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

**Generalizations of the AMMI and GGE models to understand the
interaction between genotypes and environments and between QTL
and environments**

Tatiana Oliveira Gonçalves de Assis

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

**Piracicaba
2020**

Tatiana Oliveira Gonçalves de Assis
Bachelor of Mathematics

**Generalizations of the AMMI and GGE models to understand the
interaction between genotypes and environments and between QTL
and environments**

versão revisada de acordo com a resolução CoPGr 6018 de 2011

Advisor:

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

Co-supervisor:

Prof. Dr. **PAULO CANAS RODRIGUES**

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

Piracicaba
2020

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

Assis, Tatiana Oliveira Gonçalves de

Generalizations of the AMMI and GGE models to understand the interaction between genotypes and environments and between QTL and environments / Tatiana Oliveira Gonçalves de Assis. -- versão revisada de acordo com a resolução CoPGr 6018 de 2011. -- Piracicaba, 2020 .

74 p.

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

1. Interação genótipo x ambiente 2. Ensaios multi-ambientais 3. Interação QTL x ambiente 4. Detecção de QTL . I. Título.

DEDICATION

To my children Lécio Oliveira Gonçalves de Assis and Lavinia Oliveira Gonçalves de Assis, my support at all times.

ACKNOWLEDGEMENTS

I thank to my childrens, Lécio and Lavinia, for all their patience, support and understanding in all moments.

To Frederic, who was the first person to encourage me to start graduate school and always continued to encourage and assist me.

To my parents, for support in those last moments.

To friends and colleagues at Esalq, with whom I learned a lot and provided me with great moments. Especially the girls, Janaína, Laura and Sidicleide, who always received me at their home. And, Janaína Marque e Melo, I still want to register my deep gratitude, because it was the one who did not allow me to give up in the most difficult moments.

To my advisor professor Carlos Tadeu dos Santos Dias, for the opportunity to work with him, for all the teachings and for the confidence in me.

To my co-supervisor, professor Paulo Canas Rodrigues, for his patience, teachings, professionalism and for inspiring me as a professional.

To professor Lúcio Borges de Araújo, who since graduation encouraged me to follow this path and who serves as an example to this day.

To professors Sônia Maria de Stefano Piedade, Lúcio Borges de Araújo, Danilo Augusto Sarti, Priscila Neves Faria, Natalia da Silva Martins Fonseca e José Silvio Govone for the opportunity to evaluate this work. And to professor César Gonçalves de Lima for contributions in the qualification exam.

To all LCE/ESALQ/USP professors and workers, for all the dedication and professionalism.

To all the professors at UFU, where I followed my first academic paths that were essential to complete this stage.

To the CAPES for material and financial support.

SUMMARY

Resumo	6
Abstract	7
1 Introduction	9
References	10
2 Weighted GGE model to study and understand Genotype-by-Environment Interaction and QTL-by-Environment Interaction	13
Abstract	13
2.1 Introduction	13
2.2 Material and Methods	16
2.2.1 Plant materials	16
2.2.2 GGE Analysis	18
2.2.3 Weighted GGE Analysis	19
2.2.4 Weighted GQ Analysis	20
2.2.5 Linear Mixed Models	20
2.3 Results	21
2.3.1 Results for the simulated data	21
2.3.1.1 GGE Analysis	21
2.3.1.2 Weighted GGE Analysis	21
2.3.1.3 GQ and WGQ Analyses	22
2.3.1.4 The 100 Simulated Data Sets and Comparison between Methods	22
2.3.2 Results for the Barley experiments	24
2.3.2.1 GGE Analysis	24
2.3.2.2 Weighted GGE Analysis	25
2.3.2.3 AQ and WAQ Analysis	26
2.4 Conclusion	26
References	28
3 Robust AMMI model in the study of interaction between QTL and environment: A simulation approach	33
Abstract	33
3.1 Introduction	33
3.2 Material and Methods	34
3.2.1 Plant Materials	34
3.2.2 AMMI Model	36
3.2.3 Robust AMMI Model	36
3.2.4 Robust AQ Analysis	38
3.3 Results and Discussion	38
3.3.1 No data contamination	38
3.3.2 Contaminated data	38
3.4 Conclusion	42
References	43
4 Final considerations	45
Appendix	47

RESUMO

Generalizações dos modelos AMMI e GGE para entender a interação entre genótipos e ambientes e entre QTL e ambientes

Em ensaios multi-ambientais é comum que as características genéticas dos cultivares sejam influenciadas pelos ambientes. Desta forma, o estudo de ferramentas que permitam analisar a interação entre genótipos e ambientes e entre QTLs (locus de característica quantitativa) e ambientes tem ganhado cada vez mais espaço entre os pesquisadores desta área. Porém, nem sempre dados coletados são adequados para serem utilizados com modelos já conhecidos, sendo necessário a busca por modelos mais específicos a certas situações. Nessa pesquisa, analisamos situações como dados que apresentam heterogeneidade de variância dos erros ao longo dos ambientes e também dados contaminados, que representam dados com presença de outliers. Nesses casos, modelos já conhecidos na literatura, como o modelo AMMI (modelo de efeito principal aditivo e interação multiplicativa) e o modelo GGE (modelo de efeito principal de genótipo mais interação genótipo $\times$ ambiente), não são indicados. Aqui, verificamos o uso do modelo AMMI robusto e AMMI ponderado na detecção de QTLs e na análise de interações. Também, propomos o modelo GGE ponderado e avaliamos sua eficácia, comparando com outros modelos. Foram usados dois conjunto de dados. O primeiro conjunto de dados são dados simulados de pimentão (*textit Capsicum annuum L.*) de população cruzada usando modelo de crescimento de culturas para relacionar genótipos a fenótipos de maneira não linear e o segundo, dados de cevada de população duplo haplóide Steptoe $\times$ Morex (*Hordeum vulgare L.*).

Palavras-chave: Interação genótipo $\times$ ambiente, Interação QTL $\times$ ambiente, Ensaios multi-ambientais, Detecção de QTL

ABSTRACT

Generalizations of the AMMI and GGE models to understand the interaction between genotypes and environments and between QTL and environments

In multi-environmental trials it is common for the genetic characteristics of the cultivars to be influenced by the environments. Thus, the study of tools that allow the analysis of the interaction between genotypes and environments and between QTLs (quantitative trait loci) and environments has gained more and more space among researchers in this area. However, collected data are not always suitable for use with already known models, making it necessary to search for more specific models for certain situations. In this research, we analyzed situations such as data showing heterogeneity of error variance across environments and also contaminated data, which represent data with outlying observations. In such cases, models already known in the literature, such as the AMMI (additive main-effect and multiplicative interaction) model and the GGE (genotype main-effects + genotype environment interaction) model, are not indicated. Here, we verify the use of the robust AMMI model and weighted AMMI in the detection of QTLs and in the analysis of interactions. We also propose the weighted GGE model and evaluate its effectiveness, comparing it with other models. Two data sets were used. The first data from a simulated pepper (*Capsicum annuum* L.) back cross population using a crop growth model report genotype to phenotype in a nonlinear way, and the second the doubled-haploid Steptoe $\times$ Morex barley (*Hordeum vulgare* L.) population.

Keywords: Genotype-by-environment interaction, Multi-environment trials, QTL by environment interaction, QTL detection

1 INTRODUCTION

The genotypes selection in certain environments is an important objective of genetical enhancement programs. In order to identify superior genotypes in multi-environmental tests, breeders carry out experiments in various locations and years. Thus, it is common to have interaction between genotypes and environments, which is represented by different genotypic responses in environmental variation. In contrast, researchers are increasingly looking for models that represent well the diversity of data in order to analyze these interactions (Rodrigues et al., 2014, 2016; Mohammadi et al., 2017).

The basis of the genotype and environment interaction (GEI) is the quantitative trait locus (QTL) by environment interaction (QEI), which is represented by the different expression of QTLs between environments. To analyze this type of interaction, it is essential to map QTLs related to a specific characteristic of a cultivar. Thus, it is possible that breeders act more effectively in genetic improvement, and one of the main applications of the markers is in the study of the genetic basis, since the markers allow to individualize the phenotypic effect and the monitoring of segregation (Camargo et al., 2001).

Some characteristics that are important in agriculture, such as, for example, resistance to pests, quality or even the yield of a cultivar, are quantitative and controlled by many genes. And it is, exactly these regions within genomes that contain genes associated with a quantitative characteristic, which we call QTL. The identification of these is not possible based on the phenotypic evaluation alone, an important advance being the development of molecular markers in the 1980s, which allowed the beginning of the QTL mapping work. Research indicates that the use of markers will play a key role in improving the production of cultivars by improving the efficiency of breeding programs (Kasha, 1999; Ortiz, 1998).

In order to analyze interactions between genotypes and environments and, between QTL and environments, researchers have presented results of some statistical models (Samonte et al., 2005; Forkman e Piepho, 2006; Dias e Krzanowski, 2006; Yan et al., 2007; Gauch et al., 2008; Rodrigues et al., 2014; Hadasch et al., 2018; Rodrigues, 2018). Among them, the additive main-effect and multiplicative interaction (AMMI) model and the genotype main-effects + genotype environment interaction (GGE) model, which present visual results by biplot graphs, which allows analyzing data from multi-environmental tests (METs) (Casanoves et al., 2005; Dehghani et al., 2006). There is still much research to identify which of these models are better and under what circumstances (Gauch, 2006; Yan et al., 2007), although the two methods provide similar results (Gauch, 2006). However, the GGE model is presented as the most suitable for identifying mega environments, indicating cultivars more adapted and stable to specific environments and also in the selection of representative and discriminative environments (Gauch et al., 2008; Asfaw et al., 2009; Yan, 2011). The AMMI model is indicated as the most efficient in the selection of cultivation sites, that is, identification of superior environmental conditions and also in the identification of genotypes with higher performance (Gauch et al., 2008; Yan, 2011).

However, even the literature showing good results from these models in the analysis of GEI and QEI, there are real situations in which these models are not indicated. For example, it is common, in multi-environmental trials, to identify heterogeneity of error variance across environments. In this case, the AMMI model is not indicated and Rodrigues et al. (2014) proposed the weighted AMMI model (W-AMMI) to solve this problem. Similarly, the GGE model is also not indicated in the cases where the error variance across environments is heterogeneous. Therefore, this research aims to propose the weighted GGE model, to analyze the GEI. The results, for both GEI and QEI, obtained with the W-GGE model are compared with the results from the GGE model and from the AMMI, weighted AMMI and linear mixed model. To conduct this comparison, two data sets are used: data from a simulated pepper (*Capsicum annuum* L.) back cross population using a crop growth model to relate genotypes in a non-linear way to phenotypes, and the doubled haploid Steptoe x Morex barley (*Hordeum vulgare* L.) population. Another common situation is the presence of outliers in the data, and the AMMI model is

also not indicated in these cases. Thus, the robust AMMI model (R-AMMI) (Rodrigues et al., 2016) was proposed. In this research, we used the R-AMMI models to identify their sensitivity in the detection of QTLs and in the analysis of the QEI. The performance of the proposed robust extension of the AMMI model is evaluated through Monte Carlo simulations, in which several contamination schemes are considered. The method is also compared with the results obtained by the classic AMMI and the weighted AMMI model (Rodrigues et al., 2014) for QTL detection and QEI behaviour. The data under consideration includes 100 two-way data tables with 200 genotypes and 12 environments with simulated yield of a pepper (*Capsicum annuum*) population, by using a genotype-to-phenotype crop growth model (Rodrigues et al., 2014, 2020).

The contribution with new proposals of statistical models for the analysis of genotype and environment interactions and QTL and environment interactions, facilitate the variety and efficiency of decision and recommendation to cultivate, as well as acting strongly in the selection of superior cultivars in studies of genetic improvement.

References

- Asfaw, A., Alemayehu, F., Gurum, F., and Atnaf, M. (2009). AMMI and GGE Biplot analysis for matching varieties onto soybean production environments in Ethiopia. *Scientific Research and Essay*, 4:1322–1330.
- Camargo, L. E. A., Nass, L. L., Valois, A. C. C., Melo, I. S., and Valadares-Inglis, M. C. (2001). Marcadores moleculares no melhoramento para resistência a doenças. *Recursos Genéticos e Melhoramento*, 183:995–1010.
- Dias, C. T. S. and Krzanowski, W. J. (2006). Choosing components in the additive main effect and multiplicative interaction (AMMI) models. *Scientia Agricola*, 63:169–175.
- Forkman, J. and Piepho, H. P. (2006). Statistical analysis of yield trials by AMMI and GGE. *Crop Science*, 46:1488–1500.
- Gauch, H. G., Piepho, H. P., and Annicchiarico, P. (2008). Statistical Analysis of Yield Trials by AMMI and GGE: Further Considerations. *Crop Science*, 48:866–889.
- Hadasch, S., Forkman, J., Malik, W. A., and Piepho, H. (2018). Weighted Estimation of AMMI and GGE Models. *Journal of Agricultural Biological and Environmental Statistics*, 23:255–275.
- Kasha, K. J. (1999). Biotechnology and world food supply. *Genome*, 42:642–645.
- Mohammadi, R., Armion, M., Zadhasan, E., Ahmadi, M. M., and A., A. (2017). The use of AMMI model for interpreting genotype $\times$ environment interaction in durum Wheat. *Experimental Agriculture*, 54:670–683.
- Ortiz, R. (1998). Critical role of plant biotechnology for the genetic improvement of food crops: Perspectives for the next millennium. *Electron J Biotechnol*, 1.
- Rodrigues, P. C. (2018). An overview of statistical methods to detect and understand genotype-by-environment interaction and QTL-by-environment interaction. *Biometrical Letters*, 55:123–138.
- Rodrigues, P. C., Heuvelink, E., Marcelis, L. F. M., Chapman, S., and van Eeuwijk, F. (2020). A complex trait with unstable QTLs can follow from component traits with stable QTLs: an illustration by a simulation study in pepper (submitted).

- Rodrigues, P. C., Malosetti, M., Gauch, H. G., and van Eeuwijk, F. A. (2014). A Weighted AMMI Algorithm to Study Genotype-by-Environment Interaction and QTL-by-Environment Interaction. *Crop Science*, 54:1555–1570.
- Rodrigues, P. C., Monteiro, A., and Lourenço, V. M. (2016). A Robust additive main effects and multiplicative interaction model for the analysis of genotype-by-environment data. *Bioinformatics*, 32:58–66.
- Samonte, S. O. P. B., Wilson, L. T., McClung, A. M., and Medley, J. C. (2005). Targeting Cultivars onto Rice Growing Environments Using AMMI and SREG GGE Biplot Analyses. *Crop Science*, 45:2414–2424.
- 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., Kang, M., Ma, B., Woods, S., and Paul, L. C. (2007). GGE Biplot vs. AMMI Analysis of Genotype-by-Environment Data. *Crop Science*, 47:643–655.

2 WEIGHTED GGE MODEL TO STUDY AND UNDERSTAND GENOTYPE-BY-ENVIRONMENT INTERACTION AND QTL-BY-ENVIRONMENT INTERACTION

Abstract

The influence that environments exert on genetic traits of agricultural crops can interfere with productivity. Thus, it is of paramount importance researches aimed at detecting and analysing the interactions between genotypes and environments (GEI) and between quantitative trait locus (QTL) and environment (QEI) in multi-environmental trials (METs). Two of the most fixed effects models used to model and understand GEI are the additive main effect and multiplicative interaction (AMMI) model and the genotype main effects + genotype environment interaction (GGE) model, which have proven to provide good results. However, these models are not adequate when the phenotypic data shows heterogeneous error variance throughout the environments, which is common in METs. For the AMMI model, previous work was conducted in order to overcome this limitation. In this paper, we present a generalization of the GGE model that accounts for data with heterogeneous error variance, the weighted GGE (W-GGE) model. The results, for both GEI and QEI, obtained with the W-GGE model are compared with the results from the GGE model and from the AMMI, weighted AMMI and linear mixed model. To conduct this comparison, two data sets are used: data from a simulated pepper (*Capsicum annuum* L.) back cross population using a crop growth model to relate genotypes in a non-linear way to phenotypes, and the doubled haploid Steptoe x Morex barley (*Hordeum vulgare* L.) population. Considering the number of detected QTLs and the values for their LOD scores the proposed W-GGE model has proved to provide better results than the unweighted competing methods.

Keywords: Genotype-by-environment interaction, Multi-environment trials, QTL by environment interaction, GGE model.

2.1 Introduction

The selection of cultivars is considered a great challenge in a scenario in which there is interaction between genotypes and environments because the presence of interaction alters the genotypic performance between the environments (Mohamed, 2013) and minimizes the magnitude of the association between the phenotypic and genotypic values (Alwala et al., 2010). In Multi-Environment trials (Crossa et al., 1991; Cooper e DeLacy, 1994; Samonte et al., 2005) where the field trials are performed in multiple environments or under different management conditions, it is common to change the genotypic response through the environments. These facts emphasize the importance of studies and researches that address interaction between genotype and environment, thus enabling the identification of cultivars more adapted to certain regions, where they can express their genetic potential. Thus, the study and understanding of interactions is seen as a way to improve complex characteristics in environmental gradients (Annicchiarico, 2002, 2009; Crossa, 1990; Romagosa et al., 2009; van Eeuwijk et al., 2016; Rodrigues, 2018). Statistics, geneticists and breeders have applied several methods to describe and interpret the mean GEI response in agricultural experiments (Araújo et al., 2012; Hadasch et al., 2018). Some of the methods are analysis of variance (ANOVA), linear regression adjustment (Deon et al., 2004), bilinear models and linear-bilinear models.

The additive main effects and multiplicative interaction model (AMMI) (Gauch, 1992) and genotype and genotype by environment interaction models (GGE) (Yan e Kang, 2003) have been extensively used for the analysis of genotype by environment experiments in plant breeding and variety testing

(Malik et al., 2018) and are tools widely used by breeders (Forkman e Piepho, 2006; Hadasch et al., 2017). The AMMI model applies the singular value decomposition (DVS) to the residuals of the analysis of variance (ANOVA) that considers genotypic and environmental main effect. The GGE model differs from the AMMI model in the initial stage of the analysis, because it applies the SVD to the residuals of the ANOVA that considers only genotypic main effects. Recent revisions have presented both models for genotype $\times$ environment interaction analysis (Samonte et al., 2005; Forkman e Piepho, 2006; Dias e Krzanowski, 2006; Yan et al., 2007; Gauch et al., 2008; Rodrigues et al., 2014; Hadasch et al., 2018; Rodrigues, 2018). Pereira et al. (2012) and Rodrigues et al. (2011) shown the usefulness of the AMMI model in comparison with regression related techniques, and Forkman e Piepho (2006), Gauch et al. (2008), Forkman e Piepho (2014) and Hongyu et al. (2015) provide comparison between the AMMI and the GGE models. Further developments and generalizations of the AMMI model have also been proposed (Paderewski e Rodrigues, 2018; Rodrigues, 2018; Rodrigues et al., 2016). The GGE model is presented as more suitable for the identification of mega-environments, selection of representative environments and discriminative and indication of cultivars more adapted and stable to specific environments (Yan et al., 2007; Gauch et al., 2008; Asfaw et al., 2009; Yan, 2011). The AMMI model can be used efficiently in the identification of superior environmental conditions for the farm (selection of cultivation sites) and genotypes of higher average performance (Gauch et al., 2008; Yan, 2011).

The AMMI model is not indicated for situations in which there is heterogeneity of error variance throughout the environments, which is expected when the GEI is large (Crossa e Cornelius, 1997; Edwards e Jannink, 2006). A generalization that takes into account the heterogeneity of error variance, the weighted AMMI (W-AMMI) model, was presented by Rodrigues et al., (2014), being another alternative for non-replicated data presented by Assis et al. (2018). This generalization outperforms the AMMI model and offers a reasonable approximation to the mixed model for GEI, which is generally considered an appropriate approach for these cases (Rodrigues et al., 2014). Similarly, when considering the GGE model, to analyse two-way data sets that have heterogeneity of error variance throughout the environments, can result in misleading interpretations.

Quantitative trait loci (QTLs) are regions of the genome responsible for the expression of phenotypic traits that have continuous distribution, in which these phenotypic traits can be, for example, height and weight of both animals and plants/fruits, as well as yield of a given cultivar. That is, it is defined as regions of the genome that are responsible for the variation of a quantitative trait (Doerge, 2002). Advances in the characterization of quantitative patterns that created opportunities to select QTLs were initiated by the development of molecular markers in the 1980 decade (Collard et al., 2005). In order to identify QTLs potentially involved in controlling GEI, that is, to help identifying genotypes specifically adapted to each environment, the use of the AMMI methodology has been used to study the QTL-by-environment interaction (QEI) (Romagosa et al., 1996; Gauch et al., 2011). Aiming to provide successful understanding of QEI and to make predictions of QTLs for new environments, the proposed strategy was to apply the AMMI model to phenotypic data to gain accuracy a strength from other environments and, consequently to increase the logarithmic probability (LOD) scores for the QTLs. Although the environments are sorted by parameters of the AMMI model, which assume information of the GEI, it helps revealing consistent patterns and systematic tendencies that often have ecological or biological interpretations (Rodrigues et al., 2014).

QTL mapping is defined as the process of constructing linkage maps and conducting QTL analysis to identify genomic regions associated with traits (McCouch e Doerge, 1995; Mohan et al., 1997; Paterson, 1996a,b). Thus, mapping a QTL means identifying its position in the genome and estimating its genetic effects, such as: the additive effect, dominance effect and other effects present in the adopted model (Toledo et al., 2008). The efficient detection of QTLs depends on several factors, such as its number, magnitude of its effects, heritability characteristics, and interactions between genes, frequency

of recombination between QTLs and the types of markers and degree of saturation of the map (Sergio et al., 1993; Young, 1996; Liu, 1998). The main objectives of QTLs mapping are defined as searching for QTLs detection, searching for trusted regions for QTLs locations, and estimating QTLs effects.

Although the interaction between QTL and environment has been the focus of many research articles (Romagosa et al., 1996; Crossa et al., 1999), many of the QTL studies treated interaction phenomena as a nuisance, focusing mainly on the main effects of QTL. Thus, potentially valuable QEI standards may have been neglected due to the absence of methods that would allow meaningful explorations and useful descriptions of QEI (Yan e Tinker, 2005).

The genetic basis of GEI and QEI originates in the differential expression of genes throughout the environments and can occur in three ways. One way is that a QTL can express itself in an environment without expressing itself in another, causing it to be detected inconsistently across different environments. Another way would be that a QTL can express itself strongly in a given environment, but weakly in another, causing it to present a variation in its effect throughout the environments. Finally, the case in which a QTL expresses itself in a many differentiated way in certain environments and has an opposite effect in others (Vieira et al., 2006). A good understanding of the QEI allows researchers to select better genotypes for different environmental gradients and consequently to improve crops and their productions, based on their climate and soil characteristics.

QTL analysis is based on the principle of detection an association between phenotype and genetic markers (Collard et al., 2005). Markers are used to partition the mapping population into different genotypic groups based on the presence or absence of a particular marker and to determine whether significant differences exist between groups with respect to the trait being measured (Sergio et al., 1993; Young, 1996). A significant difference between phenotypic means of the groups, depending on the markers system and type of population, indicates that the marker being used to partition the mapping population is linked to a QTL controlling the trait.

Gauch et al. (2011) suggest the AQ analysis, which is a new approach for the detection and understanding of GEI, in which the QTL scans are made based on AMMI predictions rather than QTL significant scans in the GEI. Similarly, we are proposing that QTL scans to be made on the basis of predictions via the GGE model. The weighted AQ analysis, or WAQ, consists of performing QTLs exams in the weighted AMMI models (Rodrigues et al., 2014). This approach can potentially improve QTLs detection (Gauch et al., 2011). Environments can be ordered by AMMI or W-AMMI parameters that summarize GEI and QEI with information to reveal consistent patterns and systematic trends that can often be explained in terms of environmental conditions (Gauch, 1992; Gauch et al., 2011).

This paper presents a generalization of the GGE model that accounts for heterogeneity of error variances across environments, the weighted GGE model, or W-GGE model. Moreover, we adapt the AQ and WAQ methodologies from (Gauch et al., 2011), that considered the AMMI model, to the GGE model and to the W-GGE model. Here, we will also order the environments by GGE and W-GGE parameters to better understand patters and trends in terms of environmental conditions. This will allow researchers to have a competing methodology for the case of heterogeneity of variance. The comparison between the results from considering each of the four models (AMMI, W-AMMI, GGE, W-GGE, and linear mixed model) are done for two data sets. The first data includes 100 two way data tables of yield simulated for a backcross pepper (*Capsicum annuum*) population using a crop growth (physiological) genotype-to-phenotype model (Rodrigues, 2012; Rodrigues et al., 2020). The motivation for using this crop growth model which non-linearly transforms genotypic information into phenotypic information is the possibility of simulating a biologically realistic data set, while controlling the underlying genetic architecture (Chenu et al., 2009; Tardieu e Tuberosa, 2010; van Eeuwijk et al., 2010). The second data set includes the yield for the well known Steptoe x Morex barley (*Hordeum vulgare* L.) population, originating from the North American Barley Genome Mapping Project (Hayes et al., 1993).

2.2 Material and Methods

2.2.1 Plant materials

The first set of data in this study consisted of 100 two-way data tables with 200 genotypes (G) and 12 environments (E) with yield. These data sets were simulated based on a crop growth model and assuming that the final yield results from the sum of the signal and the noise. The signal of genotype i in the environment j was simulated from a eco-physiological genotype-phenotype crop growth model (CGM) for pepper (Rodrigues, 2012; Rodrigues et al., 2020), and is a function of parameters that ultimately depend on genotype or environment, but never on the genotype and environment simultaneously.

Table 2.1: Description of the environments considered in the genotype-to-phenotype crop growth model in Eq. [2.1]. The countries were chosen to represent different environmental and practical conditions (Rodrigues, 2012; Rodrigues et al., 2020). Radiation has two levels (years) based on historical data. Temperature contains three levels of daily average temperature. The last column gives the error variance for the 100 simulated data sets. The heritability for the environments was set to be $h^2 = 0.5$.

Environment	Country	Radiation	Temperature	Error Variance
NL1-15	Netherlands	Lower	15 ⁰ C	22.4
NL1-20	Netherlands	Lower	20 ⁰ C	38.01
NL1-25	Netherlands	Lower	25 ⁰ C	29.82
NL5-15	Netherlands	Higher	15 ⁰ C	38.27
NL5-20	Netherlands	Higher	20 ⁰ C	72.81
NL5-25	Netherlands	Higher	25 ⁰ C	57.73
SP1-15	Spain	Lower	15 ⁰ C	11.85
SP1-20	Spain	Lower	20 ⁰ C	17.84
SP1-25	Spain	Lower	25 ⁰ C	17.71
SP5-15	Spain	Higher	15 ⁰ C	14.92
SP5-20	Spain	Higher	20 ⁰ C	22.61
SP5-25	Spain	Higher	25 ⁰ C	26.94

The model can be written as follows (Rodrigues, 2012; Rodrigues et al., 2020):

$$\begin{aligned}
Yield_{i,j} &= Signal_{i\bar{V}_j} + Noise_{i,j} \\
&= (Radiation\ intercepted \times Conversion\ efficiency \times Partitioning\ to\ yield)_{i\bar{V}_j} + Noise_{i,j} \\
&= \left\{ \sum_{t=t_0}^{t_f} [1 - \exp(-k_i \times LAI_{i,j,t})]_{j,t} \right\} \times \{LUE_{i,j}\} \times \{FTF_i \times [1 - W_i \times (T_j - T_{FTF})]FDMC_i\} \\
&\quad + \varepsilon_{i,j} \quad (2.1)
\end{aligned}$$

The radiation intercepted depends on:

- (i) light extinction coefficient (K_i);
- (ii) leaf area index of genotype i in environment j on day t , this is, $LAI_{i,j,t} = [a + B_i(T_j - T_{base}) \times (t - t_0)] \times Sd_j$, being a and B_j , the constant intercept and the genotype specific slope in a regression of the leaf area per stem (m^2) as a function of the temperature sum ($^{\circ}C$), T_j is the average temperature per environment j , T_{base} is the base temperature, t represents the t -th day of the growing season ($t = t_0$ s the first day of flowering) and Sd_j is the density of the stem in the environment j ;
- (iii) $I_{j,t} = RAD_{j,t} \times F_{PAR} \times Tr_j$ is the photosynthetic active radiation incident in the harvest by environment j on day t , that is, is the productt of:
 1. global radiation of day t in the environment j ($RAD_{j,t}$);

2. fraction of photosynthetic radiation (PAR) activity in global radiation (F_{PAR});
3. transmissivity of greenhouse effect in the environment j (Tr_j).

The sustained t_0 and t_f represent the beginning and end of the growing season, in days. The light use efficiency for genotype i in the environment j ($LUE_{i,j}$) depends on the temperature given by the parameter Z_i - linear slope reduction for LUE at temperatures below $20^\circ C$. The partitioning to yield, as described on the right side of the equation, is defined by a combination of the partitioned dry weight fraction of the fruits (FTF_i), fruit dry matter content ($FDMC_i$) and slope of the linear reduction in the harvest index with temperatures above $15^\circ C$ (W_i). Detailed information about his simulation model and parameter specification can be found in Rodrigues (2012) and Rodrigues (2018).

The Model (2.1) is a function of seven physiological parameters that depend on the genotype. Three environmental variables are considered (Table 3.1):

- (i) temperature with three levels: low ($15^\circ C$), medium ($20^\circ C$) and high ($25^\circ C$);
- (ii) radiation, with two levels/years, based on the history of meteorological data by country;
- (iii) country, with two levels: Spain and the Netherlands.

Each of the seven physiological parameters was simulated as a sum of the number of QTL effects (Rodrigues, 2012; Rodrigues et al., 2020) plus specific residual effect of the physiological parameter. This simulation is inspired by a system of 200 genotypes of pepper, characterized by 237 markers covering all 12 second chromosomes (Barchi et al., 2007). Eleven potential QTLs were allocated along the 12 chromosomes of the pepper genome. The simulations were performed using the `sim.cross` function of the QTL package (Broman e Sen, 2009) of the R software. This function allows simulating QTL additive for the yield components and controlling their heritability's. The main motivation for the use of a nonlinear genotype-phenotype physiological model is to ensure that the simulated data originate from a model in which we have complete information about the underlying genetic architecture and its translation into the phenotype (Rodrigues et al., 2020). The error of the equation (1), $\varepsilon_{i,j}$ simulated from the multivariate Gaussian distribution, with a vector of zero averages the matrix of variance and covariance $\sigma_{\varepsilon_j}^2 \mathbf{I}_n$, being $\sigma_{\varepsilon_j}^2$, $j = 1, \dots, J$ depend on the environment and the heritability chosen (h^2), i.e.:

$$\sigma_{\varepsilon_j}^2 = \frac{1 - h^2}{h^2} \sigma_{g_j}^2$$

Being $\sigma_{g_j}^2$ and $\sigma_{\varepsilon_j}^2$ genetic variances and error for the environment, $j, j = 1, \dots, J$, the final yield data is the result of the sum of the signal and the noise components, as in the equation (2.1). This simulation was repeated 100 times resulting in 100 two-way data tables with 200 genotypes per 12 environments. This crop growth model allows to simulate phenotypic data based on component traits with QTL information, which can then be used to access whether the original QTLs are detected in the phenotypic data. Complete details on the simulation model can be found in Rodrigues (2012) and in Rodrigues (2018).

The second data set in this study is a subset of the grain production data Steptoe x Morex (SxM) produced by the North American Barley Genome Mapping Project (Hayes et al., 1993). This subset consists of 150 genotypes conducted in 8 environments. The genotypes were characterized by 223 markers covering seven chromosomes and a subset of 116 markers was used in this study. In Rodrigues et al. (2014) the trials without replications were removed, resulting in a set of trials that were fully replicated, in a randomized block design with two blocks (trials of 1992), or partially replicated, with a complete block and a Second block containing 50 genotypes (1991 assays). The details of the eight environments are presented in (Table 2.2).

Table 2.2: The eight environments used in the Steptoe x Morex analysis, the environments are either fully replicated randomized complete block designs with two blocks, or partially replicated block designs with 50 genotypes replicated in a second block, the trials conducted in 1991 have a full replicate (block) and a second one containing only 50 genotypes, the trials conducted in 1992 have two complete replications. The last column presents the error variances per environment.

Environment†	Full replication	Error variance
ID91	No	0.62
ID92	Yes	0.42
MAN92	Yes	0.20
MIN92	Yes	0.37
MTd92	Yes	0.30
MTi91	No	0.24
MTi92	Yes	0.31
WA91	No	0.71

†ID91, Idaho 1991; ID92, Idaho 1992; MAN92, Manitoba 1992; MIN92, Minnesota 1992; MTd92, Montana dryland 1992; MTi91, Montana irrigated 1991; MTi92, Montana irrigated 1992; WA91 Washington 1991.

2.2.2 GGE Analysis

The GGE model represents the genotype main effects plus genotype and environment interaction. The GGE model is a modification in the conventional AMMI model, which was proposed by Yan et al. (2000). It groups the effect of genotype, which is an additive effect in the AMMI analysis, with the GEI, multiplicative effect, and submits these effects to an analysis that follows the SREG (Sites Regression) model. The SREG model is a multiplicative regression model for locations. The main advantage of this technique in relation to the AMMI analysis is that the GGE method always explains an intermediate portion of the sum of squares of genotypes plus GEI, in relation to the AMMI models. Similarly to the AMMI model, the GGE model can be defined as a method based on singular value decomposition (or principal component analysis) that explores multi-environmental field trials. Also, the relationships between environments, genotypes and genotype-by-environment interaction can be visualized through the biplot (Gariel, 1971). Three important aspects can be visualized in the GGE model (Yan, 2001): (i) the structure of GEI, allowing the grouping of genotypes and environments with similar behaviours and showing the genotype with the highest potential and its identification in each subgroup of environments; (ii) the interrelationship between the environments, facilitating the identification of the best environment in the evaluation of the genotypes and indicating which environment may be less favourable; and (iii) the interrelationship between genotypes, facilitating the comparison of genotypes and the ordering for the parameters of yield and stability. The GGE model can be written as follows:

$$Y_{ij} = \mu + \beta_j + \sum_{n=1}^N \lambda_n \gamma_{ni} \delta_{nj} + \varepsilon_{ij}, \quad (2.2)$$

where: Y_{ij} is the observed phenotypic value (e.g. yield) of the i -th genotype and in the j -th environment; μ is the grand mean; β_j is the effect of the j -th environment; λ_n is the singular value for interaction principal component (PC) on the n axis; γ_{ni} and δ_{nj} are the PC scores (left and right singular vectors) of the i -th genotype and j -th environment on the n axis; $N \leq \min(I - 1, J)$, with I the number of genotypes and J the number of environments (Gauch et al., 2008; Malik et al., 2018) ; and ε_{ijk} is the experimental error.

Considering the matrix formulation of the equation (2.2) can be written as:

$$\mathbf{Y} = \mathbf{1}_I \mathbf{1}_J^T \mu + \mathbf{1}_I \beta_J^T + \mathbf{U} \mathbf{D} \mathbf{V}^T + \varepsilon. \quad (2.3)$$

and the matrix (G+GEI) Θ can be written as

$$\Theta = \mathbf{U} \mathbf{D} \mathbf{V}^T$$

and can be decomposed by the product of the matrix $\mathbf{U}$, of the left singular vectors, that has I rows and r columns $r \leq \min(I, J)$ with the matrix transposed $\mathbf{V}$, of the right singular vectors, that has r rows e J columns, and with matrix $\mathbf{D}$ which is a diagonal matrix containing the r singular values (Yan e Kang, 2003; Yan e Tinker, 2006).

The main difference between the AMMI and GGE models is that the GGE model does not subtract the effects of the genotypes before performing DVS. Consequently, the multiplicative parameters resulting from DVS have different values since this decomposition is applied to different matrices.

2.2.3 Weighted GGE Analysis

In this section we propose a generalization of the GGE model, the W-GGE model or the weighted GGE model, which allows to consider the heterogeneity of the error variance throughout the environments, besides allowing the presence of missing values in the original two-way data table.

Our proposal consists in the substitution of SVD in equation (2.3) by a weighted SVD (Rodrigues et al., 2014) allowing the generalization of the GGE model in case there is heterogeneity in the variance of the error throughout the environments. Thus, using the expectation-maximization (EM) algorithm iteratively will be obtained the values for:

$$\mathbf{X}^{(t+1)} = SVD(\mathbf{W} \odot \Theta + (\mathbf{1} - \mathbf{W}) \odot \mathbf{X}^{(t)}), \quad (2.4)$$

in which $\mathbf{W}$ is a matrix of weights ($I \times J$), $W_{i,j}$, $0 \leq W_{i,j} \leq 1$, $\mathbf{1}$ matrix of ones in all positions, $\odot$ is the product of Hadamard and t is the number of the iteration. The result of this algorithm are the matrices $\mathbf{U}_N$, $\mathbf{D}_N$ and $\mathbf{V}_N$, so that $\tilde{\Theta} \approx \mathbf{U}_N \mathbf{D}_N \mathbf{V}_N^T$ being N the rank of the approximation, which must be decided before the beginning of the weighted estimation. Applying the weighted SVD (2.4) to the matrix $\tilde{\Theta}$ and replacing in (2.3), results in the W-GGE model. This generalisation can be applied to all data sets in which the GGE model is usually used.

The weight matrix $\mathbf{W}$ can be obtained by the inverse of the error variance for each environment. For the simulated pepper data, replications can be obtained by changing the seed value when running the crop growth model. For the Steptoe x Morex data, the weighted matrix $\mathbf{W}$ can be obtained as follows. When dealing with partial replicated data, cell means based on replicated observations will bring more information to the model than those based on single observations (Rodrigues et al., 2014). The scheme of weights should reflect the number of replications per cell. The matrix with weights W , for this new scheme of weights, can be calculated from the Hadamard product of two matrices.

- (i) a matrix in which the entries are column wise constant as the inverse of the error variance
- (ii) a matrix with the proportion of replications per cell.

$$\begin{bmatrix} \frac{1\sigma_1^2}{m} & \frac{1\sigma_2^2}{m} & \dots & \frac{1\sigma_J^2}{m} \\ \frac{1\sigma_1^2}{m} & \frac{1\sigma_2^2}{m} & \dots & \frac{1\sigma_J^2}{m} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1\sigma_I^2}{m} & \frac{1\sigma_2^2}{m} & \dots & \frac{1\sigma_J^2}{m} \end{bmatrix} \odot \begin{bmatrix} \frac{Nrep_{1,1}}{Nrep} & \frac{Nrep_{1,2}}{Nrep} & \dots & \frac{Nrep_{1,J}}{Nrep} \\ \frac{Nrep_{2,1}}{Nrep} & \frac{Nrep_{2,2}}{Nrep} & \dots & \frac{Nrep_{2,J}}{Nrep} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{Nrep_{I,1}}{Nrep} & \frac{Nrep_{I,2}}{Nrep} & \dots & \frac{Nrep_{I,J}}{Nrep} \end{bmatrix} \quad (2.5)$$

where I is the number of genotypes, J the number of environments, $m = \max(1\sigma_{\epsilon_j}^2, \sigma_{\epsilon_j}^2)$, $j = 1, \dots, J$ is the error variance for environment j , $Nrep_{i,j}$, $i = 1, \dots, I, j = 1, \dots, J$, the number of replications for genotype i in environment j , and $Nrep$ is the maximum number of replications in the data set.

2.2.4 Weighted GQ Analysis

In order to detect and understand the QEI, Gauch et al. (2011) proposed the AQ analysis, an approach aimed at better detecting and understanding the QEI. In this approach, QTL scans are not made directly in the means of genotype x environment interaction, but based on the predictions of the AMMI model. The GQ analysis here considers a similar framework for the GGE model, i.e. the QTL scans are obtained from the GGE predicted values. Moreover, in the weighted GQ methodology, the weighted AQ analysis is performed and then the QTLs scans are obtained based on the values W-GGE predicted values. Environments can be sorted by GGE and W-GGE parameters that summarize the interaction information between genotypes and environments, helping identifying potential environmental trends and patterns for QTLs and environments.

2.2.5 Linear Mixed Models

In order to compare the results, we also used the QTL analysis based on linear mixed model, which can be written as (Boer et al., 2007):

$$Y_{ij} = \mu + \beta_j + \sum_{q=1}^Q x_{i,q} \alpha_{q,j} + \varepsilon_{ij}, \quad (2.6)$$

with μ an intercept; β_j an environmental main effect; $x_{i,q}$ a genetic predictor related to the probability that for a genotype i at a particular genomic position q , the alleles are coming from one or the other parent; and value for interaction principal component (PC) on the n axis; $\alpha_{q,j}$ the allele substitution value for a QTL q in environment j . The allele substitution values in Eq. [2.6] are dependent on the environment and are the sum of a QTL main effect and a QEI. The residual term $\varepsilon_{i,j}$, feeds on polygenic effects, where the variance–covariance matrix for this residual should allow for heterogeneity of genetic variances and correlations across environments. When the number of environments becomes larger, the number of genetic variances and co-variances that needs to be estimated rapidly grows. To reduce the number of parameters for estimation while retaining flexibility for modelling heterogeneity of variances and co-variances, various types of more-parsimonious models have been proposed, with the so-called factor analytic (FA) models being the most popular ones (Rodrigues et al., 2014). The environmental scores in the FA model are mixed model analogues of the environmental scores in AMMI models. Also, in FA models, more than one multiplicative term can be included. The environmental scores for the FA variance–covariance model for the residual $\varepsilon_{i,j}$ Model [2.6] are more comparable to the environmental scores in the so-called GGE models (Yan e Kang, 2003), with GGE standing for genotypic main effects and GEI, than with the environmental scores in AMMI models, because in Eq. [2.6] we include the environmental main effect as in GGE models, but not both genotypic and environmental main effects as in AMMI models (Rodrigues et al., 2014).

In (Boer et al., 2007) the input for the mixed model QTL analysis consists of the genotype-by-environment means and corresponding weights, defined in the same way that was used for W-GGE model and weighted GQ analysis. The mixed model QTL mapping procedure initially scans the genome fitting models containing a single QTL only ($Q = 1$ in Model [2.6]), but this single QTL has environment-specific effects. Typically, a simple interval mapping scan is followed by one or more composite interval mapping scans. A final multi-QTL model is selected by backward elimination from a model containing all identified QTLs of the last composite interval mapping scan (Rodrigues et al., 2014). The threshold

level for the Wald test for QTL detection is based on a multiple testing correction as developed by (Liu et al., 2005). (Boer et al., 2007), (van Eeuwijk et al., 2010), and (Malosetti et al., 2013) forth bring more information about it.

2.3 Results

2.3.1 Results for the simulated data

2.3.1.1 GGE Analysis

Table 3 gives the ANOVA for the GGE model, based on one randomly-chosen realization of a genotype-by-environment two-way data table. The genotype and GEI account for 23,8% and 76,15% of the treatments sum of SS. The GGE biplot of one of the 100 simulated data sets, chosen randomly, is depicted in Figure 2 (left), with the first and second principal components explaining 73.21% and 6.58% of the sum of squares of the interaction, respectively. In this figure, the environments with the highest error variance (NL5-20 and NL5-25; Table 1) are more distant from the origin, showing extreme behaviour when compared with the other environments. Later the heterogeneity of error variance should be incorporated in a weighted analysis of the GxE data to produce a potentially more reliable and interpretable result in terms of visualization and QTL detection. This can be achieved by giving smaller weights to the environments with higher error variance as discussed previously.

Table 2.3: Analysis of variance of the genotype main effects + genotype environment interaction model 2.2 for the pepper population, simulated from the genotype-phenotype model with seven physiological parameters.

Source†	df	SS‡	MS§
Total	2399	335809.000	139.9
Genotypes	199	80076.000	402.4
GEI	2200	255733.000	116.2
PC1	210	187238.035	891.6
PC2	208	16832.574	80.9
PC3	206	11196.748	54.3
PC4	204	9311.307	45.6
PC5	202	8115.639	40.2
PC6-PC12	1070	23038.420	117.6

† GEI, genotype-by-environment interaction; IPC, interaction principal component; ‡ SS, sums of squares; § MS, mean squares.

2.3.1.2 Weighted GGE Analysis

The weighted GGE analysis was used here with the objective of avoiding that environments with high error variance are considered as outliers (Gauch et al., 2011), or that influence too much the results. The weight for each environment was determined by the inverse of the error variance (Table 2.1). In Figure 2.1 (right) depicts the W-GGE biplot, where a clear pattern in the environments is observed: the lower part of the biplot shows environments with temperatures of 25°C, the upper right side shows environments with temperatures of 15°C and the upper left side the environments with temperatures of 20°C.

Figure 2.1: Genotype main effects + genotype environment interaction model (GGE) (left) and weighted GGE (right) biplots based on simulate data. The abscissa represents the first multiplicative term and the ordinate represents the second. The open dots represent the 200 genotypes and the codes for the 8 environments are defined in Table 3.1.

2.3.1.3 GQ and WGQ Analyses

In this paper, as an adaptation of the AQ analysis presented by (Rodrigues et al., 2014), is obtained by adjusting the GGE model to the GxE two-way table of means, followed by the QTL scan in the values predicted by the GGE model. Also, the weighted adaptation of the AQ analysis that accounts for heterogeneity in the error variance is considered. These analyses are useful for datasets whose environments have high heterogeneity of error variance.

Figure 2.2 shows the QTL scan for six of the environments of the complex trait yield simulated from the physiological genotype-to-phenotype model with seven physiological parameters (Rodrigues, 2012; Rodrigues et al., 2020). The QTL scans based on the original data (dotted grey line), the QTL scans based on the GGE predicted values (grey lines), and the QTL scans based on the W-GGE predicted values (black lines), are shown. The six environments were chosen to represent the three temperature levels with the smallest and the highest variance of the error. Figure S1 shows the QTL scans for all 12 environments.

In this simulation study, a QTL is considered to be truly detected when a peak was identified in a 20 cM interval centred in its true location, i.e. when the QTL was detected less than 10 cM away from its true location. The thresholds for QTL detection were obtained based on a permutation test with 1000 permutations, at a significance level of 0.05.

The improvement in the number of detected QTLs and on the magnitude of LOD scores is clear when using the W-GGE predicted values instead of the GGE predicted values and in the original data. This pattern is visible in Fig. 2.1 and in Fig. A.1. This trend is more visible in the environments with higher error variance (environments on the left hand side of Figure 2.1).

2.3.1.4 The 100 Simulated Data Sets and Comparison between Methods

As reported by (Rodrigues et al., 2014), there is a noticeable improvement in the QTL analysis when the values predicted by the W-AMMI model are used, when compared to the use of the values predicted by the AMMI model. This improvement is both in number and accuracy in QTLs detected as well as in higher values of LOD scores. This is particularly noticeable when there is heterogeneity in the error variance along environments.

When analysing the results presented in Rodrigues et al. (2014) and in the QTL scans in Figure

Figure 2.2: QTL scans for six environments of the yield data for the pepper population, simulated from the genotype-phenotype model with seven physiological parameters. Each Line of charts represents a different level of temperature. The plots on the left hand side correspond to the lowest error variances of the simulated data and those on the right hand side, the larger ones. The dotted segments represent the analyses for the actual data the grey lines represent the QTL scans on the GGE predicted values, and the black lines the QTL scans on the W-GGE predicted values. The horizontal lines correspond to the thresholds based on a permutation test with 1000 permutations, at a significance level of 0.05. These analyses are based on random choice of a set of data among the 100 simulations. The codes for the captions of the individual scans are described in Table 3.1

2.2 in detail, it is observed that when using the original data, 15 QTLs are detected; when using the predicted values of the AMMI model, 23 QTLs are detected; when using the predicted values of the weighted AMMI model, 29 QTLs are detected (Rodrigues et al., 2014); when using the predicted values of the GGE model, 31 QTLs are detected; and when using the predicted values of the weighted GGE model, 36 QTLs are detected. There is also a relative increase in the LOD scores when the W-GGE predicted values are used, in comparison with the W-AMMI predicted values, in the QTLs detected in chromosomes 5 and 6 of the environments with the highest variance of the error (left hand side of Figure 2.2).

Figure 2.3 shows the boxplots, based on the 100 simulated two-way data tables, that summarize the number of QTLs detected by each of the methods (original data, predicted values by the GGE model, predicted values by the W-GGE model, predicted values by the AMMI model, predicted values by the W-AMMI model, and predicted values by the mixed linear model), by type of parameter of the genotype-to-phenotype crop growth model (1) (left hand side of Figure 2.3) and by type of environmental cluster (right hand side of Figure 3). Figure 2.4 shows the boxplots for the number of QTLs detected by each of the six methods for each environment.

When analysing Figures 2.3 and 2.4, it can be verified that the W-AMMI model outperforms the AMMI model, in terms of number and accuracy in the detection of QTLs (Figures 3 and 4). Similarly, the W-GGE model outperforms the GGE model (Figures 2.3 and 2.4).

The WGQ analysis and the QTL mixed model framework seem to be the best options in the presence of heterogeneity of error variance across environments. However, a few false positives were detected by the QTL mixed model framework because only seven or eight QTLs are expected per environment (Rodrigues et al., 2014, 2020), and in some environments, more than eight QTLs have been detected.

Figure 2.3: Summary of the number of detected QTLs for the GxE means (yellow), predicted values by the GGE model (blue), predicted values by the W-GGE model (purple), predicted values by the AMMI model (orange), predicted values by the W-AMMI model (red), and predicted values by the linear mixed models (brown). The plot on the left hand side shows the boxplots for the number of QTLs detected by model parameters, and the plot on the right hand side shows the number of QTLs detected by environmental cluster, described in Rodrigues et al. (2014). These values are for QTLs detected when considering an interval of 20 cM centred on the exact position of the simulated QTL. These plots are for a heritability of 0.5 in all environments.

2.3.2 Results for the Barley experiments

2.3.2.1 GGE Analysis

Table 2.4 gives the ANOVA for the GGE model, based on one randomly-chosen 395 realization of a genotype-by-environment two-way data table. The genotype and GEI account for 23,8% and 76,15% of the treatments sum of SS. The GGE biplot of one of the 100 simulated data sets, chosen randomly, is depicted in Figure 2 (left), with the first and second principal components explaining 73.21% and 6.58% of the sum of squares of the interaction, respectively. In this figure, the environments with the highest error variance (NL5-20 and NL5-25; Table 1) are more distant from the origin, showing extreme behaviour when compared with the other environments. Later the heterogeneity of error variance should be incorporated in a weighted analysis of the GxE data to produce a potentially more reliable and interpretable result in terms of visualization and QTL detection. This can be achieved by giving smaller weights to the environments with higher error variance as discussed previously.

The GGE biplot is depicted in Figure 2.5 (left). The environments with a larger error variation (Table 2.2) tend to be placed away from the origin (Rodrigues et al., 2014).

Figure 2.4: Number of detected QTLs per environment for an expected maximum of 7 (environments with temperature of 15°C or 20°C) or 8 (environments with temperature of 25°C). The boxplots are presented for for the GxE means (yellow), predicted values by the GGE model (blue), predicted values by the W-GGE model (purple), predicted values by the AMMI model (orange), predicted values by the W-AMMI model (red), and predicted values by the linear mixed models (brown). These values are for QTLs detected when a 20 cM centred on the exact position of the simulated QTL. These plots are for a heritability of 0.5 in all 511 environments.

Table 2.4: Analysis of variance of the genotype main effects + genotype environment interaction model 2 for the Steptoe x Morex yield data.

Source†	df	SS‡	MS§
Total	2099	3538.200	1.7
Genotypes	149	633.700	4.2
GEI	1050	2554.700	2.4
PC1	156	1860.583	69.3
PC2	154	244.68	86.4
PC3	152	215.729	91.1
PC4	150	125.119	94.7
PC5-IPC7	438	239.343	1.6
Intra Block Error	900	349.8	0.4

† GEI, genotype-by-environment interaction; IPC, interaction principal component; ‡ SS, sums of squares; § MS, mean squares.

2.3.2.2 Weighted GGE Analysis

The barley data considered in this paper is partially replicated, thus, the cell means bring more information when there are two observations for the genotype environment combinations. Thus, in the W-GGE analysis, we use the matrix of weights as defined by Eq. [5], where $m = \max(1\sigma_{\epsilon_j}^2)$, $\sigma_{\epsilon_j}^2$, $j = 1, \dots, 8$ are the number of replications for genotype i in environment j , and $N_{rep} = 2$, is the maximum number of replications in the data set.

Figure 2.5 (right) shows the biplot for the W-GGE model. The environments with the greatest influence in the GGE analysis, because they have a greater error variance, have more homogeneous distribution when using the W-GGE analysis, which reduces their impact on the modelling.

Figure 2.5: Genotype main effects + genotype environment interaction model (GGE) (left) and weighted GGE (right) biplots based on the Steptoe x Morex barley yield data. The abscissa represents the first multiplicative term and the ordinate represents the second. The open dots represent the 150 genotypes and the codes for the 8 environments are defined in Table 2.2.

2.3.2.3 AQ and WAQ Analysis

Figure 2.6 shows the QTL scan of the 8 environments for the original data (grey dotted line), for the GGE predicted values (grey solid line), and for the W-GGE predicted values (solid black line) model. We can perceive an increase in the values of LOD scores when the scan is done with the predicted values obtained by the GGE model instead of the Gx $\times$ E means. In this figure we can identify the QTLs of each environment for all seven chromosomes. When comparing the results presented in Figure 6 with the results from Rodrigues et al. (2014), it can be seen that when using the original data, 17 QTLs are detected; when using the GGE predicted values, 17 QTLs are detected; when using the weighted GGE predicted values, 19 QTLs are detected; when using the AMMI predicted values, 22 QTLs are detected; and when using the W-AMMI predicted values, 29 QTLs are detected. Comparing all methods, we see that there is a relative increase in the values of LOD scores when considering the predicted values by the GGE and W-GGE model, especially in the QTLs detected on chromosome 3. When comparing the results of this paper with the results of (Rodrigues et al., 2014) for the AMMI and W-AMMI predicted values, there is a superiority of the weighted AMMI model in number of detected QTLs.

2.4 Conclusion

The proposed W-GGE algorithm can work with any weighting scheme. It is a generalization of the standard GGE analysis (Yan, 2001) that is able to account for heterogeneity in both genetic variances (captured by the main components of interaction in GGE) and the error variances (using the weighted generalization of the GGE model to estimate the interaction scores) between environments in multi-environmental trials. There are particular cases in MET in which there is little heterogeneity of error variances, and in these the GGE model is still fully appropriate. When considering the data, it should be borne in mind that the weights in our W-GGE algorithm require a (re) escalation that puts them between zero and one. Thus, the proposal of the weighted GGE model allows improving the modelling of interaction between genotypes and environments when the error variance is not homogeneous throughout the environments. A simple criterion to decide which approach to use, GGE or W-GGE, for complete data and completely randomized designs or randomized complete block designs, is to calculate the residual error or variance for each environment and verify whether it is homogeneous along the environments. If the variation of the error is homogeneous along the environments, the results of the GGE model will be similar to those of the W-GGE approach and the use of the standard GGE strategy will suffice. When

Figure 2.6: QTL scan for the eight environments in the SxM yield data. Dotted grey lines represent the QTL scans for the actual data, the grey solid lines represent the QTL scans for the GGE predicted values, and the solid black lines represent the QTL scans for the W-GGE predicted values. The codes above each individual scan represent the environments and are listed in Table 2.

error variances show high heterogeneity among environments, the use of the GGE model is not advisable and the W-GGE algorithm should be used. Important characteristic in the GGE model, such as biplots proposed by (Gariel, 1971), maintains the standard characteristics and interpretation in the W-GGE algorithm. Using the predicted values of the GGE and W-GGE models to obtain the QTL scans results in a considerable improvement in the number and accuracy of detected QTLs and higher LOD scores when compared to the use of the predicted values using the AMMI and W-AMMI, respectively. Although the use of the predicted values by the AMMI and W-AMMI models presents better results for the SxM yield data set, as they detect a higher number of QTLs, there is an increase in the number of QTLs detected when values predicted by the W-GGE model replace the values predicted by the GGE model. For the simulated data, the proposed W-GGE model results in a better performance in QTL detection and study of the interaction between QTL and environment when compared to other fixed effects models, and a performance similar to the linear model. The latter sometimes detects false positives. The model choice between W-AMMI and W-GGE for cases when there is a heterogeneity of error variance across environments, should be decided in similar manner as of between AMMI and GGE models. It should be based on the researcher's objectives, taking into consideration that the GGE model is more suitable for the identification of mega-environments, selection of representative and discriminative environments and indication of more cultivars adapted and stable to specific environments (Yan et al., 2007; Gauch et al., 2008; Asfaw et al., 2009; Yan, 2011) and the AMMI model in identifying superior environmental conditions for the farm (selection of cultivation sites) and higher average genotypes (Gauch et al., 2008; Yan, 2011).

Rodrigues et al. (2014) proposed a generalization of the AQ approach (QTL scans on the AMMI predicted values), taking into account the information on the heterogeneity of error variances between environments, the WAQ analysis (QTL scans on the W-AMMI predicted values). In this paper, we

propose the GQ approach (QTL scans on the GGE predicted values) and a generalization that takes into consideration the heterogeneity of error variance across environments, the WGQ approach (QTL scans on the W-GGE predicted values).

All methods were compared between them and with the linear mixed model by considering simulation data from an eco-physiological genotype-to-phenotype crop growth model and a real barley data set. When using the GGE and W-GGE predicted values to obtain QTLs scans, there is a considerable improvement in terms of the number and accuracy in the detected QTLs and higher LOD scores, when compared with the use of AMMI and W-AMMI predicted values, respectively. Although the use of the AMMI and W-AMMI predicted values yield better results for the SxM data, an increase in the number of detected QTLs is obtained when considering the GGE and W-GGE predicted values. For the simulated data, the predicted values by the proposed W-GGE model results in a better performance in QTL detection and better understanding of the QEI, when compared with the other fixed effects models (AMMI, W-AMMI and GGE). Moreover, the QTL scans on the W-GGE predicted values have a similar performance to the linear mixed model, while the latter sometimes detects false positives.

The results of this study are promising because it allows breeders that are familiar with the GGE model to understand and easily perform the weighted GQ analysis. The performance of the W-GGE analysis increases when the data includes environments with heterogeneous error variance, namely in detecting QTLs associated with interaction. Another strong point of this method is that the WGQ analysis can be performed using the open source R software (R Development Core Team), which is one of the most widely used statistical software packages. In addition, similar to the AMMI and GGE models, the W-GGE algorithm and the WGQ analysis presented here, are applicable to a wide range of fields such as plant breeding, crop sciences, genetics, microarray experiments (Crossa et al., 2005), rDNA studies (Adams e Eisenberg, 2002), plant and microbial populations, growth across several environmental conditions (Culman et al., 2009), animal sciences (Barhdadi e Dube, 2010), and human genetics (Mukherjee et al., 2012).

References

- Adams, G. C. and Wu, N. T. and Eisenberg, B. E. (2002). Virulence and double-stranded RNA in *Sphaeropsis sapinea*. *For. Pathol*, 32:309–329.
- Alwala, S., Kwolek, T., McPherson, M., Pellow, J., and Meyer, D. (2010). A comprehensive comparison between Eberhart and Russell joint regression and GGE biplot analyses to identify stable and high yielding maize hybrids). *Field Crops Research*, 119:225–230.
- Annicchiarico, P. (2002). *Genotype x Environment Interactions – Challenges and Opportunities for Plant Breeding and Cultivar Recommendations*. P. FAO, Italy.
- Annicchiarico, P. (2009). Non-linear state space modelling of fisheries biomass dynamics by using Metropolis-Hastings within-Gibbs sampling. *Coping with and exploiting genotype $\times$ environment interactions*. In *Plant Breeding and Farmer Participation*, pages 519–564.
- Araújo, M. F. C., Araújo, L. B., Faria, P. N., and Dias, C. T. S. (2012). Statistical test for genotype and environment contribution in the genotypes $\times$ environments interaction matrix). *Interciencia*, 37:515–519.
- Asfaw, A., Alemayehu, F., Gurum, F., and Atnaf, M. (2009). AMMI and GGE Biplot analysis for matching varieties onto soybean production environments in Ethiopia. *Scientific Research and Essay*, 4:1322–1330.

- Assis, T. O. G., Dias, C. T. S., and Rodrigues, P. C. (2018). A weighted AMMI algorithm for nonreplicated data). *Pesquisa Agropecuária Brasileira*, 53:557–565.
- Barchi, L., Bonnet, J., Boudet, C., Signoret, P., Nagy, I., Palloix, S. L. A., and Lefebvre, V. (2007). A High-Resolution. Intraspecific Linkage Map of Pepper (*Capsicum Annuum* L.) and Selection of Reduced Recombinant Inbred Line Subsets for Fast Mapping. *Genome* 50, 1:51–60.
- Barhdadi, A. and Dube, M. P. (2010). Testing for gene-gene interaction with AMMI models. *Stat. Appl. Genet. Mol. Biol.*
- Boer, M. P., Wright, D., Feng, L. Z., Podlich, D. W., Luo, L., Cooper, M., and van Eeuwijk, F. (2007). A mixed-model quantitative trait loci (QTL) analysis for multiple-environment trial data using environmental covariables for QTL-by-environment interactions, with an example in maize. *Genetics*, 177:1801–1813.
- Broman, K. W. and Sen, S. (2009). *A Guide to QTL Mapping with R/Qtl*. New York.
- Chenu, K., Chapman, S. C., Tardieu, S. C., McLean, G., Welcker, C., and Hammer, G. (2009). Simulating the yield impacts of organ-level quantitative trait loci associated with drought response in maize: A “gene-to-phenotype” modeling approach. *Genetics*, 183:1507–1523.
- Collard, B C Y abd Jahufer, M. Z. Z., Brouwer, J. B., and Pang, E. C. K. (2005). Na introductio to markers quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: The basic concepts. *Euphytica*, pages 142–169.
- Cooper, M. and DeLacy, I. H. (1994). Relationships among analytical methods used to study genotypic variation and genotype-by-environment interaction in plant breeding multi-environment experiments. *Theoretical and Applied Genetics*, 88:561–572.
- Crossa, J. (1990). Statistical analyses of multilocation trials. *Advances in Agronomy*, 44:55–85.
- Crossa, J., Burgueno, J., Autran, D., Vielle-Calzada, J. P., Cornelius, P. L., Garcia, N., Salamanca, F., and Arenas, D. (2005). Using linear-bilinear models for studying gene expression $\times$ treatment interaction in microarray experiments. *Journal of Agricultural, Biological, and Environmental Statistics*, 10:337–353.
- Crossa, J. and Cornelius, P. L. (1997). Statistical analyses of multilocation trials. *Crop Science*, 37:406–415.
- Crossa, J., Fox, P. N., Pfeiffer, W. H., Rajaram, S., and Gauch, H. G. (1991). Statistical analyses of multilocation trials. *Theoretical and Applied Genetics*, 81:27–37.
- Crossa, J., Vargas, M., van Eeuwijk, F. A., Jiang, C., Edmeades, G. O., and Hoisington, D. (1999). Interpreting genotype $\times$ environment interaction in tropical maize using linked molecular markers and environmental covariables. *Theoretical and Applied Genetics*, 99:611–625.
- Culman, S. W., Bukowski, R., Gauch, H. G., Cadillo-Quiroz, H., and Buckley, D. H. (2009). T-REX: Software for the processing and analysis of T-RFLP data. *BMC Bioinformatics*, 10:337–353.
- Deon, M., De Resende, V., and Thompson, R. (2004). Factor Analytic Multiplicative Mixed Models in the Analysis of Multiple Experiments. *Revista de Matemática e Estatística*, 22:31–52.
- Dias, C. T. S. and Krzanowski, W. J. (2006). Choosing components in the additive main effect and multiplicative interaction (AMMI) models. *Scientia Agricola*, 63:169–175.

- Doerge, R. W. (2002). Mapping and analysis of quantitative trait loci in experimental populations. *Nature Reviews Genetics*, 46:820–833.
- Edwards, J. W. and Jannink, J. L. (2006). Bayesian modeling of heterogeneous error and genotype X environment interaction variances. *Crop Science*, 46:820–833.
- Forkman, J. and Piepho, H. P. (2006). Statistical analysis of yield trials by AMMI and GGE. *Crop Science*, 46:1488–1500.
- Forkman, J. and Piepho, H. P. (2014). Parametric bootstrap methods for testing multiplicative terms in GGE and AMMI models. *Biometrics*, 70:639–647.
- Gariel, K. R. (1971). Biplot graphic display of matrices with application to principal component analysis. *Biometrika*, 58:453–467.
- Gauch, H. G. (1992). *Statistical analysis of regional yield trials: AMMI analysis of factorial designs*. Amsterdam.
- Gauch, H. G., Piepho, H. P., and Annicchiarico, P. (2008). Statistical Analysis of Yield Trials by AMMI and GGE: Further Considerations. *Crop Science*, 48:866–889.
- Gauch, H. G., Rodrigues, P. C., Munkvold, J. D., Heffner, E. L., and Sorrells, M. (2011). Two new strategies for detecting and understanding QTL $\times$ environment interactions. *Crop Science*, 51:96–113.
- Hadasch, S., Forkman, J., Malik, W. A., and Piepho, H. (2018). Weighted Estimation of AMMI and GGE Models. *Journal of Agricultural Biological and Environmental Statistics*, 23:255–275.
- Hadasch, S., Forkman, J., and Piepho, H. (2017). Cross-Validation in AMMI and GGE Models: A Comparison of Methods. *Crop Science*, 57:264–274.
- Hayes, P. M., Liu, B. H., Knapp, S. J., Chen, F., Jones, B., Blake, T., Franckowiak, J., Rasmusson, D., Sorrells, M., Ullrich, S. E., Wesenberg, D., and Kleinhofs, A. (1993). Quantitative trait locus effects and environmental interaction in a sample of North-American barley germ plasm. *Theoretical and Applied Genetics*, 87:392–401.
- Hongyu, K., Silva, F. L., Oliveira, A. C. S., Sarti, D. A., Araújo, L. B., and Dias, C. T. S. (2015). Comparação entre os modelos AMMI e GGE biplot para os dados de ensaios Multi-Ambientais. *Revista Brasileira de Biometria*, 33:139–155.
- Li, J. and Ji, L. (2005). Adjusting multiple testing in multilocus analyses using the eigen values of a correlation matrix. *Heredity*, 95:221–227.
- Liu, B. H. (1998). *Statistical genomics: linkage, mapping, and QTL analysis*.
- Malik, W. A., Hadasch, S., Forkman, J., and Piepho, H. P. (2018). Nonparametric resampling methods for testing multiplicative terms in AMMI and GGE models for multi-environment trials. *Crop Science*, 58:752–761.
- Malosetti, M., Ribaut, J. M., and van Eeuwijk, F. A. (2013). The statistical analysis of multi-environment data: Modeling genotype-by-environment interaction and its genetic basis. *Frontiers in Physiology*.
- McCouch, S. R. and Doerge, R. W. (1995). QTL mapping in rice. *Trends Genet*, 11:482–487.
- Mohan, M., Nair, S., Bhagwat, A., Krishna, T. G., Yano, M., and Bhatia, c R. and Sasaki, T. (1997). Genome mapping. molecular markers and marker-assisted selection in crop plants. *Mol Breed*, 3:87–103.

- Mukherjee, B., Ko, Y. A., VanderWeele, T., Roy, A., Parkd, S. K., and Chene, J. (2012). Principal interactions analysis for repeated measures data: Application to gene-gene and gene-environment interactions. *Statistics in Medicine*, 31:2531–2551.
- Paderewski, J. and Rodrigues, P. C. (2018). Constrained AMMI model: Application to Polish winter-wheat post-registration data. *Crop Science*, 58:1458–1469.
- Paterson, A. H. (1996a). *Making genetic maps*. Landes Company, San Diego, California.
- Paterson, A. H. (1996b). *Mapping genes responsible for differences in phenotype*. Landes Company, San Diego, California.
- Pereira, D. and Rodrigues, P. C., Mejza, S., and Mexia, J. T. (2012). A comparison between Joint Regression Analysis and the AMMI model: a case study with barley. *Journal of Statistical Computation and Simulation*, 82:193–207.
- Rodrigues, P. C. (2012). New Strategies to Detect and Understand Genotype-by-Environment Interactions and QTL-by-Environment Interactions - PhD Thesis. *Nova University of Lisbon*.
- Rodrigues, P. C. (2018). An overview of statistical methods to detect and understand genotype-by-environment interaction and QTL-by-environment interaction. *Biometrical Letters*, 55:123–138.
- Rodrigues, P. C., Heuvelink, E., Marcelis, L. F. M., Chapman, S., and van Eeuwijk, F. (2020). A complex trait with unstable QTLs can follow from component traits with stable QTLs: an illustration by a simulation study in pepper (submitted).
- Rodrigues, P. C., Malosetti, M., Gauch, H. G., and van Eeuwijk, F. A. (2014). A Weighted AMMI Algorithm to Study Genotype-by-Environment Interaction and QTL-by-Environment Interaction. *Crop Science*, 54:1555–1570.
- Rodrigues, P. C., Monteiro, A., and Lourenço, V. M. (2016). A Robust additive main effects and multiplicative interaction model for the analysis of genotype-by-environment data. *Bioinformatics*, 32:58–66.
- Rodrigues, P. C., Pereira, D., and Mexia, J. T. (2011). A comparison between JRA and AMMI: the robustness with increasing amounts of missing data. *Scientia Agricola*, 68:679–686.
- Romagosa, I., Ullrich, S. E., Han, F., and Hayes, P. M. (1996). Use of the additive main effects and multiplicative interaction model in QTL mapping for adaptation in barley. *Theoretical and Applied Genetics*, 93:30–37.
- Romagosa, I., van Eeuwijk, F. A., and Thomas, W. T. B. (2009). Statistical analyses of genotype by environment data. *Cereals Handbook of Plant Breeding*, pages 291–231.
- Samonte, S. O. P. B., Wilson, L. T., McClung, A. M., and Medley, J. C. (2005). Targeting Cultivars onto Rice Growing Environments Using AMMI and SREG GGE Biplot Analyses. *Crop Science*, 45:2414–2424.
- Sergio, P. E., Jarquin, D., and Crossa, J. (1993). Mapping poligenes. *Annual Review of Genetics*. London, 27:205–233.
- Tardieu, F. and Tuberosa, R. (2010). Dissection and modelling of abiotic stress tolerance in plants. *Current Opinion in Plant Biology*, 13:206–212.

- Toledo, E. R., Leandro, R., Júnior, C., and Souza, A. (2008). Mapeamento de QTLs : uma abordagem Bayesina. *Revista Brasileira de Biometria*, 26:107–114.
- van Eeuwijk, F. A., Bink, M. C. A. M., and Chenu, K. and Chapman, S. C. (2010). Detection and use of QTL for complex traits in multiple environments. *Current Opinion in Plant Biology*, 13:193–205.
- van Eeuwijk, F. A., Bustos-Kortsas, D. V., and Malosetti, M. (2016). What Should Students in Plant Breeding Know About the Statistical Aspects of Genotype $\times$ Environment Interactions? *Current Opinion in Plant Biology*, 56:2119–2140.
- Vieira, E. A., Nodari, R. O., Carvalho, F. I. F., and Fialho, J. F. (2006). *Mapeamento Genético de Caracteres Quantitativos e Sua Interação Com o Ambiente*. Documentos Embrapa, Cerrados.
- Yan, W. (2001). GGE biplot - a windows application for graphical analysis of multienvironmental trial data and other types of two-way data. *Agronomy Journal*, 93:1111–1118.
- 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., Hunt, L. A., Sheng, Q., and Z., S. (2000). Cultivar Evaluation and Mega-Environment Investigation Based on the GGE Biplot. *Crop Science*, 40.
- Yan, W., Kang, M., Ma, B., Woods, S., and Paul, L. C. (2007). GGE Biplot vs. AMMI Analysis of Genotype-by-Environment Data. *Crop Science*, 47:643–655.
- Yan, W. and Kang, M. S. (2003). *GGE biplot analysis: A graphical tool for breeders, geneticists and agronomists*. CRC Press, Florida.
- Yan, W. and Tinker, N. A. (2005). A Biplot Approach for Investigating QTL-by Environment Patterns. *Molecular Breeding*, 15:31–43.
- Yan, W. and Tinker, N. A. (2006). Biplot Analysis of Multi-Environment Trial Data: Principles and Applications. *Canadian Journal of Plant Science*, 86:623–645.
- Young, N. D. (1996). QTL mapping and quantitative disease resistance in plants. *Annual Review Phytopathology*, 34:479–501.

3 ROBUST AMMI MODEL IN THE STUDY OF INTERACTION BETWEEN QTL AND ENVIRONMENT: A SIMULATION APPROACH

Abstract

Data obtained from multi-environmental trials (METs) often shows, not only genotype-by-environment interactions (GEI), but also quantitative trait loci (QTL)-by-environment interactions (QEI) that interfere with productivity. Therefore, it is important for researchers and practitioners to find ways and develop tools to improve the results for decision making. Robust statistical methods stand out to deal with contaminated data, preventing the model parameters from being underestimated or overestimated when the data contains outlying observations. In this work the robust additive main effects and multiplicative interaction (R-AMMI) model is evaluated for the case where the data is contaminated with outlying observations. The R-AMMI model will be applied to simulated data from an eco-physiological genotype-to-phenotype crop growth model (CGM), being the results compared with the standard AMMI model, considering different scenarios for data contamination.

Keywords: Robust AMMI model; Genotype-by-environment interaction; QTL detection; QTL-by-environment interaction; crop growth model.

3.1 Introduction

Genotype-by-environment interaction (GEI) are present when the genotype responses to different levels of environmental stress are not consistent, which is quite common in multi-environmental trials (METs). The study and understand of these interactions can reveal differences of genotype performance across environments, being useful to analyze the stability of genotypes and the value of test sites (Dia et al., 2016). The basis of the GEI is the quantitative trait locus (QTL)-by-environment interaction (QEI), which occurs when QTLs effects differ across environments. A QTL analysis is a form of genetic analysis of quantitative traits or characteristics that are controlled by several genes. It uses lines that have been genotyped on many markers throughout the genome, and testes for association between the alleles on each marker and the phenotype expression of individuals with those alleles. An association between a phenotype and marker allele shows evidence that a location close to the marker contributes to a genetic variation in the phenotype (Donohue, 2016). Thus, with the development of molecular markers and mapping techniques, researchers can take a step further and analyze the entire genome to discover the specific location of genes that influence a quantitative trait, such as yield. Thus, a good understanding of these interactions allows the selection of the best genotypes under different environmental conditions.

GEI can be quantified using various procedures based on the evaluation of genotypes under multiple environments. These methods are divided into univariate and multivariate stability statistics. The most used univariate methods are based on regression of the average value of each genotype in the environmental index or marginal average of the environments (Yates, 1938; Finlay e N., 1963; Eberhart e Russel, 1966). The GEI multivariate analysis is an alternative and includes complementary methods to evaluate the genotype stability (Crossa, 1990), such as a widely used method to investigate the GEI, the additive main effects and multiplicative interaction (AMMI) model (Gauch, 1992). The AMMI model uses the analysis of variance (ANOVA) to estimate the genotypic and environmental main effects and the singular value decomposition (SVD), applied to the residuals of the ANOVA, to estimate the components that characterize the interaction. This method has proven to be effective because it captures a large part of the sum of squares of the interaction, separates the main and the interaction effects and can provide a meaningful interpretation of the data (Ebdon e Gauch Jr., 2002). In other words, it is capable of extracting a large part of the GEI and, therefore, it is more efficient in the analysis of the

interaction pattern (Gauch, 1992; Gauch e Richard, 1997; Ebdon e Gauch Jr., 2002). The results of the AMMI analysis have been useful to support decisions related to improvement of breeding programs, such as specific adaptation and environment selection. In this way, it stands out for being able to identify genotypes with specific adaptability, as well as stable genotypes under certain environmental conditions (Mohammadi et al., 2017). Another highlight of the results of the analyzes with the AMMI model is that they can be expressed in graphs called biplots that show the genotype scores and environments scores based on the SVD (Gauch e Richard, 1997).

However, the AMMI model is not indicated in some cases, such as when there is the presence of peripheral information resulting from measurement errors or due to individual intrinsic characteristics, thus generating a lack of normality. In these cases, robust statistical techniques are recommended, which are designed for cases in which one or more hypotheses of the traditional method are not satisfied. Robust methods are accessible as they are able to restrict the influence of outliers by preventing the model parameters from being underestimated or overestimated (Nargis, 2005).

Since data contamination is commonly identified in real data analysis, the research on statistical methods that allow valid results is of great importance, even if the model assumptions are o violated (Lourenço et al., 2011). In this context, the robust AMMI model (Rodrigues et al., 2016) can be used to overcome the problem of analyzing and model contaminated data with outlying observations and to better understand the GEI. In the robust AMMI model, the ANOVA linear fit is replaced by a robust fit (Huber, 1964) and the standard SVD is replaced by a robust SVD approach (see a review in P. e Todorov (2013)). However, the context in which the robust AMMI model was developed (Rodrigues et al., 2016) was only for the study of GEI.

In this chapter we generalize the use of the AMMI model to model and understand QEI and for QTL detection. The performance of the proposed robust extension of the AMMI model is evaluated through Monte Carlo simulations, in which several contamination schemes are considered. The method is also compared with the results obtained by the classic AMMI and the weighted AMMI model (Rodrigues et al., 2014) for QTL detection and QEI behaviour. The data under consideration includes 100 two-way data tables with 200 genotypes and 12 environments with simulated yield of a pepper (*Capsicum annuum*) population, by using a genotype-to-phenotype crop growth model (Rodrigues et al., 2014, 2020). The data was contaminated in a similar way as described in Rodrigues et al. (2016).

3.2 Material and Methods

3.2.1 Plant Materials

Data in this study consists on 100 two-way data tables with 200 genotypes (G) and 12 environments (E) with yield. These data sets were simulated based on a crop growth model and assume that the yield results from the sum of the signal and the noise. The signal of genotype i in the environment j was simulated from a physiological genotype-phenotype crop growth model (CGM) for pepper (Rodrigues, 2012; Rodrigues et al., 2020), and it is a function of parameters that ultimately depend on genotype or environment, but never on the genotype and environment simultaneously. Complete details about the simulation model can be found in Rodrigues (2012) and Rodrigues et al. (2020).

The model can be written as follows (Rodrigues, 2012; Rodrigues et al., 2020):

$$\begin{aligned}
 Yield_{i,j} &= Signal_{i\bar{V}_j} + Noise_{i,j} \\
 &= (Radiation\ intercepted \times Conversion\ efficiency \times Partitioning\ to\ yield)_{i\bar{V}_j} + Noise_{i,j} \\
 &= \left\{ \sum_{t=t_0}^{t_f} [1 - \exp(-k_i \times LAI_{i,j,t})]_{j,t} \right\} \times \{LUE_{i,j}\} \times \{FTF_i \times [1 - W_i \times (T_j - T_{FTF})]FDMC_i\} + \varepsilon_{i,j} \quad (3.1)
 \end{aligned}$$

Table 3.1: Description of the environments considered in the genotype-to-phenotype crop growth model in Eq. [3.1]. The countries were chosen to represent different environmental and practical conditions (Rodrigues, 2012; Rodrigues et al., 2020). Radiation has two levels (years) based on historical data. Temperature contains three levels of daily average temperature. The last column gives the error variance for the 100 simulated data sets. The heritability for the environments was set to be $h^2 = 0.5$.

Environment	Country	Radiation	Temperature	Error Variance
NL1-15	Netherlands	Lower	15 ⁰ C	22.4
NL1-20	Netherlands	Lower	20 ⁰ C	38.01
NL1-25	Netherlands	Lower	25 ⁰ C	29.82
NL5-15	Netherlands	Higher	15 ⁰ C	38.27
NL5-20	Netherlands	Higher	20 ⁰ C	72.81
NL5-25	Netherlands	Higher	25 ⁰ C	57.73
SP1-15	Spain	Lower	15 ⁰ C	11.85
SP1-20	Spain	Lower	20 ⁰ C	17.84
SP1-25	Spain	Lower	25 ⁰ C	17.71
SP5-15	Spain	Higher	15 ⁰ C	14.92
SP5-20	Spain	Higher	20 ⁰ C	22.61
SP5-25	Spain	Higher	25 ⁰ C	26.94

The radiation intercepted depends on:

- (i) light extinction coefficient (K_i);
- (ii) leaf area index of genotype i in environment j on day t , this is, $LAI_{i,j,t} = [a + B_i(T_j - T_{base}) \times (t - t_0)] \times Sd_j$, being a and B_j , the constant intercept and the genotype specific slope in a regression of the leaf area per stem (m^2) as a function of the temperature sum ($^{\circ}C$), T_j is the average temperature per environment j , T_{base} is the base temperature, t represents the t -th day of the growing season ($t = t_0$ is the first day of flowering) and Sd_j is the density of the stem in the environment j ;
- (iii) $I_{j,t} = RAD_{j,t} \times F_{PAR} \times Tr_j$ is the photosynthetic active radiation incident in the harvest by environment j on day t , that is, is the product of:
 1. global radiation of day t in the environment j ($RAD_{j,t}$);
 2. fraction of photosynthetic radiation (PAR) activity in global radiation (F_{PAR});
 3. transmissivity of greenhouse effect in the environment j (Tr_j).

The sustained t_0 and t_f represent the beginning and end of the growing season, in days. The light use efficiency for genotype i in the environment j ($LUE_{i,j}$) depends on the temperature given by the parameter Z_i - linear slope reduction for LUE at temperatures below 20⁰C. The partitioning to yield, as described on the right side of the equation, is defined by a combination of the partitioned dry weight fraction of the fruits (FTF_i), fruit dry matter content ($FDMC_i$) and slope of the linear reduction in the harvest index with temperatures above 15⁰C (W_i).

The Model (3.1) is a function of seven physiological parameters that depend on the genotype. Three environmental variables are considered (Table 3.1):

- (i) temperature with three levels: low (15⁰C), medium (20⁰C) and high (25⁰C);
- (ii) radiation, with two levels/years, based on the history of meteorological data by country;
- (iii) country, with two levels: Spain and the Netherlands.

The 100 simulated data sets were subject to a contamination process according to Rodrigues et al. (2016). Having the good non-contaminated data, a percentage of contamination is introduced in the two-way original data table so as to be consistent with the known shift and point mass outliers described in Rocke and Woodru (1996): the bad data is generated from the normal distributions $N(\mu_j + k\sigma_j^2, \sigma_j^2)$ (pure shift outliers), $N(\mu_j, \sigma_j^2/100)$ (pure point-mass outliers) and $N(\mu_j + k\sigma_j^2, \sigma_j^2/100)$ (shift-point-mass outliers) where $k = 4; 5; 6; 7$ and 10 units, and μ_j and σ_j^2 are taken as the sample phenotypic mean and sample phenotypic variance according to the correspondent environment j . The “bad data” replaces a percentage of the good data from the two-way table at the assigned positions for three scenarios:

- (i) positions randomly selected in the two-way table thus randomly assigning contamination in different environments for distinct genotypes (2; 5 & 10% scattered contamination);
- (ii) positions randomly assigned in only one of the environments (5; 10; 15 & 20% single-environment contamination); and
- (iii) a mix between (i) and (ii), i.e., one of the environments is assigned a percentage of contamination (5; 10; 15 & 20%) and then scattered contamination is considered in the rest of the two-way table, excluding this environment and up to a given total percentage (2; 5 & 10%) - a more realistic scenario (mixed contamination) Rodrigues et al. (2016).

3.2.2 AMMI Model

The AMMI model is widely used for yield analysis in multi-environmental trials (METs). It combines the characteristics of analysis of variation (ANOVA) and singular value decomposition (SVD), estimating the main additive effects and the interaction multiplicative effects (Rodrigues et al., 2014). The model can be written as follows:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \theta_{jk} + \sum_{n=1}^N \lambda_n \gamma_{ni} \delta_{nj} + \varepsilon_{ijk} \quad (3.2)$$

where: Y_{ijk} is the observed value of the i -th genotype in the j -th environment and in the k -th block; μ is the grand mean; α_i is the effect of the i -th genotype; β_j is the effect of the j -th environment; θ_{jk} is the effect of the k -th block in the j -th environment; λ_n is the singular value of the interaction principal component in the n -axis; γ_{ni} and δ_{nj} are the PC scores of the i -th genotype and j -th environment for the n -axis; $N \leq \min(I - 1, J - 1)$, with I the number of genotypes and J the number of environments, is the number of interaction principal components; and ε_{ijk} is the experimental error.

The matrix formulation of equation (1) can be given by:

$$\mathbf{Y} = \mathbf{1}_I \mathbf{1}_J^T \mu + \alpha \mathbf{1}_I^T + \mathbf{1}_J^T \beta + \mathbf{U} \mathbf{D} \mathbf{V}^T + \varepsilon \quad (3.3)$$

where $\mathbf{Y}$ is the matrix ($I \times J$) with the yield (or other phenotypic trait of interest); $\mathbf{1}$ is a vector with the value one in all positions; μ is the matrix ($I \times J$) with the grand mean in all positions; α is the vector of genotypic main effects; β is the vector of environment main effects; $\mathbf{U}$ the matrix ($I \times N$) whose columns contain the left singular vectors of the interaction; $\mathbf{D}$ is the matrix ($N \times N$) which diagonal contains the singular value, $\mathbf{V}$ is the matrix ($J \times N$) whose columns contain the right singular vectors of the interaction, and ε is the matrix ($I \times J$) with the experimental errors.

3.2.3 Robust AMMI Model

The presence of data contamination with outlying observations is common in real data collected in many areas of knowledge. In such cases, the use of robust methodologies is indicated for a better

modelling that avoids an over-impact of outliers. Thus, the robust AMMI model presents itself as a good alternative to the classical AMMI model, with the main advantage the ability of reducing the influence of outliers in the model (Rodrigues et al., 2016). To obtain robust regression estimators, the first stage of the robust AMMI model, it is necessary to have a model that describes most of the information in the sample, so we must assign weights in the calculation of the estimates (Rousseeuw e Annick, 1987).

Huber (1964) proposed the class of M-estimators in the context of regression models, thus being a pioneer in studies of robust estimation. The class of M-estimators has been extended to all probability distributions and generalizes the maximum likelihood method, producing consistent and asymptotically normal estimators (Heritier et al., 2009). Here we use the Huber M estimator to estimate the additive main effects of the robust AMMI model (R-AMMI, Rodrigues et al., 2016). Thus, the additive main effects of the R-AMMI model are determined by minimizing the sum of the objective function Huber (1973):

$$\sum_{i=1}^n \rho(\epsilon_i) = \sum_{i=1}^n \rho(y_i - \hat{y}_i)$$

with ϵ the residuals, and $\rho(\cdot)$ the function that maps the contribution of each residual to the objective function, being some of its properties continuity and symmetrical, with a minimum value of zero (Andersen et al., 2008; Rousseeuw e Annick, 1987). Thus, the Huber M estimator is obtained when:

$$\rho(\epsilon_i) = \begin{cases} \frac{1}{2}\epsilon_i^2, & \text{se } |\epsilon_i| \leq c \\ c(|\epsilon_i| - \frac{1}{2}c), & \text{se } |\epsilon_i| > c \end{cases}$$

The value of c is called the tuning constant for Huber M estimators. The lower the value of c , the less sensitive the presence of outliers will be the estimator. However there is a cost to efficiency when errors are normally distributed. The value of the constant is generally chosen in order to obtain acceptable efficiency when the errors have a normal distribution. Huber (1973) indicated that $c = 1.345 \times \sigma$, with σ being the standard deviation of the error, which produces an efficiency of 95% considering that the errors have a normal distribution. However, in this work we use $c = 1.345$ to calculate the additive main effects, that is, we consider that the standard deviation of the errors is equal to one.

Huber's M estimate defines a weight function such that $w(\epsilon) = \frac{\rho'(\epsilon)}{\epsilon}$ and $w_i = w(\epsilon_i)$. In this way, the estimated equation can be written as $\sum w_i \epsilon_i x'_{ij} = 0$.

The weight matrix is a function of the residuals, thus, the equation is solved using iteratively re-weighted least squares. The weight function for the Huber M estimator defines that observations with low residual value obtain a weight equal to one and the higher the value of the residual, the less weight will be assigned to the sample. The standard weighting function is defined as (Huber, 1973).

$$\rho(\epsilon_i) = \begin{cases} 1, & \text{se } |\epsilon_i| \leq c \\ c(\frac{c}{|\epsilon_i|}), & \text{se } |\epsilon_i| > c \end{cases} \quad (3.4)$$

The R-AMMI model proposed by Rodrigues et al. (2016) consists of the adjustment of a robust regression model based on the Huber M estimator (Huber, 1981), which uses the weight function defined in equation 3 for the main additive effects of the model estimation. Sequentially, the robust singular values decomposition defined by Hawkins et al. (2002) is applied to the residual matrix to estimate the multiplicative effects of the R-AMMI model. The robust singular value decomposition uses the L1 norm, instead of the usual L2 (least squares) norm, to calculate the robust approximation of a rectangular matrix (Rodrigues et al., 2016).

3.2.4 Robust AQ Analysis

Knowing that the AQ analysis is obtained by fitting the AMMI model to the $G \times E$ table of means, followed by making QTL scans on the AMMI predicted values for each environment (Gauch et al., 2011) we propose the RAQ analysis is a analysis where the AMMI analysis is replaced by the robust AMMI analysis proposed before. In this approach, QTL scans are not made directly in the means of genotype $\times$ environment interaction, but based on the predictions of the R-AMMI model. This analyses are particularly useful to analyse data sets with outlying observations.

3.3 Results and Discussion

3.3.1 No data contamination

Figure 3.1 shows the QTL scan for on random, out of the simulated 100, without data contamination. The dotted segments represent the QTL scans for the original data, the gray lines represent the QTL scans for the AMMI predicted values and the black lines represent the proposed QTL scans based on the R-AMMI predicted values. Although some trends can be found in this figure, the summary for the number of detected QTLs in all 100 data sets is shown in Figure 3.2, that includes a comparison between the number of QTLs detected by: (i) raw data, (ii) AMMI predicted values (Gauch et al., 2011), (iii) weighted AMMI predicted values (Rodrigues et al., 2014), and (iv) robust AMMI predicted values. The left hand side plot gives the boxplots for the model parameters (Rodrigues, 2012; Rodrigues et al., 2020) and the right hand side plot gives the boxplots for main clusters of environments. Figure 3.3 gives the boxplots for individual environments. Based on the analysis of the boxplots in Figures 3.2 and 3.3, it can be seen that the number of QTLs detected using the AMMI, W-AMMI and R-AMMI predicted values is superior to the number of QTLs detected on the raw data. As pointed out by (Gauch et al., 2011) and (Rodrigues et al., 2014), the use of AMMI related models to obtain predicted values that are subjected to QTL scans, results in a strengthening of the individual environments with information coming for the other environments. In this case, and as expected because of no data contamination, the distribution of the number of QTLs based on the AMMI predicted values is similar to the distribution of the number of QTLs based on the R-AMMI predicted values.

3.3.2 Contaminated data

When the data is contaminated with outlying observations, it is expected to observe differences between the number of QTLs detected based on the AMMI predicted values and the number of QTLs detected based on the R-AMMI predicted values. Figure 3.4 shows the QTL scan for on random, out of the simulated 100, with 5% scattered contamination with shift point-mass outliers equal to 4. The summary for the number of detected QTLs in all 100 data sets is shown in Figures 3.2 and 3.3, that present a comparison between the number of QTLs detected by: (i) raw data, (ii) AMMI predicted values, (iii) weighted AMMI predicted values, and (iv) robust AMMI predicted values. Based on the analysis of the boxplots in 3.2 and 3.3, when there is 5% scattered contamination with shift point-mass outliers equal to 4, the overall number of QTLs detected by with the R-AMMI predicted values is higher than the other methods. Moreover, it is visible that for the parameters FTF+FDMC (main effects parameters), the W-AMMI predicted values results in an higher number of QTLs, for the parameters W+Z (interaction parameters, temperature based parameters) the W-AMMI and R-AMMI predicted values result in a similar higher number of QTLs, and for the parameter LUE (main effects parameter), the AMMI and R-AMMI predicted values result in a similar higher number of QTLs (left hand side of Figure 3.2). When considering clusters of environments, in the clusters NL1, SP1 and SP2 the R-AMMI predicted values results in an overall higher number of detected QTLs, while for the cluster NL5, the

Figure 3.1: QTL scan for one random realization of the yield data for the pepper population, simulated from the genotype-phenotype model with seven physiological parameters. Each row of charts represents a different level of temperature. The dotted segments represent the QTL scans for the original data, the gray lines represent the QTL scans for the AMMI predicted values and the black lines represent the proposed QTL scans based on the R-AMMI predicted values. The horizontal lines correspond to the thresholds based on a permutation test with 1000 permutations, at a significance level of 0.05. These analyses are based on random dataset among the 100 simulations, with no data contamination. The codes for the environments are as defined above.

Figure 3.2: Boxplots for the number of detected QTLs in all 100 data sets, for the raw data (yellow), AMMI predicted values (orange), weighted AMMI predicted values (red), and robust AMMI predicted values (green). The left hand side plot gives the boxplots for the model parameters and the right hand side plot gives the boxplots for the main clusters of environments. These values are for QTLs detected when considering an interval of 20 cM centred on the exact position of the simulated QTL. These plots are for a heritability of 0.5 in all environments.

higher number of detected QTLs was obtained from the W-AMMI predicted values (right hand side of Figure 3.5). Figure 3.3 reinforces the results obtained in Figure 3.5, and in both plots is clear that the

Figure 3.3: Boxplots for the number of detected QTLs in all 100 data sets, for the raw data (yellow), AMMI predicted values (orange), weighted AMMI predicted values (red), and robust AMMI predicted values (green). The boxplots are for individual environments, for an expected maximum of seven (environments with temperature of 15°C or 20°C) or eight (environments with temperature of 25°C). These values are for QTLs detected when considering an interval of 20 cM centred on the exact position of the simulated QTL. These plots are for a heritability of 0.5 in all environments.

number of QTLs detected based on the raw data is underestimated and a better accuracy for the true number of QTLs is obtained from the QTL scans on the R-AMMI predicted values.

Figure 3.4: QTL scan for one random realization of the yield data for the pepper population, simulated from the genotype-phenotype model with seven physiological parameters. Each row of charts represents a different level of temperature. The dotted segments represent the QTL scans for the original data, the gray lines represent the QTL scans for the AMMI predicted values and the black lines represent the proposed QTL scans based on the R-AMMI predicted values. The horizontal lines correspond to the thresholds based on a permutation test with 1000 permutations, at a significance level of 0.05. These analyses are based on random dataset among the 100 simulations, with 5% scattered contamination with shift point-mass outliers equal to 4. The codes for the environments are as defined above.

The other contamination scenarios were analysed in a similar manner and with similar results. Figures 1S-11S show the QTL scan for on random, out of the simulated 100, being the summary for the

Figure 3.5: Boxplots for the number of detected QTLs in all 100 data sets, for the raw data (yellow), AMMI predicted values (orange), weighted AMMI predicted values (red), and robust AMMI predicted values (green). The left hand side plot gives the boxplots for the model parameters and the right hand side plot gives the boxplots for the main clusters of environments. These values are for QTLs detected when considering an interval of 20 cM centred on the exact position of the simulated QTL. These plots are for a heritability of 0.5 in all environments. These results are based on random dataset among the 100 simulations, with 5% scattered contamination with shift point-mass outliers equal to 4.

Figure 3.6: Boxplots for the number of detected QTLs in all 100 data sets, for the raw data (yellow), AMMI predicted values (orange), weighted AMMI predicted values (red), and robust AMMI predicted values (green). The boxplots are for individual environments, for an expected maximum of seven (environments with temperature of $15^{\circ}C$ or $20^{\circ}C$) or eight (environments with temperature of $25^{\circ}C$). These values are for QTLs detected when considering an interval of 20 cM centred on the exact position of the simulated QTL. These plots are for a heritability of 0.5 in all environments. These results are based on random dataset among the 100 simulations, with 5% scattered contamination with shift point-mass outliers equal to 4.

number of detected QTLs in all 100 data sets is shown in Figures 16S-38S, that present a comparison between the number of QTLs detected by:

- (i) raw data;
- (ii) AMMI predicted values;
- (iii) weighted AMMI predicted values;

(iv) robust AMMI predicted values.

The considered contamination scenarios are:

- (i) 5% scattered contamination with shift outliers equal to 4 (QTL scans in Figure A.2 and comparative boxplots in Figures A.17 and A.18);
- (ii) 5% scattered contamination with shift outliers equal to 7 (QTL scans in Figure A.3 and comparative boxplots in Figures A.19 and 1A.20);
- (iii) 5% scattered contamination with shift outliers equal to 10 (QTL scans in Figure 3S and comparative boxplots in Figures A.21 and A.22);
- (iv) 5% scattered contamination with shift point-mass outliers equal to 7 (QTL scans in Figure 4S and comparative boxplots in Figures A.23 and A.24);
- (v) 5% scattered contamination with shift point-mass outliers equal to 10 (QTL scans in Figure 5S and comparative boxplots in Figures A.25 and A.26);
- (vi) 10% single environment contamination with shift outliers equal to 4 (QTL scans in Figure 6S and comparative boxplots in Figures A.27 and A.28);
- (vii) 10% single environment contamination with shift outliers equal to 7 (QTL scans in Figure 7S and comparative boxplots in Figures A.29 and A.30);
- (viii) 10% single environment contamination with shift outliers equal to 10 (QTL scans in Figure 8S and comparative boxplots in Figures A.31 and A.32);
- (ix) 10% single environment contamination with shift point-mass outliers equal to 4 (QTL scans in Figure 9S and comparative boxplots in Figures A.33 and A.34);
- (x) 10% single environment contamination with shift point-mass outliers equal to 7 (QTL scans in Figure 10S and comparative boxplots in Figures A.35 and A.36);
- (xi) 10% single environment contamination with shift point mass outliers equal to 10 (QTL scans in Figure 11S and comparative boxplots in Figures A.37 and A.38), respectively.

Figure A.39 shows that, both for scattered contamination and single environment contamination, the number of detected QTLs shows the same behaviour as in Figure 3.5.

A comparative analysis can also be done by means of biplots. Figure A.13 shows the biplots for the AMMI and R-AMMI models with two interaction principal components for one random two-way data table without data contamination, in comparison with data contaminated with mixed contamination. Figures 13S-15S show the biplots for the all considered contamination scenarios where a comparison is made between the AMMI, W-AMMI and R-AMMI models. Based on the analysis of these biplots, it is clear that in the presence of data contamination, the R-AMMI biplot shows more similar behaviour to the baseline AMMI biplot for no data contamination.

3.4 Conclusion

Data contaminated with outlying observations is common in real life multi-environment experiments, either because of measurement errors or atypical observations influenced by, e.g., pests and diseases. This anomalous observations might then have a large influence in the analysis using standard statistical models. This chapter proposes a generalization of the AQ approach by taking into account the presence of outliers in the analyzed data, the RAQ approach, which has the advantage of being easily understood by researchers already familiar with the AMMI methodology. Using the proposed methodology we have to, overall, for contaminated data, the number of true positive QTLs detected by the QTL scans on the predicted values of the R-AMMI model outperform the number of true positive QTLs detected by the QTL scans on the predicted values of the AMMI and W-AMMI models. The latter uses the WAQ approach, proposed by Rodrigues et al. (2014), as a generalization of the AQ approach, which takes into account the heterogeneity of error variance across environments. If we compare the number of QTLs

detected in the row data, they are always much lower than the number of QTLs detected via AMMI, W-AMMI and R-AMMI models. Thus, we have to use the mentioned models to improve the detection of QTLs. And, like the AMMI model, the R-AMMI algorithm and RAQ analysis are fully applicable to various scenarios, such as plant breeding, crop sciences, genetics, microarray experiments, plant and microbial populations.

References

- Andersen, S., Harrison, G. W., Lau, M. I., and Rutstrom, E. E. (2008). Eliciting Risk and Time Preferences. *Econometrica*, 76:583–618.
- Crossa, J. (1990). Statistical analyses of multilocation trials. *Advances in Agronomy*, 44:55–85.
- Dia, M., Wehner, T. C., and Arellano, C. (2016). Analysis of Genotype $\times$ Environment Interaction (G $\times$ E) Using SAS Programming. *Agronomy Journal*, 108:1838–1852.
- Ebdon, J. S. and Gauch Jr., H. G. (2002). Additive Main Effect and Multiplicative Interaction Analysis of National Turfgrass Performance Trials: I. Interpretation of Genotype $\times$ Environment Interaction. *Crop Science*, 42:489–496.
- Eberhart, S. A. and Russel, W. A. (1966). Stability parameters for comparing varieties. *Crop Science*, 6:36–40.
- Finlay, K. W. and N., W. O. (1963). The analysis of adaptation in a plant breeding programme. *Australian Journal of Agricultural Research*, 14:742–754.
- Gauch, H. G. (1992). *Statistical analysis of regional yield trials: AMMI analysis of factorial designs*. Amsterdam.
- Gauch, H. G. and Richard, W. Z. (1997). Identifying Mega-Environments and Targeting Genotypes. *Crop Science*, 37:311–326.
- Gauch, H. G., Rodrigues, P. C., Munkvold, J. D., Heffner, E. L., and Sorrells, M. (2011). Two new strategies for detecting and understanding QTL $\times$ environment interactions. *Crop Science*, 51:96–113.
- Hawkins, D. M., Liu, L., and Young, S. S. (2002). Robust Singular Value Decomposition. *Technical Report NISS*.
- Heritier, S., Cantoni, E., Copt, S., and Victoria-Feser, M. (2009). *Robust Methods in Biostatistics*. Wiley-Blackwell.
- Huber, P. J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics*, 35:73–101.
- Huber, P. J. (1973). Robust Regression: Asymptotics, Conjectures and Monte Carlo. *Annals of Statistics*, 1:799–821.
- Lourenço, V. M., Pires, A. M., and M., K. (2011). Robust linear regression methods in association studies. *Bioinformatics*, 27:815–821.
- Mohammadi, R., Armion, M., Zadhasan, E., Ahmadi, M. M., and A., A. (2017). The use of AMMI model for interpreting genotype $\times$ environment interaction in durum Wheat. *Experimental Agriculture*, 54:670–683.

- Nargis, S. (2005). Robust Methods in Logistic Regression. *University of Canberra. Division of Business L and Is*.
- P., F. and Todorov, V. (2013). Robust Tools for the Imperfect World. *Information Sciences*, 245:4–20.
- Rodrigues, P. C. (2012). New Strategies to Detect and Understand Genotype-by-Environment Interactions and QTL-by-Environment Interactions - PhD Thesis. *Nova University of Lisbon*.
- Rodrigues, P. C., Heuvelink, E., Marcelis, L. F. M., Chapman, S., and van Eeuwijk, F. (2020). A complex trait with unstable QTLs can follow from component traits with stable QTLs: an illustration by a simulation study in pepper (submitted).
- Rodrigues, P. C., Malosetti, M., Gauch, H. G., and van Eeuwijk, F. A. (2014). A Weighted AMMI Algorithm to Study Genotype-by-Environment Interaction and QTL-by-Environment Interaction. *Crop Science*, 54:1555–1570.
- Rodrigues, P. C., Monteiro, A., and Lourenço, V. M. (2016). A Robust additive main effects and multiplicative interaction model for the analysis of genotype-by-environment data. *Bioinformatics*, 32:58–66.
- Rousseeuw, P. J. and Annick, M. L. (1987). *Robust regression and outlier detection*. New York.
- Yates, F. and Cochran, W. (1938). The analysis of groups of experiments. *The Journal of Agricultural Science*, 28:556–580.

4 FINAL CONSIDERATIONS

The use of several methodologies, such as AMMI, W-AMMI, GGE, W-GGE and R-AMMI models enrich the analysis of genotype by environment interaction and QTL by environment interaction. Since, often data from multi-environment experiments do not present ideal conditions, such as heterogeneity of the error variances across environments or the presence of data contamination with outlying observations, the proposal of new models that account for these problems in the data are of great importance. Those methods allow researchers and breeders to obtain more accurate results and better diagnoses regarding the identification of superior genotypes for given environmental conditions.

In this thesis it I provided three major contributions for the analysis of GEI and QEI: (i) the proposal of the weighted GGE model to model data with heterogeneity of error variances across environments; (ii) to improve the LOD scores for the QTLs based on the weighted AMMI predicted values; and (iii) to propose the use of the robust AMMi predicted values for QTL detection. The proposed methodologies were compared with standard models with real and simulated data.

APPENDIX

Appendix chapter 1

Figure A.1: QTL scans for the 12 environments of the yield data for pepper simulated from the physiological genotype-to-phenotype model with seven physiological parameters (Rodrigues et al. 2012). Each column represents a different level of temperature. The first two rows correspond to the highest error variance in this realization of the simulation and the last two rows to the lowest. The dashed grey line represents the scans for the actual data, the grey line the scans for the GGE predicted values, and the black line the scans for the W-GGE predicted values. All scans are based on composite interval mapping. The horizontal lines correspond to the thresholds based on a permutation test with 1000 permutations, at the 0.05 significance level. These scans are based on one randomly chosen realization out of the 100 simulations. The codes for the captions of the individual scans are described in Table 1.

Appendix chapter 2

Figure A.2: QTL scan for one random realization of the yield data for the pepper population, simulated from the genotype-phenotype model with seven physiological parameters. Each row of charts represents a different level of temperature. The dotted segments represent the QTL scans for the original data, the gray lines represent the QTL scans for the AMMI predicted values and the black lines represent the proposed QTL scans based on the R- AMMI predicted values. The horizontal lines correspond to the thresholds based on a permutation test with 1000 permutations, at a significance level of 0.05. These analyses are based on random dataset among the 100 simulations, with 5% scattered contamination with shift equal 4. The codes for the environments are as defined above.

Figure A.3: QTL scan for one random realization of the yield data for the pepper population, simulated from the genotype-phenotype model with seven physiological parameters. Each row of charts represents a different level of temperature. The dotted segments represent the QTL scans for the original data, the gray lines represent the QTL scans for the AMMI predicted values and the black lines represent the proposed QTL scans based on the R-AMMI predicted values. The horizontal lines correspond to the thresholds based on a permutation test with 1000 permutations, at a significance level of 0.05. These analyses are based on random dataset among the 100 simulations, with 5% scattered contamination with shift equal 7. The codes for the environments are as defined above.

Figure A.4: QTL scan for one random realization of the yield data for the pepper population, simulated from the genotype-phenotype model with seven physiological parameters. Each row of charts represents a different level of temperature. The dotted segments represent the QTL scans for the original data, the gray lines represent the QTL scans for the AMMI predicted values and the black lines represent the proposed QTL scans based on the R-AMMI predicted values. The horizontal lines correspond to the thresholds based on a permutation test with 1000 permutations, at a significance level of 0.05. These analyses are based on random dataset among the 100 simulations, with 5% scattered contamination with shift equal 10. The codes for the environments are as defined above.

Figure A.5: QTL scan for one random realization of the yield data for the pepper population, simulated from the genotype-phenotype model with seven physiological parameters. Each row of charts represents a different level of temperature. The dotted segments represent the QTL scans for the original data, the gray lines represent the QTL scans for the AMMI predicted values and the black lines represent the proposed QTL scans based on the R-AMMI predicted values. The horizontal lines correspond to the thresholds based on a permutation test with 1000 permutations, at a significance level of 0.05. These analyses are based on random dataset among the 100 simulations, with 5% scattered contamination with shift point-mass outliers equal to 7. The codes for the environments are as defined above.

Figure A.6: QTL scan for one random realization of the yield data for the pepper population, simulated from the genotype-phenotype model with seven physiological parameters. Each row of charts represents a different level of temperature. The dotted segments represent the QTL scans for the original data, the gray lines represent the QTL scans for the AMMI predicted values and the black lines represent the proposed QTL scans based on the R-AMMI predicted values. The horizontal lines correspond to the thresholds based on a permutation test with 1000 permutations, at a significance level of 0.05. These analyses are based on random dataset among the 100 simulations, with 5% scattered contamination with shift point-mass outliers equal to 10. The codes for the environments are as defined above.

Figure A.7: QTL scan for one random realization of the yield data for the pepper population, simulated from the genotype-phenotype model with seven physiological parameters. Each row of charts represents a different level of temperature. The dotted segments represent the QTL scans for the original data, the gray lines represent the QTL scans for the AMMI predicted values and the black lines represent the proposed QTL scans based on the R-AMMI predicted values. The horizontal lines correspond to the thresholds based on a permutation test with 1000 permutations, at a significance level of 0.05. These analyses are based on random dataset among the 100 simulations, with 10% single contamination with shift equal to 4. The codes for the environments are as defined above.

Figure A.8: QTL scan for one random realization of the yield data for the pepper population, simulated from the genotype-phenotype model with seven physiological parameters. Each row of charts represents a different level of temperature. The dotted segments represent the QTL scans for the original data, the gray lines represent the QTL scans for the AMMI predicted values and the black lines represent the proposed QTL scans based on the R-AMMI predicted values. The horizontal lines correspond to the thresholds based on a permutation test with 1000 permutations, at a significance level of 0.05. These analyses are based on random dataset among the 100 simulations, with 10% single contamination with shift equal to 7. The codes for the environments are as defined above.

Figure A.9: QTL scan for one random realization of the yield data for the pepper population, simulated from the genotype-phenotype model with seven physiological parameters. Each row of charts represents a different level of temperature. The dotted segments represent the QTL scans for the original data, the gray lines represent the QTL scans for the AMMI predicted values and the black lines represent the proposed QTL scans based on the R-AMMI predicted values. The horizontal lines correspond to the thresholds based on a permutation test with 1000 permutations, at a significance level of 0.05. These analyses are based on random dataset among the 100 simulations, with 10% single contamination with shift equal to 10. The codes for the environments are as defined above.

Figure A.10: QTL scan for one random realization of the yield data for the pepper population, simulated from the genotype-phenotype model with seven physiological parameters. Each row of charts represents a different level of temperature. The dotted segments represent the QTL scans for the original data, the gray lines represent the QTL scans for the AMMI predicted values and the black lines represent the proposed QTL scans based on the R-AMMI predicted values. The horizontal lines correspond to the thresholds based on a permutation test with 1000 permutations, at a significance level of 0.05. These analyses are based on random dataset among the 100 simulations, with 10% single contamination with shift point-mass outliers equal to 4. The codes for the environments are as defined above.

Figure A.11: QTL scan for one random realization of the yield data for the pepper population, simulated from the genotype-phenotype model with seven physiological parameters. Each row of charts represents a different level of temperature. The dotted segments represent the QTL scans for the original data, the gray lines represent the QTL scans for the AMMI predicted values and the black lines represent the proposed QTL scans based on the R- AMMI predicted values. The horizontal lines correspond to the thresholds based on a permutation test with 1000 permutations, at a significance level of 0.05. These analyses are based on random dataset among the 100 simulations, with 10% single contamination with shift point-mass outliers equal to 7. The codes for the environments are as defined above.

Figure A.12: QTL scan for one random realization of the yield data for the pepper population, simulated from the genotype-phenotype model with seven physiological parameters. Each row of charts represents a different level of temperature. The dotted segments represent the QTL scans for the original data, the gray lines represent the QTL scans for the AMMI predicted values and the black lines represent the proposed QTL scans based on the R-AMMI predicted values. The horizontal lines correspond to the thresholds based on a permutation test with 1000 permutations, at a significance level of 0.05. These analyses are based on random dataset among the 100 simulations, with 10% single contamination with shift point-mass outliers equal to 10. The codes for the environments are as defined above.

Figure A.13: Biplots obtained for a simulated data set with mixed contamination (1st column) and without mixed contamination (2nd column).

Figure A.16: Biplots obtained through models H-AMMI (Rodrigues et al., 2016), R-AMMI and AMMI (columns 1 - 3, respectively) for a randomly simulated data set with 10% single environment contamination with shift and point-mass outliers.

Figure A.17: Boxplots for the number of detected QTLs in all 100 data sets, for the raw data (yellow), AMMI predicted values (orange), weighted AMMI predicted values (red), and robust AMMI predicted values (green). The left hand side plot gives the boxplots for the model parameters and the right hand side plot gives the boxplots for the main clusters of environments. These values are for QTLs detected when considering an interval of 20 cM centred on the exact position of the simulated QTL. These plots are for a heritability of 0.5 in all environments. These results are based on random dataset among the 100 simulations, with 5% scattered contamination with shift equal to 4.

Figure A.18: Boxplots for the number of detected QTLs in all 100 data sets, for the raw data (yellow), AMMI predicted values (orange), weighted AMMI predicted values (red), and robust AMMI predicted values (green). The boxplots are for individual environments, for an expected maximum of seven (environments with temperature of 15°C or 20°C) or eight (environments with temperature of 25°C). These values are for QTLs detected when considering an interval of 20 cM centred on the exact position of the simulated QTL. These plots are for a heritability of 0.5 in all environments. These results are based on random dataset among the 100 simulations, with 5% scattered contamination with shift equal to 4.

Figure A.19: Boxplots for the number of detected QTLs in all 100 data sets, for the raw data (yellow), AMMI predicted values (orange), weighted AMMI predicted values (red), and robust AMMI predicted values (green). The left hand side plot gives the boxplots for the model parameters and the right hand side plot gives the boxplots for the main clusters of environments. These values are for QTLs detected when considering an interval of 20 cM centred on the exact position of the simulated QTL. These plots are for a heritability of 0.5 in all environments. These results are based on random dataset among the 100 simulations, with 5% scattered contamination with shift equal to 7.

Figure A.20: Boxplots for the number of detected QTLs in all 100 data sets, for the raw data (yellow), AMMI predicted values (orange), weighted AMMI predicted values (red), and robust AMMI predicted values (green). The boxplots are for individual environments, for an expected maximum of seven (environments with temperature of 15°C or 20°C) or eight (environments with temperature of 25°C). These values are for QTLs detected when considering an interval of 20 cM centred on the exact position of the simulated QTL. These plots are for a heritability of 0.5 in all environments. These results are based on random dataset among the 100 simulations, with 5% scattered contamination with shift equal to 7.

Figure A.21: Boxplots for the number of detected QTLs in all 100 data sets, for the raw data (yellow), AMMI predicted values (orange), weighted AMMI predicted values (red), and robust AMMI predicted values (green). The left hand side plot gives the boxplots for the model parameters and the right hand side plot gives the boxplots for the main clusters of environments. These values are for QTLs detected when considering an interval of 20 cM centred on the exact position of the simulated QTL. These plots are for a heritability of 0.5 in all environments. These results are based on random dataset among the 100 simulations, with 5% scattered contamination with shift equal to 10.

Figure A.22: Boxplots for the number of detected QTLs in all 100 data sets, for the raw data (yellow), AMMI predicted values (orange), weighted AMMI predicted values (red), and robust AMMI predicted values (green). The boxplots are for individual environments, for an expected maximum of seven (environments with temperature of 15°C or 20°C) or eight (environments with temperature of 25°C). These values are for QTLs detected when considering an interval of 20 cM centred on the exact position of the simulated QTL. These plots are for a heritability of 0.5 in all environments. These results are based on random dataset among the 100 simulations, with 5% scattered contamination with shift equal to 10.

Figure A.23: Boxplots for the number of detected QTLs in all 100 data sets, for the raw data (yellow), AMMI predicted values (orange), weighted AMMI predicted values (red), and robust AMMI predicted values (green). The left hand side plot gives the boxplots for the model parameters and the right hand side plot gives the boxplots for the main clusters of environments. These values are for QTLs detected when considering an interval of 20 cM centred on the exact position of the simulated QTL. These plots are for a heritability of 0.5 in all environments. These results are based on random dataset among the 100 simulations, with 5% scattered contamination with shift point-mass outliers equal to 7.

Figure A.24: Boxplots for the number of detected QTLs in all 100 data sets, for the raw data (yellow), AMMI predicted values (orange), weighted AMMI predicted values (red), and robust AMMI predicted values (green). The boxplots are for individual environments, for an expected maximum of seven (environments with temperature of 15°C or 20°C) or eight (environments with temperature of 25°C). These values are for QTLs detected when considering an interval of 20 cM centred on the exact position of the simulated QTL. These plots are for a heritability of 0.5 in all environments. These results are based on random dataset among the 100 simulations, with 5% scattered contamination with shift point-mass outliers equal to 7.

Figure A.25: Boxplots for the number of detected QTLs in all 100 data sets, for the raw data (yellow), AMMI predicted values (orange), weighted AMMI predicted values (red), and robust AMMI predicted values (green). The left hand side plot gives the boxplots for the model parameters and the right hand side plot gives the boxplots for the main clusters of environments. These values are for QTLs detected when considering an interval of 20 cM centred on the exact position of the simulated QTL. These plots are for a heritability of 0.5 in all environments. These results are based on random dataset among the 100 simulations, with 5% scattered contamination with shift point-mass outliers equal to 10.

Figure A.26: Boxplots for the number of detected QTLs in all 100 data sets, for the raw data (yellow), AMMI predicted values (orange), weighted AMMI predicted values (red), and robust AMMI predicted values (green). The boxplots are for individual environments, for an expected maximum of seven (environments with temperature of 15°C or 20°C) or eight (environments with temperature of 25°C). These values are for QTLs detected when considering an interval of 20 cM centred on the exact position of the simulated QTL. These plots are for a heritability of 0.5 in all environments. These results are based on random dataset among the 100 simulations, with 5% scattered contamination with shift point-mass outliers equal to 10.

Figure A.27: Boxplots for the number of detected QTLs in all 100 data sets, for the raw data (yellow), AMMI predicted values (orange), weighted AMMI predicted values (red), and robust AMMI predicted values (green). The left hand side plot gives the boxplots for the model parameters and the right hand side plot gives the boxplots for the main clusters of environments. These values are for QTLs detected when considering an interval of 20 cM centred on the exact position of the simulated QTL. These plots are for a heritability of 0.5 in all environments. These results are based on random dataset among the 100 simulations, with 10% single contamination with shift equal to 4.

Figure A.28: Boxplots for the number of detected QTLs in all 100 data sets, for the raw data (yellow), AMMI predicted values (orange), weighted AMMI predicted values (red), and robust AMMI predicted values (green). The boxplots are for individual environments, for an expected maximum of seven (environments with temperature of 15°C or 20°C) or eight (environments with temperature of 25°C). These values are for QTLs detected when considering an interval of 20 cM centred on the exact position of the simulated QTL. These plots are for a heritability of 0.5 in all environments. These results are based on random dataset among the 100 simulations, with 10% single contamination with shift equal to 4.

Figure A.29: Boxplots for the number of detected QTLs in all 100 data sets, for the raw data (yellow), AMMI predicted values (orange), weighted AMMI predicted values (red), and robust AMMI predicted values (green). The left hand side plot gives the boxplots for the model parameters and the right hand side plot gives the boxplots for the main clusters of environments. These values are for QTLs detected when considering an interval of 20 cM centred on the exact position of the simulated QTL. These plots are for a heritability of 0.5 in all environments. These results are based on random dataset among the 100 simulations, with 10% single contamination with shift equal to 7.

Figure A.30: Boxplots for the number of detected QTLs in all 100 data sets, for the raw data (yellow), AMMI predicted values (orange), weighted AMMI predicted values (red), and robust AMMI predicted values (green). The boxplots are for individual environments, for an expected maximum of seven (environments with temperature of 15°C or 20°C) or eight (environments with temperature of 25°C). These values are for QTLs detected when considering an interval of 20 cM centred on the exact position of the simulated QTL. These plots are for a heritability of 0.5 in all environments. These results are based on random dataset among the 100 simulations, with 10% single contamination with shift equal to 7.

Figure A.31: Boxplots for the number of detected QTLs in all 100 data sets, for the raw data (yellow), AMMI predicted values (orange), weighted AMMI predicted values (red), and robust AMMI predicted values (green). The left hand side plot gives the boxplots for the model parameters and the right hand side plot gives the boxplots for the main clusters of environments. These values are for QTLs detected when considering an interval of 20 cM centred on the exact position of the simulated QTL. These plots are for a heritability of 0.5 in all environments. These results are based on random dataset among the 100 simulations, with 10% single contamination with shift equal to 10.

Figure A.32: Boxplots for the number of detected QTLs in all 100 data sets, for the raw data (yellow), AMMI predicted values (orange), weighted AMMI predicted values (red), and robust AMMI predicted values (green). The boxplots are for individual environments, for an expected maximum of seven (environments with temperature of 15°C or 20°C) or eight (environments with temperature of 25°C). These values are for QTLs detected when considering an interval of 20 cM centred on the exact position of the simulated QTL. These plots are for a heritability of 0.5 in all environments. These results are based on random dataset among the 100 simulations, with 10% single contamination with shift equal to 10.

Figure A.33: Boxplots for the number of detected QTLs in all 100 data sets, for the raw data (yellow), AMMI predicted values (orange), weighted AMMI predicted values (red), and robust AMMI predicted values (green). The left hand side plot gives the boxplots for the model parameters and the right hand side plot gives the boxplots for the main clusters of environments. These values are for QTLs detected when considering an interval of 20 cM centred on the exact position of the simulated QTL. These plots are for a heritability of 0.5 in all environments. These results are based on random dataset among the 100 simulations, with 10% single contamination with shift point-mass outliers equal to 4.

Figure A.34: Boxplots for the number of detected QTLs in all 100 data sets, for the raw data (yellow), AMMI predicted values (orange), weighted AMMI predicted values (red), and robust AMMI predicted values (green). The boxplots are for individual environments, for an expected maximum of seven (environments with temperature of 15°C or 20°C) or eight (environments with temperature of 25°C). These values are for QTLs detected when considering an interval of 20 cM centred on the exact position of the simulated QTL. These plots are for a heritability of 0.5 in all environments. These results are based on random dataset among the 100 simulations, with 10% single contamination with shift point mass equal to 4.

Figure A.35: Boxplots for the number of detected QTLs in all 100 data sets, for the raw data (yellow), AMMI predicted values (orange), weighted AMMI predicted values (red), and robust AMMI predicted values (green). The left hand side plot gives the boxplots for the model parameters and the right hand side plot gives the boxplots for the main clusters of environments. These values are for QTLs detected when considering an interval of 20 cM centred on the exact position of the simulated QTL. These plots are for a heritability of 0.5 in all environments. These results are based on random dataset among the 100 simulations, with 10% single contamination with shift point-mass outliers equal to 7.

Figure A.36: Boxplots for the number of detected QTLs in all 100 data sets, for the raw data (yellow), AMMI predicted values (orange), weighted AMMI predicted values (red), and robust AMMI predicted values (green). The boxplots are for individual environments, for an expected maximum of seven (environments with temperature of 15°C or 20°C) or eight (environments with temperature of 25°C). These values are for QTLs detected when considering an interval of 20 cM centred on the exact position of the simulated QTL. These plots are for a heritability of 0.5 in all environments. These results are based on random dataset among the 100 simulations, with 10% single contamination with shift point mass equal to 7.

Figure A.37: Boxplots for the number of detected QTLs in all 100 data sets, for the raw data (yellow), AMMI predicted values (orange), weighted AMMI predicted values (red), and robust AMMI predicted values (green). The left hand side plot gives the boxplots for the model parameters and the right hand side plot gives the boxplots for the main clusters of environments. These values are for QTLs detected when considering an interval of 20 cM centred on the exact position of the simulated QTL. These plots are for a heritability of 0.5 in all environments. These results are based on random dataset among the 100 simulations, with 10% single contamination with shift point-mass outliers equal to 10.

Figure A.38: Boxplots for the number of detected QTLs in all 100 data sets, for the raw data (yellow), AMMI predicted values (orange), weighted AMMI predicted values (red), and robust AMMI predicted values (green). The boxplots are for individual environments, for an expected maximum of seven (environments with temperature of 15°C or 20°C) or eight (environments with temperature of 25°C). These values are for QTLs detected when considering an interval of 20 cM centred on the exact position of the simulated QTL. These plots are for a heritability of 0.5 in all environments. These results are based on random dataset among the 100 simulations, with 10% single contamination with shift point mass equal to 10.

Figure A.39: Boxplots for the number of detected QTLs in all 100 data sets, for the raw data (yellow), AMMI predicted values (orange), weighted AMMI predicted values (red), and robust AMMI predicted values (green). The plots gives the boxplots for the model parameters. The data presented in the first line refer to the single contamination and the second line refer to the scattered contamination.