Model*, edited by SAMMUT, C. and G. I. WEBB, Springer US.
- BENAGLIA, T., D. CHAUVEAU, D. HUNTER, and D. YOUNG, 2009 mixtools: An r package for analyzing mixture models. *Journal of Statistical Software, Articles* **32**: 1–29.
- BRAME, R., D. S. NAGIN, and L. WASSERMAN, 2006 Exploring some analytical characteristics of finite mixture models. *Journal of Quantitative Criminology* **22**: 31–59.
- CASELLA, G. and R. BERGER, 2002 *Statistical Inference*. Duxbury advanced series in statistics and decision sciences, Thomson Learning.
- CEPPELINI, B. R., M. SINISCALCO, and C. A. B. SMITH, 1955 The estimation of gene frequencies in a random-mating population. *Annals of Human Genetics* **20**: 97–115.
- CULLEN, A. and H. FREY, 1999 *Probabilistic Techniques in Exposure Assessment*. Plenum Press, New York.
- D'AGOSTINO, R. B. and M. STEPHENS, 1986 *Goodness of fit techniques*. Marcel Dekker Inc, New York.
- DAY, N., 1969 Estimating the components of a mixture of normal distributions. *Biometrika* **56**: 436–474.
- DELIGNETTE-MULLER, M. and C. DUTANG, 2015 Fitdistrplus: an r package for fitting distributions. *Journal of Statistical Software, Articles* **64**: 1–34.
- DEMPSTER, A. P., N. M. LAIRD, and D. B. RUBIN, 1977 Maximum likelihood from incomplete data via the em algorithm. *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B* **39**: 1–38.
- EFRON, B. and R. TIBSHIRANI, 1994 *An introduction to bootstrap*. Chapman Hall, USA.
- G. J. McLACHLAN, G. J. M., K. E. BASFORD, 1987 *Mixture Models*. Statistics: A Series of Textbooks and Monographs, CRC Press, 0th edition.
- HOCHSTETLER, A., D. J. PETERS, and M. DELISI, 2016 Classifying risk development and predicting parolee recidivism with growth mixture models. *American Journal of Criminal Justice* **41**: 602–620.
- JAMES, B., 1981 *Probabilidade: um curso em nível intermediário*. Projeto Euclides, Instituto de Matemática Pura e Aplicada, CNPq.
- KARIAN, A. and J. DUDEWICZ, 2011 *Handbook of fitting statistical distributions with R*. CRC Press, New York.
- LINDSAY, B. G. and M. STEWART, 2010 pp. 129–138 in *Mixture Models*, edited by DURLAUF, S. N. and L. E. BLUME, Palgrave Macmillan UK.

- LUCEÑO, A., 2006 Fitting the generalized pareto distribution to data using maximum goodness-of-fit estimators. *Comput. Stat. Data Anal.* **51**: 904–917.
- MCLACHLAN, G., L. B.-T. JONES, and R. BEAN, 2006 A simple implementation of a normal mixture approach to differential gene expression in multiclass microarrays. *Bioinformatics* **22**: 1608–1615.
- MOOD, A., F. GRAYBILL, and D. BOES, 1974 *Introduction to the Theory of Statistics*. McGraw-Hill international editions: Statistics series, McGraw-Hill.
- PFANZAGL, J., 1990 pp. 48–52 in *Mixture models*, Springer US.
- ROUSSAS, G., 2003 *An Introduction to Probability and Statistical Inference*. Statistics/Mathematics, Elsevier Science.
- SCHLATTMANN, P., 2009 *Medical Applications of Finite Mixture Models*. Statistics for Biology and Health, Springer-Verlag Berlin Heidelberg, first edition.
- SUNDBERG, R., 1974 Maximum likelihood theory for incomplete data from an exponential family. *Scandinavian Journal of Statistics* **1**: 49–58.
- SUNDBERG, R., 1976 An iterative method for solution of the likelihood equations for incomplete data from exponential families. *Communications in Statistics - Simulation and Computation* **5**: 55–64.
- TSE, Y., 2009 *Nonlife Actuarial Models: Theory, Methods and Evaluation*. International Series on Actuarial Science, Cambridge University Press.
- UNKNOWN, 2016 *Unsupervised Machine Learning in Python: Master Data Science and Machine Learning with Cluster Analysis, Gaussian Mixture Models, and Principal Components Analysis*. LazyProgrammer.
- VOSE, D., 2010 *Quantitative Risk Analysis. A Guide to Monte Carlo Simulation Modelling*. John Wiley.
- WU, C. F. J., 1983 On the convergence properties of the em algorithm. *Ann. Statist.* **11**: 95–103.
- XU, L. and M. I. JORDAN, 1996 On convergence properties of the em algorithm for gaussian mixtures. *Neural Computation* **8**: 129–151.

4 SIMULATION OF GENOTYPE X ENVIRONMENT INTERACTION DATA IN A BLOCK EXPERIMENTAL DESIGN

4.1 Introduction

Simulation of data has been proved a valuable tool for science and evaluation of statistical models.

The use of computational simulations can be verified in the following studies:

- CULLEN (1999) in peaks of energy in residence electric supply context
- ZACKHARY *et al.* (2009) which shows uses of the monte carlo methods in hotel profit optimization problems
- MUÑOZ (2003) that shows the use of methods in supply chain management questions
- BARTEZZAGHI and ZOTTERI (1999) who solves a problem in sporadically and irregular ("lumpy demands")
- VOSE (2010) that shows applications of the techniques in risk analysis sector

Another important factor that facilitates the use of simulations is the possibility of computational implementation of the methodologies.

In statistical modelling context we can use simulated data with pre established properties to check if a model that claims to detect such properties from data really does so.

The main objective of this article is to propose a methodology to simulate data from probabilistic data sets fitted in Chapter 3 in a block design context taking into consideration interactions between genotype and environment factors.

Simulations will be based in computing that uses monte carlo methods in R software.

A table will be used to generate the relationship between genotypes and environments. Details will be provided in Section 4.2.1.

The main result is a dataset called *simf* containing observations from 10 genotypes in 5 environments and 4 blocks experimental design.

This data set will be used as input for models in future chapters. The aim is to test the capacity of the chosen models to detect specific behaviours, such as adaptability on specific or across different environments.

Notwithstanding the length of this chapter it is an extremely important part of this thesis so it was written solely.

4.2 Methodology

4.2.1 Table to generate the genotypes x environments interaction

This section describes the table which will be used to simulate a dataset in block design via monte carlo method and comprises the main objective of this Chapter.

The model underlying the experiment and that will guide the simulations is shown in Equation 2.1

$$y_{ijh} = \mu + g_i + e_j + (ge)_{ij} + b_{h(j)} + \epsilon_{ijh} \quad (4.1)$$

Where:

- y_{ijr} is the response of i -th genotype in j -th environment for block r and $i = 1...g, j = 1...e$ and $r = 1...b$
- μ is constant, frequently the general mean
- g_i is the effect of the genotype i
- e_j is the effect of the environment j
- $(ge)_{ij}$ is the interaction factor for genotype i in the environment j
- $b_{r(j)}$ is the effect of block h inside environment j .
- ϵ_{ijr} is the experimental error associated to genotype i , inside environment j in the block r , assumed $N(\mu, \sigma^2)$

Figures 4.1 and 4.2 sections a) and b) are key to understand the methodology proposed in this article. It illustrates the process that comprises the methodology visually and schematically.

In the schema we can see that the backbone of the methods are constituted by two statistical references stages associated to a computational simulation stage, the procedures used were the following:

- **STATISTICAL REFERENCE 1:** To use fitted models from Chapter 3 as base. Then simulate 5000 observations from the related probabilistic model underneath it. Choose randomly or pre determine a quantile from these models to represent a genotype. The sources models are gwc for environments a1, a4, and a5, gws for a3 and gwn for environment a2.
- **STATISTICAL REFERENCE 2:** To use the quantile choosen in step 1 as mean or location parameter in simulations of data. The variance of such simulations will be the same as the source models.
- **COMPUTATIONAL SIMULATION:** The model to be used in simulations is the same as the source model from which a genotype was taken as specific quantile. The variance is the same from the model fitted in the Chapter 3. By example, if the source model is a logistic ditribution, as the one fitted for center region, the model used in simulations is logistic too. Then simulate 5000 observations from a model with parameters stated in stage 1 for the genotypes. Sample 4 single observations from the simulated data to represent the data for each of the 4 blocks considered in experimental design.

To force the specific adaptability of a given genotype to an environment, a quantile is pre determined This will be one of the main characteristics of interest for us in future chapters.

By example, we will be interested in the ability of GGEBiplots and AMMI to detect such specific adaptations. When it is not required to determine a specific adaptation the quantile is chosen randomly from the source model.

The relationship between environments and quantiles chosen to represent the genotypes is shown in Table 4.2 which comprises different possible relationships of chosen quantiles mentioned in 4.2.1 and environments. The models fitted in Chapter 3 and stated as source models in 4.2.1 are in Table 4.1.

By example, lets consider genotype g1, which will have a specific adaptation to environment a1, this means that this environment is the most appropriate for cropping this genotype. Also, this genotype has yield above the mean of the probability model fitted to environment a1 (center), chosen as the specific environment for this genotype.

In this case we choose this genotype to be represented by the 90th percentile of probability model fitted to center and south region, as well as the same quantile of the model 2 for normal mixture fitted to north region. The idea is to create a genotype that performs above the average and is stable across all the environments considered.

Tabela 4.1: Relationship among data sets from chapter

environment	dataset	related
		probabilistic model / parameters
a1,a4,a5	gwc	logistic scale = 1106.83 kg location = 8392.22 kg
a2	gwn	normal mixture mean1 = 4640.68 kg mean2 = 8425.84 kg standard deviation1=1239.72 kg standard deviation2=1239.72 kg
a3	gws	normal mean = 11562.492 kg standard deviation = 1620.763 kg

Figure 4.1: Reasoning used to generate simulations

The number of simulated data from mixtures is 5000 from each component. This will be represented as model1 and model2 in Table 4.2.

4.2.2 Results

The information presented in 4.2 was used as input in R to generate the simulated dataset `simf2`.

This dataset will be used in future chapters to study the capacities of the models to detect genotype x environment interactions.

Figure 4.2: Reasoning used to generate simulations

Table 4.2: Simulation guidelines

genotype	specific environment	quantile center 1	quantile center 2	quantile center 3	quantile north	quantile south
g1	a1 (center1)	90 determined	90 determined	90 determined	90 determined	90 determined
g2	a2 (north)	random 10	random 90	random 60	90 determined from model2	random 90
g3	a3 (south)	random 50	random 10	random 60	random model1 40	random 90
g4	a4 (center 2)	random 20	determined 10	random 70	random model2 40	random 80
g5	a5 (center 3)	random 40	random 20	determined 90	random model1 60	random 30
g6	a1 (center 1)	determined 90	random 70	random 80	random model2 70	random 40
g7	a4 (center 2)	random 40	determined 90	random 80	random model2 60	random 70
g8	a3 (south)	random 50	random 30	random 70	random model1 70	determined 90
g9	a4 (center 2)	random 70	determined 90	random 80	random model1 70	random 50
g10	a2 (north)	random 10	random 50	random 30	determined model 1 90	random 60

The resulting data set is comprised of five levels for the environment factor, 10 levels for genotypes evaluated in a 4 block statistical design.

4.3 Conclusions

This article showed how the proposed methodology was used to generate the dataset `simf2`, comprising 10 genotypes, from 5 environments in randomized block design containing 4 repetitions. The simulated data holds specific properties that will allow the evaluation of models for genetic x environment interaction evaluated in future chapters.

References

- BARTEZZAGHI, E. R. V. and G. ZOTTERI, 1999 A simulation framework for forecasting uncertain lumpy demand. *International Journal of Production Economics* **2**: 499–510.
- CULLEN, A. K., 1999 *Forecasting Electricity Demand using Regression and Monte Carlo Simulation Under Conditions of Insufficient Data*. West Virginia University, New York.
- MUÑOZ, F., 2003 *A Bayesian Framework For Modeling Demand In Supply Chain Simulation Experiments*. Instituto Tecnológico Autónomo de México, Cidade do México.
- VOSE, D., 2010 *Quantitative Risk Analysis. A Guide to Monte Carlo Simulation Modelling*. John Wiley.
- ZACKHARY, A., A. ATIYA, H. EL-SHISHINY, and N. GAYAR, 2009 *Forecasting Hotel Arrivals and Occupancy Using Monte Carlo Simulation*. Cairo University, Cairo.

5 OVERVIEW ON GENE-ENVIRONMENT INTERACTION

5.1 Introduction

The aim of this short article is to contextualize readers with non technical background about gene x environment interactions.

5.2 The biological universality of genotype and environment interactions

The relationship between individuals and environment plays a central role in understanding of biological phenomena.

Comprehension of how an individual cell or set of organisms behave in different ecosystems has been object of research and inquire across scientific history

In his seminal work Richard Wolterek proposed the concept of norm of reaction, to illustrate the patterns of phenotypic expression of a genotype across a set of different environments HARWOOD (1996). A comprehensive discussion about norm of reaction and phenotypic distribution can be found in GRIFFITHS *et al.* (2000).

The gene x environmental interaction is a concept with several applications within biology. This can be illustrated by the study of determinants of several environmental factors in development of human disorders such as schizophrenia, as shown in PICCHIONI and MURRAY (2007) and VAN OS (2004), as well as in the study of the IQ determinants.

LI *et al.* (2018) describes the genomic and environment factors in norm of reaction characterization. In some cases the concept of phenotypic plasticity is applied.

5.3 Classification of genotype x environment interaction

In the specific case of plant breeding, the definition of interaction of genotype and environment is the one proposed in SINEBO *et al.* (1972)

This definition states that gxe interaction presence is shown by the variability of genotypes response to different environments, implying in a change of the relative position of a genotype given a specific characteristic across several environments.

The presence of interactions is an important factor of confusion in new cultivars development decisions.

According to ROBERTSON (1959) GxE interaction can be:

- simple: is function of the difference in the the variability between genotypes inside environemnts
- complex: when the correlations between the measures of a given genotype inside different environments is absent.

The second kind is the most complex from plant breeding point of view.

According to CRUZ and REGAZZI (1997) the phenotypic expression is given by Equations 5.1 and 5.2 in cases of presence and absence of interactions.

$$P = G + E \quad (5.1)$$

$$p = G + E + GE \quad (5.2)$$

One of the most important techniques to manage GxE interaction is the realization of Multi Environmental Trials as the one described in 2.

The phenomenon of Genotype x Environment Interaction can be seen graphically in Figure 5.1. This figure is made using the simulated data set simf2.

Contemplating such figure we can check the relationship among phenotype and genotype in the senses that the scale of the variability between environments and genotypes varies according to the enviroments.

The same picture also allows the diagnosis of the classification proposed by SQUILASSI (2003).

- **Absence of interaction:** as the situation of genotype g9 between environments e4 and e5. There is zero interaction once the response is virtually the same.
- **Simple interaction quantitative:** the relative difference between genotype performance and ranking of genotype remains the same across environments. This is the case when comparing genotypes g8 and g3 across environments e1 and e2.
- **Cross interaction qualitative:** in the case of g1 /g2 and across environments e2 and e3, by example, there is a difference in magnitude and in ranking of genotypes across environments.

The cross interactions are the most common kind of interaction verified in real MET analysis and the most complex to manage in plant breeding. This leads to several problems in recommendation of genotypes based on means.

Figure 5.1: Genotype x Environment interaction

5.4 Historical perspectives on gene x environment modelling via AMMI models

In this section we aim to give a brief historical perspective on AMMI modelling of genotype x environment interactions. The idea is to organize and expand the reasoning exposed in HONGYU (2015).

YATES (1938) was one of the first authors to state GxE factor as linearly stated in the Equation 5.4.

$$(ge)_{ij} = \epsilon_i e_j + d_{ij} \quad (5.3)$$

Where:

- ϵ_i is the linear regression coefficient from genotype-i to environmental mean e_j
- d_{ij} is a deviation

In terms of statistical tests 5.4 stated a test based on expression of Equation 5.4.

$$(ge)_{ij} = K g_i e_j \quad (5.4)$$

Where:

- K is a constant

A sequel in the reasoning of mathematical modelling of gxe factor developed by MANDEL (1961) and expressed in the Equation 5.4, which proposes sets of lines that could be further tested for concurrency among them HONGYU (2015) quoting DIAS (2005).

$$(ge)_{ij} = \lambda \gamma_i e_j \quad (ge)_{ij} = \lambda \gamma_i \alpha_j \quad (5.5)$$

Noticing that in this model:

- Idea is to check if both and genotypes has the same coefficients or not.

By its turn GOLLOB (1968) and MANDEL (1969) quoted by HONGYU (2015) proposed a bi-linear term expressed in Equation 5.4.

$$(ge)_{ij} = \sum_{k=1}^n \lambda_k \gamma_{ik} \alpha_{jk} \quad (5.6)$$

Where:

- $\lambda_1 \geq \lambda_2, \dots, \lambda_n$
- γ_{ik}, α_{jk} satisfies $\sum_i \gamma_{ik} \gamma_{ik'} = \sum_i \alpha_{ik} \alpha_{ik'} = 0 \quad k \neq k' \quad \text{and} \quad \sum_i \gamma_{ik}^2 = \sum_i \alpha_{jk}^2 = 1$

Furthermore the mathematical reasoning led to the bilinear model presented in Equation 5.8. HONGYU (2015) states that this model is a generalization of the model of regression on the means but more flexible. HONGYU (2015) quoting DIAS (2005) argues that this model was called Additive Main Effects and Multiplicative Interaction Model - AMMI GAUCH (1986). The model can be stated as in the Equations 5.7, 5.8 and 5.9.

$$y_{ijr} = \mu + g_i + e_j + \sum_{k=1}^n \lambda_k \gamma_{ik} \alpha_{jk} + \rho_{ij} + b_{r(j)} + \epsilon_{ijr} \quad (5.7)$$

Where:

- y_{ijr} = is value for i-th genotype inside environment j in r-th block and $i = 1, 2, \dots, g$, $j = 1, 2, \dots, e$ and $r = 1, 2, \dots, h$;
- μ is a constant that corresponds to the general means
- g_i is the effect of genotype i
- e_j is the effect of the environment j
- λ_k is the square root of eigen value k from matrix $(\mathbf{GE})(\mathbf{GE})^t$ or $(\mathbf{GE})^t(\mathbf{GE})$ with $k = 1, 2, \dots, p$ and $p = \min\{(g-1)(e-1)\}$
- γ_{ik} is the i-th element of column vector γ_k associated to λ_k
- ρ_{ij} is the additional residual value that contains the two multiplicative effects terms not included in the model
- $b_{r(j)}$ is the effect of block r inside environment j ϵ_{ijk} is the experimental error associated to genotype i environment j and block r assumed iid normally distributed with $(\mu, \sigma^2/r)$ and r the number of blocks
- $n < p$ with $p = \text{rank of matrix GE}$

and GE is the matrix

$$\mathbf{GE} = \begin{pmatrix} \bar{y}_{1,1} & \bar{y}_{1,2} & \cdots & \bar{y}_{1,e} \\ \bar{y}_{2,1} & \bar{y}_{2,2} & \cdots & \bar{y}_{2,e} \\ \vdots & \vdots & \ddots & \vdots \\ \bar{y}_{g,1} & \bar{y}_{g,2} & \cdots & \bar{y}_{g,e} \end{pmatrix} \quad (5.8)$$

$$Y_{ijk} = \underbrace{\mu + g_i + e_j}_{\text{additive}} + \underbrace{\sum_{k=1}^p \lambda_k \gamma_{ik} \alpha_{jk}}_{\text{multiplicative}} + \epsilon_{ijk} \quad (5.9)$$

$$Y_{ijk} = \underbrace{\mu + g_i + e_j}_{\text{additive}} + \underbrace{\sum_{k=1}^n \lambda_k \gamma_{ik} \alpha_{jk} \sum_{k=n+1}^p \lambda_k \gamma_{ik} \alpha_{jk}}_{\text{multiplicative}} + \epsilon_{ijk} \quad (5.10)$$

5.5 Conclusion

This short article showed the evolution of the reasoning on GxE interactions modelling. As shown simple models evolved on time leading to the use of more sophisticated multiplicative additive modelling.

References

- CRUZ, C. and A. REGAZZI, 1997 *Modelos biométricos aplicados ao melhoramento genético*. UFV, Viosa.
- DIAS, C., 2005 *Mtodo para a escolha de componentes em modelo de efeito principal aditivo e interao multiplicativa*. Ph.D. thesis, Universidade de São Paulo, Piraciaba.
- GAUCH, H. G., 1986 Model selection and validation for yield trials with interaction. *Biometrics* pp. 705–715.
- GOLLOB, H., 1968 A statistical model which combines feature of factor analitic and analysis of variance techniques. *Psychometrika* **33**: 73–115.
- GRIFFITHS, A., J. MILLER, D. SUZUKI, R. LEWONTIN, and W. GELBART, 2000 *Introduction to Genetic Analysis*. W. H. Freeman.
- HARWOOD, J., 1996 Weimar culture and biological theory: A study of richard woltreck (1877–1944). *History of Science* **34**: 347–377, PMID: 11616300.
- HONGYU, K., 2015 *Comparação do GGE biplot-ponderado e AMMI-ponderado com outros modelos de interação genótipo $\times$ ambiente*. Ph.D. thesis, USP, Piracicaba.
- LI, X., T. GUO, Q. MU, X. LI, and J. YU, 2018 Genomic and environmental determinants and their interplay underlying phenotypic plasticity. *Proceedings of the National Academy of Sciences* **115**: 6679–6684.
- MANDEL, J., 1961 Journal of the american statistical association. *Journal of Statistical Software* **56**: 878–888.
- MANDEL, J., 1969 The partitioning of interactions in analysis os variance. *Journal of Research of the National Bureau of Standards, Series B*. **73**: 309–328.
- PICCHIONI, M. M. and R. M. MURRAY, 2007 Schizophrenia. *BMJ* **335**: 91–95.
- ROBERTSON, A., 1959 Experimental design on the mesurement of herdabilities and genetic correlations. *Biometrical genetics* .
- SINEBO, W., B. LAKEW, and A. FEYISSA, 1972 Genotype environment interaction: its implications in forest tree improvement. *Proceedings* .
- SQUILASSI, M., 2003 *Interao gentipos com ambientes*. Embrapa Tabuleiros Costeiros, Aracaju.
- VAN OS, J., 2004 Does the urban environment cause psychosis? *British Journal of Psychiatry* **184**: 287–288.
- YATES, W., F.; COCHRAN, 1938 The analysis of groups of experiments. *Journal of Agricultural Science* **28**: 556–580.

6 AMMI, WEIGHTED AMMI AND GGE BILOT METHODS

6.1 Introduction

One constant interest in statistics, design and analysis of experiments is the study of main effects of factors and, in many cases, on the interactions among different levels of factors.

In the context of Multi Environmental Trials, described in 2 and in plant breeding, one of the main issues to be evaluated is the interaction between the factors genotype and environment.

HONGYU *et al.* (2014), HONGYU *et al.* (2015) and DUARTE and VENCovsky (1999) show the role and importance of studying genotype x environment interactions for decisions in plant breeding. By its turn, YAN and KANG (2003) describes the role of models in cultivar recommendation.

Several models can be used to study interactions among factors and to evaluate genotype over environment interactions. This is the case of Additive main effects and multiplicative interaction - AMMI models GAUCH (1992) and GGE model YAN and KANG (2003), as well as their variants, such as the weighted AMMI HONGYU *et al.* (2015).

By the other hand, comparison of models can supply different parameters and information about the same key interest question.

Every model claims to answer a particular question. To check if this is valid we can use simulated data, with well established properties, such as the ones described in chapter 4 about simulation.

This paper aims to evaluate the behaviour of AMMI, weighted AMMI and GGE analysis in the context of the simulated dataset simf2 described in chapter 4.

We will describe the methodology underlying Analysis of Variance, AMMI modelling, weighted AMMI modelling and GGE analysis. Results of using such methods in the described dataset simf2 from 4 will be discussed.

6.2 Methodology

6.2.1 Analysis of Variance

Analysis of variance is a statistical procedure to verify the effect of two or more levels of treatments. Each experimental design has its own analysis of variance.

To accommodate reasoning we present first a single factor design as shown in

$$y_{ij} = \mu + \tau_i + \epsilon_{ij} \quad (6.1)$$

Where:

- μ is the general mean
- τ_i is the effect of treatment τ
- ϵ_{ij} is the expected mean error associated with the measure $\epsilon_{ij} \sim N(0, \sigma^2)$
- $SQ = \sum_{i=1}^I \sum_{j=1}^J y_{ij}^2 - C$
- $C = \frac{\sum_{i=1}^I \sum_{j=1}^J y_{ij}^2}{IJ}$
- $SQT = \frac{\sum_{i=1}^I y_{i.}^2}{J} - C$
- $\sum_{i=1}^I y_{i.}$ is the total of treatment i
- $SQ_{res} = SQ_{Total} - SQ_{treat}$

For which we are interested in testing the following hypotheses:

$$\begin{cases} H_0 \equiv \tau_1 = \tau_2 = \dots = \tau_i \equiv^{\mu+\tau_i} \mu_1 = \mu_2 = \dots = \mu_i \\ H_1 \equiv \tau_i \neq \tau_j \equiv^{\mu+\tau_i} \mu_i \neq \mu_j \quad \text{for at least one contrast} \end{cases} \quad (6.2)$$

The analysis of variance in the last case decomposes the total variance in different parts, named between treatments and experimental error (inside). In general these two variations can be measured by the sum of squares of each related component.

To test hypothesis H_o it's possible to use the test F in parametric contexts under the assumptions of normality.

It's possible to show that $MSTret/MSRes$ has a distribution F on (i-j) and I(J-1) degrees of freedom under iid and normality assumptions.

When the same experiments are repeated inside several environments a joint analysis can be made. In our case the main interest is to check if the interaction between genotype and environment are statistically significant.

The basic model of ANOVA for a block experiment conducted in several environments is in Equation 6.3 and its relative ANOVA table is:

Table 6.1: Analysis of Variance Table considering genotype x environment interactions

Source of Variation	Df	Sq	Mean Sq
blocks inside environments	e(r-1)	SQ_{bie}	MS_{Bie}
Genotypes	(g-1)	SQ_g	MS_g
Environments	(e-1)	SQ_e	MS_e
GxE Interaction	(g-1)(e-1)	SQ_{gxe}	MS_{gxe}
Residuals	e(g-1)(r-1)	SQ_{res}	MS_{gxe}
Total	(ger-1)	SQ_{res}	

$$y_{ijh} = \mu + g_i + e_j + (ge)_{ij} + b_{h(j)} + \epsilon_{ijh} \quad (6.3)$$

Equation 6.3 describes the joint analysis of variance for detecting genotype by environment interactions.

In this context of joint analysis, the mean analysis requires estimates of mean residual variance $QM_{EM} = \frac{\sum_j SQ_{Resj}}{\sum_{j=1} GF_{Res}}$ and $j = 1, 2, \dots, e$

6.2.2 AMMI Model

According to HONGYU (2015) and HONGYU *et al.* (2015) in data containing g genotypes and e environments we have g x e interaction factors to be estimated. Each effect is unfolded in its relative components

HONGYU (2015) citing GAUCH (1992), explains that the mean of a variable for a given genotype in an environment is the result of additive effects (additive main effects). The interaction effects by its turn are a result of a multiplication of contributions of the genotypes and environments (multiplicative interaction effects).

HONGYU (2015) states that given a set of g genotypes tested inside e environments with r repetitions can be shown in 6.2.2.

$$\mathbf{GE} = \begin{pmatrix} \bar{y}_{1,1} & \bar{y}_{1,2} & \cdots & \bar{y}_{1,e} \\ \bar{y}_{2,1} & \bar{y}_{2,2} & \cdots & \bar{y}_{2,e} \\ \vdots & \vdots & \ddots & \vdots \\ \bar{y}_{g,1} & \bar{y}_{g,2} & \cdots & \bar{y}_{g,e} \end{pmatrix} \quad (6.4)$$

HONGYU (2015) states that in cases when interactions are statistically significant the SQ_{gei} should be decomposed to avoid additional sum of squares in residuals. This is made by using the analytic factor from Gollob stated in the following expressions.

$$(ge)_{ij} = \sum_{k=1}^p \lambda_k \gamma_{ik} \alpha_{jk} \quad (6.5)$$

where:

- $\lambda_1 \geq \lambda_2, \dots, \lambda_p$
- $\gamma_{ik} \alpha_{jk}$ follows ortonomalization conditions $\sum_i \gamma_{ik} \gamma_{ik'} = \alpha_{jk} \alpha_{jk'} = 0$ for $k \neq k'$ and $\sum_i \gamma_{ik}^2 = \sum_j \alpha_{jk}^2 = 1$

The mathematical model underlying AMMI models is shown in Equation 6.2.2, 6.2.2 and 6.2.2, based in VENKOVSKY (1999) and HONGYU (2015).

$$y_{ijr} = \mu + g_i + e_j + \sum_{k=1}^n \lambda_k \gamma_{ik} \alpha_{jk} + \rho_{ij} + b_{r(j)} + \epsilon_{ijr} \quad (6.6)$$

Where

- y_{ijr} = is value for i-th genotype inside environment j in r-th block and $i = 1, 2, \dots, g$, $j = 1, 2, \dots, e$ and $r = 1, 2, \dots, h$;
- μ is a constant that corresponds to the general means
- g_i is the effect of genotype i
- e_j is the effect of the environment j
- λ_k is the square root of eigen value k from matrix $(\mathbf{GE})(\mathbf{GE})^t$ or $(\mathbf{GE})^t(\mathbf{GE})$ with $k = 1, 2, \dots, p$ and $p = \min\{(g-1)(e-1)\}$
- γ_{ik} is the i-th element of column vector γ_k associated to λ_k
- ρ_{ij} is the additional residual value that contains the two multiplicative effects terms not included in the model
- $b_{r(j)}$ is the effect of block r inside environment j ϵ_{ijk} is the experimental error associated to genotype i environment j and block r assumed iid normally distributed with $(\mu, \sigma^2/r)$ and r the number of blocks
- $n < p$ with $p = \text{rank of matrix GE}$

$$Y_{ijk} = \underbrace{\mu + g_i + e_j}_{\text{additive}} + \underbrace{\sum_{k=1}^p \lambda_k \gamma_{ik} \alpha_{jk}}_{\text{multiplicative}} + \epsilon_{ijk} \quad (6.7)$$

$$Y_{ijk} = \underbrace{\mu + g_i + e_j}_{\text{additive}} + \underbrace{\sum_{k=1}^n \lambda_k \gamma_{ik} \alpha_{jk}}_{\text{multiplicative}} + \sum_{k=n+1}^p \lambda_k \gamma_{ik} \alpha_{jk} + \epsilon_{ijk} \quad (6.8)$$

HONGYU (2015) citing GAUCH (1992), argues that the term $(ge)_{ij}$ is represented by a summation of p terms, each of the λ_k expressed in terms of the same unit of Y_{ij} by an effect of genotype (γ_{ik}) and an environmental factor (α_{jk}), both of them being dimensionless, or $\sum_{k=1}^n \lambda_k \gamma_{ik} \alpha_{jk}$. The term λ_k describes the information relative to the interaction in the k-plot, therefore, the summation of p plots comprises $(SQGEI) = \sum_{k=1}^n \lambda_k^2$ and the effects γ_{ik} and α_{jk} represent the weights of genotypes i and environment j in this specific part of the interaction.

HONGYU (2015) and GAUCH (2013) state that AMMI models do not recover all $SQGEI$, but just the part that is most determined by genotype and environment (rows and columns of matrix GE), the model aims to describe the deterministic component. HONGYU (2015) and VENKOVSKY (1999) state that AMMI models separate the pattern $\sum_{k=1}^n \lambda_k \gamma_{ik} \alpha_{jk}$ from the noise $\sum_{k=n+1}^p \lambda_k \gamma_{ik} \alpha_{jk}$.

HONGYU (2015) states that each member from the matrix in Equation 6.9 can be given by the relation in Equation 6.2.2. For details about the single value decomposition of matrix GE we recommend HONGYU (2015).

$$(\hat{ge})_{ij} = Y_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y}_{..} \quad (6.9)$$

HONGYU (2015), GAUCH (2013) and DUARTE and VENKOVSKY (1999) state that AMMI models explain the sum of squares of the interaction by an approximation of rank n to the matrix GE, in general with n much less than p, this aims to explain the GEI by a reduced number of axes, resulting in an informative model without the loss of degrees of freedom.

According to HONGYU *et al.* (2014), HONGYU *et al.* (2015) and HONGYU (2015) procedures can be used to determine the number of axes to be retained in order to explain the patterns related to the interaction. The authors state that one method is to quantify the number of degrees of freedom related to $SQGEI$ to each member of AMMI family. Thus, we can get the mean square related to each plot. After that an F-test can be performed to evaluate the significance of each component related to the mean square error.

HONGYU (2015) states that a stop rule is required to select which AMMI model to be used ($AMMI_0, AMMI_1, \dots, AMMI_n$) is based in a F test for the axes of interaction.

HONGYU (2015), CORNELIUS (1993) state the statistic described in Equation 6.2.2 which is considered a robust method. Under the hypothesis that there are no more than n multiplicative terms for the interaction the statistics follows a F distribution with parameters $(g - 1 - n)(e - 1 - n)$ degrees of freedom for residuals. In this case HONGYU (2015) states that this can be a good test to get the n+1 first components of the interaction.

$$F_r = \frac{SQGEI - \sum_{k=1}^n \lambda_k^2}{f_2 QM_{Res}} \quad (6.10)$$

where

- λ_k is the k-th eigen value of the matrix $(GE)(GE)^t$ and $QMres$ is the mean square of the residuals

6.2.3 Weighted AMMI model

HONGYU (2015) states that in cases when a double entry table of data containing genotypes and environments has heterocedasticity among environments, the values of the table must have different weights for its related residual squares in the estimation processes.

The same author quoting ROBERTSON (2014) states that when the variance and covariance structure for the error differs significantly among the environments this should be considered in AMMI analysis.

In this paper, we use the generalization proposed in HONGYU *et al.* (2014), HONGYU *et al.* (2015) and HONGYU (2015) named as weighted AMMI and stated in Equation 6.2.3.

$$y_{ijr}/s_j^2 = \mu + g_i + e_j + \sum_{k=1}^n \lambda_k \gamma_{ik} \alpha_{jk} + \rho_{ij} + b_{r(j)} + \epsilon_{ijr} \quad (6.11)$$

where

- y_{ijr} = is value for i-th genotype inside environment j in r-th block and $i = 1, 2, \dots, g$, $j = 1, 2, \dots, e$ and $r = 1, 2, \dots, h$;
- s_j^2 is the variance of experimental error of environment j or means square residual for the related environment
- μ is a constant that corresponds to the general means
- g_i is the effect of genotype i
- e_j is the effect of the environment j
- λ_k is the square root of eigen value k from matrix $(\mathbf{GE})(\mathbf{GE})^t$ or $(\mathbf{GE})^t(\mathbf{GE})$ with $k = 1, 2, \dots, p$ and $p = \min\{(g-1)(e-1)\}$
- γ_{ik} is the i-th element of column vector γ_k associated to λ_k
- ρ_{ij} is the additional residual value that contains the two multiplicative effects terms not included in the model
- $b_{r(j)}$ is the effect of block r inside environment j ϵ_{ijk} is the experimental error associated to genotype i environment j and block r assumed iid normally distributed with $(\mu, \sigma^2/r)$ and r the number of blocks

6.2.4 Genotype main effects + genotype environment interaction - GGE methods

Sites regression models

Sites regression is a fundamental concept originated by the kind of multi environmental trials described in 2. The reasoning behind this method is to explore the GEI interaction via the fundamental relationship expressed in Equation 6.2.4 and its fundamental consequence: $P - E = G + GE$ HONGYU (2015) quoting YAN and KANG (2003).

$$p = G + E + GE \quad (6.12)$$

HONGYU (2015) and HONGYU *et al.* (2014), HONGYU *et al.* (2015) argues that Sites regression methods are based in biplot methods. In this sense, HONGYU (2015) argues that the methods of sites regression is similar to AMMI modelling, the difference being that, in Sites regression the principal effects of genotypes are considered together with GEI effect, meanwhile, in AMMI models they are considered additive effects.

Sites regression can be mathematically posed as in Equation 6.13.

$$\bar{Y}_{ij} = \mu + G_i + E_j + (GE)_{ij} \quad \bar{Y}_{ij} - \mu - E_j = G_i + (GE)_{ij} \quad (6.13)$$

Where:

- $\bar{Y}_{ij}$ is the general mean of i-th genotype inside the j-th environment
- μ is the general mean
- G_i is the effect of i-th genotype
- E_j is the effect of environment j
- GE_{ij} is the effect of the genotype x environment interaction

HONGYU (2015) states that sites regressions is as shown in Equation:

$$\bar{Y}_{ij} - \mu - E_j = g_{i1}e_{1j} + g_{i2}e_{2j} + \epsilon_{ij} \quad (6.14)$$

These relationships can be rewritten as in Equation 6.14 and Equation 6.15 which can be used to describe a single value decomposition as shown in HONGYU (2015) quoting YAN and KANG (2003).

$$\bar{Y}_{ij} - \mu - E_j = \lambda_1 \mu_{i1} v_{j1} + \lambda_2 \mu_{i2} v_{j2} + \epsilon_{ij} \quad (6.15)$$

Where:

- $\lambda_1 \quad \lambda_2$ are the greatest eigenvalues of first and second principal components
- μ_{i1} and μ_{i2} are the eigen vectors of genotype i for Principal components 1 and 2 and v_{j1} and v_{j2} the eigen vectors of environment j to Principal Components 1 and 2.

GGE Biplot

HONGYU (2015), YAN and KANG (2003), YAN (2006), YAN and HOLLAND (2010) and YAN (2011) state that the GGE biplot model is based in a graphical representation of a matrix of data.

These biplots models are built based on the first and second principal components that use Site Regression modelling.

The same authors state that, in conditions when the first principal components is highly correlated to the main effects of genotypes, it can be viewed as a proportion of yield given by genotype effect. HONGYU (2015), HONGYU *et al.* (2014) and HONGYU *et al.* (2015) argue that the second principal component represents the GEI part of the variable.

HONGYU (2015), HONGYU *et al.* (2014) and HONGYU *et al.* (2015) state also that three main aspects can be obtained via gge biplot YAN and KANG (2003):

- Grouping of environments with similar behaviours showing the genotypes with greater potential and its identification inside each subgroup of environments

- inter relationship between environments indicating the best environment to test the cultivars and indicating which environment is less favorable
- inter relationship among genotypes making possible their comparisons and ranking

HONGYU (2015), YAN and KANG (2003) state that, When different genotypes are adapted to different groups of environments and the variation between groups is greater than inside groups, we have the formation of a mega environment.

HONGYU *et al.* (2014), HONGYU *et al.* (2015) and TRETHOWAN (2001) show that this definition of mega environments can be viewed as an important decision framework for plant breeders.

HONGYU (2015) states that, in mega-environments studies, the mean of the related graphic is not associated to the general mean but, to the environment mean instead. HONGYU (2015), HONGYU *et al.* (2014), HONGYU *et al.* (2015) and YAN and KANG (2003) state that the cossin of an angle between two environments can be seen as the correlation among them.

HONGYU (2015), HONGYU *et al.* (2014), HONGYU *et al.* (2015), YAN (2001), YAN (2002) and YAN and KANG (2003) state that GGE biplot is built by means of the simple dispersion of g_{i1} and g_{i2} for genotypes and e_{1j} and e_{2j} for environments. HONGYU (2015) states that the Equation 6.13 is environment centered given the origin to one of the main factors to be considered in GGE analysis: Evaluations should consider only one variable by MET. example given: grains weight.

To deal with this issue YAN and KANG (2003) proposed the correction presented in Equation 6.16

$$\frac{Y_{ij} - \bar{\mu} - E_j}{d_j} = g_{i1}e_{1j} + g_{i2}e_{2j}\epsilon_{ij} \quad (6.16)$$

Where

- $d_j = \sqrt{\frac{1}{g-1} \sum_{i=1}^g (Y_{ij} - \bar{Y}_{.j})^2}$ is considered the standard phenotypic deviation for environment j
- $\bar{Y}_{ij}$ is the response of genotype i and environment j
- $\bar{Y}_{.j}$ is the mean of environment j
- g is the number of tested genotypes

HONGYU (2015) states that for MET with just one characteristic, the previously cited correction assumes that environments have the same importance, thus, do not allow to detect differences among environments.

A second transformation can be obtained via

$$\bar{Y}_{ij}/s_j - \mu - E_j = g_{i1}e_{1j} + g_{i2}e_{2j} + \epsilon_{ij} \quad (6.17)$$

Where:

- $s_j = \sqrt{\frac{1}{m(r-1)} \sum_{i=1}^m \sum_{k=1}^y (Y_{ijk} - \bar{Y}_{.ij})^2}$
- $\bar{Y}_{.ij}$ is the mean of genotyoe i and environment j in repetition k inside environment j.

Single values decomposition in GGE biplots

HONGYU (2015) and YAN and KANG (2003) state that single value decomposition is very important in biplot analysis returning the principal components providing measures of the variability captured by each of these component.

$$P_{ge} = U_{g,r} S_{r,r} V_{e,t}^T [r \leq \min(g, e)] \quad P_{ij} = \sum_{k=1}^r u_{ik} \lambda_k v_{kj} \quad (6.18)$$

Where

- P_{ge} is a interactions matrix (G+GEi)
- $U_{g,r}$ is a matrix containing g rows and r columns which characterizes genotypes
- $V_{e,t}$ has r rows and e columns describes environments
- $S_{r,r}$ is the diagonal matrix containing r sigle values.
- Single value decomposition decomposes the matrix in r principal components each of them containing genptype vector u_i , vector environment v_j and single values λ decomposed in P_{ij} expression
- $\lambda_k \geq \lambda_{k+1}$,
- r is the number of principal components to represent P $r \leq \min(g, e)$. When $r < g$ there are association between genotypes and if $g < r$ we have association between environments.

HONGYU (2015) and YAN and KANG (2003) affirm that the single value decomposition related to Equation 6.18 and its related items. HONGYU (2015) and YAN and KANG (2003), state that the principal components cannot be directly computed before making the partition of the single values in genotypes and environments. This partition is described in Equation 6.18

For a biplot (dimension=2) HONGYU (2015) state the following the Equation 6.2.4.

$$P_{ij} = (\lambda_1^s u_{i1}) \lambda_1^{1-s} v_{j1} + (\lambda_2^s u_{i2}) \lambda_2^{1-s} v_{j2} + \epsilon_{ij} \quad (6.19)$$

Where

- $s=[0;0.5;1]$ is the factor of partition of the single values
- $\lambda_1^s u_{i1}$ is used as abcissa and $\lambda_2^s u_{i2}$ for genotypes
- $\lambda_1^{1-s} v_{j1}$ as abcissa $\lambda_2^{1-s} v_{j2}$ as ordinate for environments

HONGYU (2015) and YAN and KANG (2003) inform that the factors in previous equations are used to re-scale scores of the rows and columns for a better visual interpretation of the biplot. In this sense, if $s=1$ the single value partition is centered in genotypes, if $s=0$ the allocation is made completely to environment scores, if the value is $s=0.5$ squared roots of λ_k will be allocated for both genotype and environment scores.

HONGYU (2015) states that partition centered in genotypes, or in environments, allow the evaluation of genotypes and environments of test.

HONGYU (2015) and YAN and KANG (2003) state that the approximation of any element of matrix P, of rank 2, can be estimated, visually, by the inner product of genotype and environment and the cossin of the angle (inner product property of biplot).

6.2.5 Information criteria

HONGYU (2015), HONGYU *et al.* (2015) and YAN (2006) describe an information criteria to evaluate the accuracy of a biplot in the graphical description of a two entry data table.

Given a specific double entry data table, containing g genotypes and e environments, the number of principal components to represent such table is $k = \min(e, g - 1)$ for GGE Biplot and $\min(e - 1, g - 1)$ for AMMI models HONGYU (2015), HONGYU *et al.* (2015) and YAN (2006) state that, if there is no correlation among environments, all the k PCS should be completely independent and the proportion of total variance explained by each principal component is $1/k$ HONGYU (2015). k is the rank of GE matrix.

When some correlation among environments is present, the proportion of variance explained by the first principal components should be greater than $1/k$, and the variance explained by other pcs should be less than $1/k$ HONGYU (2015), HONGYU *et al.* (2015) and YAN (2006).

This information criteria can be measured as in Equation:

$$\frac{\text{varibilitycapturedbyPC}_i}{\text{totalvariabilty}} \times k \quad (6.20)$$

Where

- $k = \min(e, g - 1)$ for GGE
- $k = \min(e - 1, g - 1)$ for AMMI

If the information criteria is such as $IR > 1$, we have some association among environments, if it is approximately 1 there are no patterns, and when IR is less than 1, no information is present.

6.3 Results

As a protocol for proceeding AMMI analysis, we will use the recommendations presented by GAUCH (2013), who suggests the following steps (i) analysis of variance, (ii) model diagnosis, (iii) mega-environment delineation, and (iv) agricultural recommendations.

The same reasoning will be applied to weighted AMMI and an analogous process to GGE analysis.

6.3.1 AMMI Analysis of simulated dataset

Analysis of Variance

The ANOVA table for dataset simf2, containing 10 genotypes simulated in 5 environments is shown in 6.2.

From the mentioned table, we can deduce that environments, genotypes and repetition inside environments, as well as the interaction GxE, were all statistically significant, justifying the posterior analysis of interaction.

The coefficient of variation is 4,74%.

The table of means of genotypes across environments is presented in Table 6.3 and in Table 6.5. Data show that the sum of squares of the interaction is 246221662.25 which comprises 21.8% of the total data variability.

We can check that, as expected, because of the patterns used in simulations, g1 was the highest mean production genotype followed by g2, g4, g5 and g6.

Regarding environments, e3 and e5 are the ones with greater mean of yield. This is exactly the behaviour of the simulated datasets.

Figure 6.1: Interaction Plot for simf2

Tabela 6.2: Anova for Simulated data set simf2 Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
ENV	4	617225742	154306436	15.192	3.721e-05 ***
REP(ENV)	15	1152357920	10157195	45.572	< 2.2e-16 ***
GEN	9	104580085	11620009	52.135	< 2.2e-16 ***
ENV:GEN	36	243429709	6761936	30.339	< 2.2e-16 ***
Residuals	135	30089062	222882		

Another way of seeing the GEI interaction is in Figure 6.1.

Considering the matrix stated in Equation 6.4 , we proceed the interaction analysis by the related decomposition in single values $GE = USV^t$. The genotypes means are shown in Table 6.5. This matrix will have rank $p = \min(9, 4)$ so the interaction can be decomposed in 4 principal components.

The decomposition of such matrices is presented in Matrices S, U and V.

U matrix contains the singular vector to the left of the interaction, S is a diagonal matrix containing singular values and V a matrix whose columns describe singular values to the right of the interaction.

$$S = \begin{pmatrix} 6355.64 & 0 & 0 & 0 \\ 0 & 3625.09 & 0 & 0 \\ 0 & 0 & 2147.54 & 0 \\ 0 & 0 & 0 & 1646.21 \end{pmatrix}$$

Tabela 6.3: Means of AMMI models

Line	ENV	GEN	Prod	RESIDUAL
1	e1	g1	10961.80	251.41
2	e1	g10	8088.94	-606.71
3	e1	g2	8625.06	-1037.03
4	e1	g3	9509.42	514.38
5	e1	g4	8625.06	-1379.67
6	e1	g5	9230.74	687.62
7	e1	g6	10961.80	1524.67
8	e1	g7	9230.74	-635.67
9	e1	g8	9509.42	10.41
10	e1	g9	10072.42	670.60
11	e2	g1	9751.12	1477.24
12	e2	g10	5965.96	-293.18
13	e2	g2	9751.12	2525.54
14	e2	g3	3681.58	-2876.96
15	e2	g4	9234.70	1666.48
16	e2	g5	4687.14	-1419.48
17	e2	g6	8825.58	1824.96
18	e2	g7	8472.29	1042.39
19	e2	g8	5040.42	-2022.08
20	e2	g9	5040.42	-1924.89
21	e3	g1	13295.11	-458.81
22	e3	g10	12784.19	1045.01
23	e3	g2	13295.11	589.49
24	e3	g3	13295.11	1256.53
25	e3	g4	12619.96	-428.30
26	e3	g5	10308.60	-1278.05
27	e3	g6	10759.10	-1721.56
28	e3	g7	12085.10	-824.84
29	e3	g8	14456.07	1913.54
30	e3	g9	12352.35	-93.00
31	e4	g1	10961.80	-634.92
32	e4	g10	9509.42	-72.56
33	e4	g2	9509.42	-1039.00
34	e4	g3	10434.40	553.03
35	e4	g4	10961.80	70.75
36	e4	g5	10434.40	1004.95
37	e4	g6	9509.42	-814.04
38	e4	g7	10961.80	209.06
39	e4	g8	10434.40	49.07
40	e4	g9	10961.80	673.65
41	e5	g1	10961.80	-634.92
42	e5	g10	9509.42	-72.56
43	e5	g2	9509.42	-1039.00
44	e5	g3	10434.40	553.03
45	e5	g4	10961.80	70.75
46	e5	g5	10434.40	1004.95
47	e5	g6	9509.42	-814.04
48	e5	g7	10961.80	209.06
49	e5	g8	10434.40	49.07
50	e5	g9	10961.80	673.65

Table 6.4: Coordinates of Simf2 AMMI biplot

	type	Prod	PC1	PC2	PC3	PC4
g1	GEN	11186.33	21.13	-2.96	11.63	-0.00
g10	GEN	9171.59	-5.29	19.32	-1.20	0.00
g2	GEN	10138.03	33.72	25.90	5.36	0.00
g3	GEN	9470.98	-40.07	5.62	7.24	-0.00
g4	GEN	10480.66	21.75	8.05	-25.83	0.00
g5	GEN	9019.06	-18.41	-30.47	-14.96	0.00
g6	GEN	9913.07	28.09	-29.57	26.12	0.00
g7	GEN	10342.35	14.13	-5.31	-19.17	-0.00
g8	GEN	9974.94	-29.04	22.69	12.35	0.00
g9	GEN	9877.76	-26.00	-13.27	-1.53	0.00
e1	ENV	9481.54	-11.26	-32.38	33.84	0.00
e2	ENV	7045.03	71.95	4.37	-2.56	0.00
e3	ENV	12525.07	-22.13	48.85	13.18	0.00
e4	ENV	10367.87	-19.28	-10.42	-22.23	0.00
e5	ENV	10367.87	-19.28	-10.42	-22.23	0.00

Table 6.5: Genotype x Environment means of genotypes from data simf2

genotype	e1	e2	e3	e4	e5
g1	10961.80	9751.12	13295.11	10961.80	10961.80
g2	8625.06	9751.12	13295.11	9509.42	8625.06
g3	9509.42	3681.58	13295.11	10434.40	9509.42
g4	8625.06	9234.70	12619.96	10961.80	8625.06
g5	9230.74	4687.14	10308.60	10434.40	9230.74
g6	10961.80	8825.58	10759.10	9509.42	10961.80
g7	9230.74	8472.29	12085.10	10961.80	9230.74
g8	9509.42	5040.42	14456.07	10434.40	9509.42
g9	10072.42	5040.42	12352.35	10961.80	10072.42
g10	8088.94	5965.96	12784.19	9509.42	8088.94

$$U = \begin{pmatrix} 0.24 & -0.04 & 0.33 & 0.02 \\ -0.07 & 0.30 & -0.17 & -0.22 \\ 0.39 & 0.42 & -0.02 & -0.33 \\ -0.52 & 0.08 & 0.29 & 0.19 \\ 0.28 & 0.21 & -0.19 & 0.60 \\ -0.18 & -0.48 & -0.19 & 0.36 \\ 0.33 & -0.51 & 0.50 & -0.22 \\ 0.21 & -0.06 & -0.40 & 0.11 \\ -0.40 & 0.35 & 0.31 & -0.02 \\ -0.28 & -0.27 & -0.46 & -0.51 \end{pmatrix}$$

$$V = \begin{pmatrix} -0.23 & -0.60 & 0.59 & -0.19 \\ 0.83 & 0.14 & 0.20 & 0.24 \\ -0.40 & 0.76 & 0.25 & -0.11 \\ 0.11 & -0.10 & -0.61 & -0.64 \\ -0.31 & -0.19 & -0.43 & 0.70 \end{pmatrix}$$

The decomposition of sum of squares of interaction from the principal components is shown in Table 6.6. The sum of squares of each component comprises the total sum of squares of interaction. In this case, all the principal components are significant. In the same table we can see, that the percentage

Table 6.6: Proportion of variance measured by the principal components

	percent	acum	Df	Sum.Sq	Mean.Sq
PC1	66.40	66.40	12.00	161576760.49	13464730.04
PC2	21.60	88.00	10.00	52565081.27	5256508.13
PC3	7.60	95.50	8.00	18447774.50	2305971.81
PC4	4.50	100.00	6.00	10840092.47	1806682.08

of variation captured by each principal component, are 66,40; 21,60;7,60 and 4,50 respectively. We can check that pC1 and pc2 summarizes 88% of the total variability of the interaction.

In next stage, we proceed to the graphical analysis of AMMI modelling, the coordinates of such graphics are in Table 6.4.

Figure 6.2: AMMI biplots for simf2 datasets

The main graphical representations of AMMI modelling of dataset simf 2 are shown in Figure 6.2.

On the left side of the figure, we have the representation of PC1 x PC2 biplot. At the right, we have the picture that contrasts PC1 and yield of genotypes.

These figures serve from the properties of AMMI models to analyze the variability dispersion of genotypes and of several important characteristics about environments.

The variability of main effects of genotype and environments is displaced in horizontal axis of Figure 6.2. PC1xPC2 represents the multiplicative factors related to GEI and is represented in vertical axis.

The interpretation of PC1xPC2 biplot for AMMI allows the evaluation of stability of genotypes over environments, by the evaluation of magnitude of genotypes and environments scores in the interaction axis.

HONGYU (2015), HONGYU *et al.* (2015) and ? state that lower scores imply in more stable genotypes across all environments. Visually, these genotypes will be closer to the origin of x,y coordinates. Genotypes closer to the environment vector are adapted to such environments.

In the case of dataset simf2, the genotype closest to the center is g10, therefore, is the most stable across environments from the set considered.

By the other hand, the graphic shown in Figure 6.2 b, allows to contrast PC1 and yield of grains. By this graph we can see that, g1, g2, g4, g6 and g7, are higher than the mean genotypes.

They should be recommended to be produced but, considering, their stability to specific environments.

In case of genotype g10, in despite of being lower than average it has great stability over environments and its improvement should be strongly recommended via breeding methods, ex given: exploitation of sister lines of its parental inbred lines.

The environment whose behaviour was the best considering these genotypes was the south. Table 6.7 shows the specific adaptation of each genotype to the environments considered.

Table 6.7: Specific adaptations of simf2 given by AMMI analysis of simf2

genotypes	specific env
g1	a1(center)
g2	a2(north)
g3	a3(south)
g4	a4(center 2)
g5	a5 (center 3)
g6	a1 (center 1)
g7	a4(center
g8	a3(south)
g9	a4(center
g10	a2(north)

We also note that given the directions of environment vectors e2 and e3 are divergent (exactly as it should be once they represent north and south regions). By its turn, we can check the formation of three groups made by (e1,e4,e5) e2 and e3.

All these properties shown by AMMI analysis reflected the properties induced in simulations of dataset simf2, described in the Chapter about simulation of dataset, showing that AMMI modelling is indeed efficient considering fixed effects for genotype and environment.

This is the case in final breeding processes, when the breeder deliberately chooses specific genotypes and environments to delineate the METs.

Weighted AMMI analysis of simulated dataset

The analysis of Variance for weighted AMMI is the same as the ANOVA for AMMI and is shown in Table 6.8. The effect of the GEI interaction is present and significant.

The use of weighted AMMI is motivated by situations of heterogeneous environments, this heterogeneity is expressed by the heteroscedasticity among environments. In this situation, the variance of the environment error is given by the mean squares of residuals.

The means and genotypes for weighted AMMI are in Tables 6.9 and 6.10. The values are in a different scale from the original data because of the weighting produced.

Table 6.8: Weighted AMMI ANOVA

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
ENV	4	617225742.25	154306435.56	15.19	0.0000
REP(ENV)	15	152357920.23	10157194.68	45.57	0.0000
GEN	9	104580085.15	11620009.46	52.14	0.0000
ENV:GEN	36	243429708.74	6761936.35	30.34	0.0000
Residuals	135	30089061.71	222881.94		

The single value decomposition for GE weighted AMMI is shown in matrices S, U and V.

$$S = \begin{pmatrix} 0.00413 & 0 & 0 & 0 \\ 0 & 0.00304 & 0 & 0 \\ 0 & 0 & 0.00247 & 0 \\ 0 & 0 & 0 & 0.00112 \end{pmatrix}$$

$$U = \begin{pmatrix} 0.15 & -0.26 & 0.04 & -0.27 \\ -0.38 & 0.38 & -0.02 & 0.06 \\ -0.54 & 0.32 & -0.25 & -0.12 \\ 0.16 & 0.08 & 0.59 & 0.11 \\ 0.30 & 0.28 & -0.26 & 0.40 \\ 0.21 & -0.24 & -0.03 & 0.54 \\ -0.48 & -0.67 & 0.04 & 0.15 \\ 0.26 & 0.05 & -0.44 & -0.02 \\ 0.06 & 0.24 & 0.53 & -0.20 \\ 0.27 & -0.17 & -0.21 & -0.62 \end{pmatrix}$$

$$V = \begin{pmatrix} -0.13 & -0.77 & 0.41 & -0.14 \\ -0.29 & 0.08 & -0.16 & 0.83 \\ -0.19 & 0.63 & 0.55 & -0.26 \\ -0.28 & 0.01 & -0.71 & -0.47 \\ 0.89 & 0.05 & -0.10 & 0.05 \end{pmatrix}$$

The W-AMMI biplots are shown in Figure 6.3. The graph on the left describes the variability captured by first and second principal components and the the figure on the left describes the variability captured by PC1 versus the yield of each genotype.

The understanding of the specific adaptation of each genotype to the different environments is the same as the conclusions made by AMMI models. The main results from W-AMMI for simf2 dataset are the same from AMMI analysis. Thus, w-AMMI captured all the properties of the simulated dataset being therefore a good recommendation of model to study GEI.

Figure 6.3: AMMI biplots for simf2 datasets

GGE Analysis of Simulated dataset

In this section we will cover the four objectives of GGE methodology: investigation of mega environments, genotype evaluation, test environments and comprehension of GEI.

GGE Biplot:

The GGE biplot for simulated data set is present in Figure 6.4: In x axis is shown the scores for First principal component and in y axis the scores of PC2. The graphic comprises 81.58% of the total variability. The genotypes are displayed as g1 to g10 and the five environments from e1 to e5. The interpretations of the graphic respects the rules shown by HONGYU (2015), HONGYU *et al.* (2015) and YAN (2006).

- if the angle of a genotype and a vector of environment is $< 90^\circ$ the genotype is greater than the mean

Figure 6.4: GGE Biplot for simf2

- if the angle is $> 90^\circ$ then it is less than average
- if the angle is close to 90° then the genotype is nearer to the mean

The Figure 6.4 can be used to:

- genotype classification in any environment
- classify environments regarding any given genotype performance

For example, g1 and g6 have obtuse angles considering environment e2 and e4 therefore they had less than average yield.

Mega environment analysis

HONGYU (2015) and HONGYU *et al.* (2015) state that a mega environment is a group of environments or sub regions where a set of genotypes have similar behaviour and adaptability. The biplot "Which won where" is an important tool to make this analysis.

This graph for simf2 data is shown in Figure 6.5. As we can see, there are irregular polygons and lines that arise from the origin of the biplot, intercepting each side of the polygons perpendicularly. The vertices of such polygons are the genotypes markers whose distance is the greatest from the origin. Thus, all the genotypes are inside the polygon HONGYU (2015), HONGYU *et al.* (2015) quoting YAN (2011).

According to HONGYU (2015), a line that crosses perpendicularly one side of the polygon represents an hypothetical environment and the two genotypes that define this side of the polygon represent good yield. The lines that irradiate from origin divide the biplot in sectors and for each sector there is a genotype at the vertex, and such genotype shows the best behaviour for environments in this sector HONGYU (2015), YAN and KANG (2003) and YAN (2011).

From the last paragraph we can see, from Figure 6.5, that g1, g2, g10, g3, g5 and g6 represent the vertices of the polygon.

The GGE biplot which won where what is shown in Figure 6.5.

Mean x stability

The GGE biplot that shows the means x stability is presented in 6.6.

Figure 6.5: Which won where what biplot for simf2

Figura 6.6: Means x stability gge biplot for simf2

HONGYU (2015) states that an ideal genotype should have high mean performance and produce at a range of several environments. The GGE biplot that shows discrimination and representativeness is in Figure 6.7, being a good tool to evaluate these aspects of genotypes.

HONGYU (2015) declares that the arrow that passes through the origin of biplot is the mean average environment (MAE). The arrow points to a greater average performance of genotypes. The line that is perpendicular to MAE indicates less stability over environments. In this graph PVS=1 and the single values are totally directed to genotypic scores HONGYU (2015) and YAN (2011).

Figure 6.6 allows the ranking of genotypes as $g1 > g6 > g2 > g7 > g4 > g9 > g10 > g5 > g8 > g3$.

Mean x stability

HONGYU (2015) states that the Discriminative and representativeness gge biplot allow the discrimination of genotypes and representation of target environments.

HONGYU (2015), HONGYU *et al.* (2015) and YAN (2011) state that when data are centered in environments the length of the vector of an environment is proportional to the standard deviation of

Figure 6.7: Discrimination representative of ggbiplot for simf2

genotypes means, this value being equal to phenotypic variance σ_p in a test environment and can be used as a value of discrimination of the environment.

HONGYU (2015), HONGYU *et al.* (2015) and YAN (2011) argue that environments with shorter vectors are less discriminative, in other words, all genotypes have a tendency of performing similarly and thus little or no information about genotypic differences can be relevant in this environment, therefore, this environment should not be used as tester.

In the case of simf2, all vectors of environments had a length that implies they are good environments to discriminate the genotypes. Environments that have small angles with MAE are better representatives of the target environment.

Regarding environment stability and performance the relationship among genotypes and environments via GGE analysis was the same as in Table 6.7 showing the ability of GGE analysis in detecting the properties of the data simulated with specific properties.

6.3.2 Comparison between AMMI, w-AMMI and GGE biplots

Basically, the results corroborate HONGYU (2015), YAN and KANG (2003), YAN (2011) indicating that the PC1 for GGE biplot is, in general, correlated with genotype factor allowing high precision in visual representation of scores PC1 versus the genotypic factors.

HONGYU (2015), YAN and KANG (2003), YAN (2011) and FRITSCHKE-NETO *et al.* (2010) state that this correlation is biologically plausible once there is a high correlation between PC1 and grains yield.

HONGYU (2015), HONGYU *et al.* (2015) and GAUCH (2006) state that AMMI model have higher accuracy in visualizing patterns in experiments when compared to GGE biplots.

HONGYU (2015), HONGYU *et al.* (2015) and GAUCH (2006) state that one criticism to be made to GGE is that this methodology does not separate the G effect from GEI, but YAN (2011) declares that if the breeder's interest in selecting the genotypes is not based on separation of G and GEI factors GGE biplot can be used.

Now we compare AMMI, W-AMMI and GGE Biplots. The main criteria of comparison will be:

- capacity of detecting the intrinsic properties of the simulated data

- capacity to determine general and specific stability of the genotypes
- information criteria described in methodology

Regarding the first item all the three models returned results that show the properties used in simulation of the dataset. Therefore, we can conclude, that the three models returned the specific adaptation of the genotypes to the environments chosen as specific in simulations answering also to the second item.

By its turn Table 6.11 shows the information criteria used to compare models. The intuition behind the values is that the results of the three models have a great correlation. This high correlation indicates that all the models captured a high value of the variability related to GEI and are correlated among them, indicating that their conclusions are very similar.

Finally, one strong and new result shown by this paper is that AMMI, W-AMMI and GGE analysis can be used independently of the fact that the data studied is normal distributed.

6.4 Conclusions

This article described how a dataset simulated with specific properties of genotypes yield over different environments was used to feed AMMI, W-AMMI and GGE models.

This methodology allowed the determination of the capacity of the models to detect the simulated properties of dataset.

The results indicate that AMMI, W-AMMI and GGE biplot methods captured all the specific properties of data set and modelled efficiently the GEI interaction.

The correlation among the information criteria used to compare such models was high, reinforcing the results that the models should be recommended to evaluate the GEI interaction and to evaluate the specific and broad adaptations of genotypes across a set of environments.

Table 6.9: Means of environments and genotypes Weighted AMMI

	ENV	GEN	Y
1	e1	g1	0.01
2	e1	g10	0.01
3	e1	g2	0.01
4	e1	g3	0.01
5	e1	g4	0.01
6	e1	g5	0.01
7	e1	g6	0.01
8	e1	g7	0.01
9	e1	g8	0.01
10	e1	g9	0.01
11	e2	g1	0.00
12	e2	g10	0.00
13	e2	g2	0.00
14	e2	g3	0.00
15	e2	g4	0.00
16	e2	g5	0.00
17	e2	g6	0.00
18	e2	g7	0.00
19	e2	g8	0.00
20	e2	g9	0.00
21	e3	g1	0.01
22	e3	g10	0.01
23	e3	g2	0.01
24	e3	g3	0.01
25	e3	g4	0.01
26	e3	g5	0.01
27	e3	g6	0.01
28	e3	g7	0.01
29	e3	g8	0.01
30	e3	g9	0.01
31	e4	g1	0.01
32	e4	g10	0.01
33	e4	g2	0.01
34	e4	g3	0.01
35	e4	g4	0.01
36	e4	g5	0.01
37	e4	g6	0.01
38	e4	g7	0.01
39	e4	g8	0.01
40	e4	g9	0.01
41	e5	g1	0.03
42	e5	g10	0.02
43	e5	g2	0.02
44	e5	g3	0.03
45	e5	g4	0.03
46	e5	g5	0.03
47	e5	g6	0.02
48	e5	g7	0.03
49	e5	g8	0.03
50	e5	g9	0.03

Table 6.10: genotypes means weithed AMMI

	type	Y	PC1	PC2	PC3	PC4
g1	GEN	0.01	0.01	-0.01	0.00	-0.01
g10	GEN	0.01	-0.02	0.02	-0.00	0.00
g2	GEN	0.01	-0.03	0.02	-0.01	-0.00
g3	GEN	0.01	0.01	0.00	0.03	0.00
g4	GEN	0.01	0.02	0.02	-0.01	0.01
g5	GEN	0.01	0.01	-0.01	-0.00	0.02
g6	GEN	0.01	-0.03	-0.04	0.00	0.00
g7	GEN	0.01	0.02	0.00	-0.02	-0.00
g8	GEN	0.01	0.00	0.01	0.03	-0.01
g9	GEN	0.01	0.02	-0.01	-0.01	-0.02
e1	ENV	0.01	-0.01	-0.04	0.02	-0.00
e2	ENV	0.00	-0.02	0.00	-0.01	0.03
e3	ENV	0.01	-0.01	0.03	0.03	-0.01
e4	ENV	0.01	-0.02	0.00	-0.04	-0.02
e5	ENV	0.03	0.06	0.00	-0.00	0.00

Tabela 6.11: Information criteria for AMMI, W AMMI and GGE models

PC	AMMI	IR	W-AMMI	IR	GGE	IR
1	66.4	3.32	50.6	2.53	62.94	3.147
2	21.6	1.08	27.4	1.37	19.45	0.9725
3	7.6	0.038	18.2	0.91	13.25	0.6625
4	4.5	0.0225	3.8	0.19	4.36	0.218

References

- CORNELIUS, P. L., 1993 Statistical tests and retention of terms in the additive main effects and multiplicative interaction model for cultivar trials. *Crop Science* **33**: 1186–1193.
- DUARTE, J. and R. VENCovsky, 1999 *Interação Genótipos Ambiente: uma introdução à modelagem AMMI*. SBG, Ribeirão Preto.
- FRITSCHÉ-NETO, R., G. V. MIRANDA, R. O. DELIMA, and H. N. SOUZA, 2010 Factor analysis and sreg gge biplot for the genotype environment interaction stratification in maize. *Cincia Rural* **14**: 1043–1048.
- GAUCH, H., 1992 Statistical analysis of regional yield trias: Ammi analysis of factorial designs. Elsevier .
- GAUCH, H. G., 2006 Statistical analysis of yield trials by ammi and gge. *Crop Science* **46**: 1488–1500.
- GAUCH, H. G., 2013 A simple protocol for ammi analysis of yield trials. *Crop Science* **53**: 1860–1869.
- HONGYU, K., 2015 *Comparação do GGE biplot-ponderado e AMMI-ponderado com outros modelos de interação genótipo $\times$ ambiente*. Ph.D. thesis, USP, Piracicaba.
- HONGYU, K., M. GARCIA-PENA, L. ARAÚJO, and C. DIAS, 2014 Statistical analysis of yield trials by ammi analysis of genotype x environment interaction. *Biometrical Letters* pp. 89–102.
- HONGYU, K., L. SILVA, F. A. OLIVEIRA, D. A. SARTI, and C. T. S. DIAS, 2015 Comparação entre os modelos ammi e gge biplot para dados de ensaios multi-ambientes. *Revista Brasileira de Biometria* .
- ROBERTSON, A., 2014 A weighted ammi algorithm to study genotype-by-environment interaction and qtl-by-environment interaction. *Crop Science* pp. 1555–1570.
- TRETHOWAN, J. G. M. R. S., R.M.; CROSSA, 2001 Relationships among bread wheat international yield testing locations in dry areas. *Crop Science* **41**: 1461–1469.
- VENKOVSKY, R., 1999 *Interação genótipo x ambiente: uma introdução a análise AMMI*. Ph.D. thesis, SBG, Ribeirão Preto: Sociedade Brasileira de Genética.
- YAN, A., W; TINKER, 2006 Biplot analysis of multi environment trial data: principles and applications. *Canadian Journal of Plant Science* **86**: 623–645.
- YAN, W., 2001 A windows application for graphical analysis of multienvironment trial data and other types two-way data. *Agronomy Journal* **93**: 1111–1118.
- YAN, W., 2002 Singular-value partition for biplot analysis of multi-environment trial data. *Agronomy Journal* **94**: 990–996.
- YAN, W., 2011 Gge biplot vs. ammi graphs for genotype-by-environment data analysis. *Journal of the Indian Society of Agricultural Statistics* **65**: 181–193.
- YAN, W. and J. B. HOLLAND, 2010 A heritability-adjusted gge biplot for test environment evaluation. *Crop Science* **171**: 355–369.
- YAN, W. and M. KANG, 2003 *GGE Biplot Analysis: A Graphical Tool for Breeders, Geneticists, and Agronomists*. Boca Raton.

APÊNDICES

Summarized meaning of the text

1. information generation capabilities are attributed to models
 - model derivation or proving methods (derivation of models)
 - checking the behaviour of the models when facing data with specific well known a priori information
2. Phenoma can be modelled via probability distribution
 - model uncertainty
 - use simulation via well known parameters fitted to data
 - the simulated data, with known properties and parameters can be used in simulations
3. Specific experiments can be designed to accomplish specific tasks
 - specific question can be answered via experimentation
4. Plant breeding is a crucial tool for the development of new cultivars
 - it comprises interdisciplinary aspects
 - the products need to have their properties checked
 - an economic specific sector is created by these activities
 - can be addressed in general or via specific aspects such as the ones in 5
5. Genes and environment relate themselves in a dynamic and interactive way
 - genotype characteristic differs in different environments
 - these interactions have impact in products developed in the context shown by 4
6. AMMI, W-AMMI and GGE models
 - can be compared
 - claims to model similar phenomena related such as the one in 5
 - as models their behaviour can be addressed as in contexts

CODES

R coding

R coding will be available at [github](#). any specific doubts can be sent to danilo.estat@gmail.com