Paulo Jorge Canas Rodrigues

Mestre

New strategies to detect and understand genotype-by-environment interactions and QTL-by-environment interactions

Dissertação para obtenção do Grau de Doutor em
Estatística e Gestão do Risco, especialidade em Estatística

Orientador: Stanislaw Mejza, Full Professor, Poznan University of Life Sciences, Poland
Co-orientador: João Tiago Mexia, Jubilee Full Professor, FCT-UNL, Portugal

Júri:

Presidente: Prof. Doutor Fernando José Pires Santana
Arguente(s): Prof. Doutor Hans-Peter Piepho
Prof. Doutora Ana Maria Nobre Vilhena Pires Parente

Vogais: Prof. Doutora Maria Antónia Amaral Turkman
Prof. Doutor Carlos Manuel Agra Coelho
Prof. Doutor Stanislaw Mejza
Prof. Doutor João Tiago Praça Nunes Mexia

Fevereiro de 2012

Paulo Jorge Canas Rodrigues

Mestre

New strategies to detect and understand genotype-by-environment interactions and QTL-by-environment interactions

Dissertação para obtenção do Grau de Doutor em
Estatística e Gestão do Risco, especialidade em Estatística

Orientador: Stanislaw Mejza, Full Professor, Poznan University of Life Sciences, Poland
Co-orientador: João Tiago Mexia, Jubilee Full Professor, FCT-UNL, Portugal

Júri:

Presidente: Prof. Doutor Fernando José Pires Santana
Arguente(s): Prof. Doutor Hans-Peter Piepho
Prof. Doutora Ana Maria Nobre Vilhena Pires Parente

Vogais: Prof. Doutora Maria Antónia Amaral Turkman
Prof. Doutor Carlos Manuel Agra Coelho
Prof. Doutor Stanislaw Mejza
Prof. Doutor João Tiago Praça Nunes Mexia

FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

Fevereiro de 2012

Acknowledgements

The last four and half years were (mostly) great and challenging! I had the great chance to work with many people and to visit many places during this journey. It is a pleasure to thank now to those who made the end of this thesis possible, either because of their scientific or emotional support.

First of all I would like to thank to my Mentors who share their knowledge with me and contributed greatly for my development as a researcher and as a person:

- to my supervisors Stanislaw Mejza and João Tiago Mexia. Professor Mexia and Professor Mejza were with me since the beginning, and have done everything they could to help me in any way I needed. I'm deeply grateful to them for all the help and unconditional support. I'm very glad I can have them by my side, as friends and collaborators. They are always there!

- to Hugh G. Gauch. Hugh definitively was one of the most important people for me in these years! His contribution to my scientific and personal development was immeasurable. He was great and even without knowing me or asking for any reference he accepted me as his guest in Cornell and shared his office with me. Hugh spent a lot of time with me discussing real science and the reasons why we do that. I'm deeply thankful for having the chance of meeting and really get to know him. Hugh is the best person I have ever met and represents what all people (including researchers) should be like. I have no doubts I'm a better person now because of him.

- to Fred van Eeuwijk. I had the chance to work with Fred in Wageningen for about two years, which really contributed for my development in many ways. We had great meetings where I learned a lot from our discussions and from his great knowledge in this topic. I'm really thankful to Fred for giving me the supervision, to wrap up this thesis and for the great support in the final stage of this thesis. He contributed greatly for a better outcome.

I also would like to thank all my Co-Authors which in one way or in other helped me to finish the chapters of this thesis and improve my knowledge on several topics.

Besides my Mentors and Co-Authors, I had the chance to meet great people and make very good friends I hope to keep. In CMA we had a great environment created by a few good friends: Miguel Fonseca, Miguel de Carvalho, Elsa, Agostinho, Sandra and Vanda. In Poznan I have also found great people I can now call friends, namely Kasia, Ania and Aga. They were great and I felt at home in Poznan. In Wageningen I really enjoyed the company and conversations with Paul, Sabine, Nurudeen, Maria João, Marcos, Alba, Noor and the great "football people" who work in the research group of Organic Farming Systems. I'll definitively miss our matches! I thank all of them and many others not mentioned for the enjoyable time I spend with them!

I must not forget the financial support which allowed me to reach this end. I would like to thank Fundação para a Ciência e Tecnologia (Portuguese Foundation for Science and Technology), of Ministério da Ciência, Tecnologia, e Ensino Superior, Portugal, for my doctoral grant SFRH/BD/35994/2007, which lasted for four years. I would also like to thank the project N N310 447838 supported by Ministry of Science and Higher Education, Poland, for further financial support.

I would also like to thank all my other friends for all the support and for being always there! Thanks to my parents and my sister for their support in a number of ways. Last, but definitely not the least, I owe my deepest gratitude to Ana Teresa for all the unconditional support all the way, all the time!

Resumo

Interação entre genótipo e ambiente (GEI) é frequente em ensaios multi-localização, e traduz-se por diferentes respostas dos genótipos em diferentes ambientes. Com o desenvolvimento das marcas moleculares e técnicas de mapeamento, os investigadores podem analisar todo o genoma para detectar as localizações específicas dos genes que influenciam a característica quantitativa de interesse. Estas localizações são denominadas de *quantitative trait locus* (QTL) e, quando estes QTLs apresentam diferentes respostas em diferentes ambientes, estamos perante interações entre QTL e ambiente (QEI), que é a base da GEI. Uma boa compreensão destas interações permite aos investigadores selecionar melhores genótipos para diferentes condições ambientais e, consequentemente, melhorar colheitas em paises desenvolvidos e, especialmente, em países em desenvolvimento. Nesta tese de doutoramento pretendo apresentar novas estratégias para melhorar a deteção e perceção de QTLs, especialmente QTLs associados a QEI no contexto de ensaios multi-localização, utilizando e fornecendo *open source software*.

Na primeira parte desta tese é apresentada uma comparação entre dois dos métodos mais usados na análise da GEI: a análise conjunta de regressões (JRA) e o modelo de efeitos principais aditivos e interação multiplicativa (AMMI). Esta comparação é realizada em termos de "robustez" com o aumento da proporção de valores omissos, e em termos da obtenção dos genótipos dominates/vencedores. Nos capítulos seguintes são apresentados métodos com duas e três etapas onde os modelos AMMI são usados para aumentar a precisão dos dados fenotípicos, e os respectivos *scores* usados para ordenar os ambientes na procura de padrões ecológicos ou biológicos. A primeira destas abordagens (duas etapas) é apropriada quando a variância do erro é constante ao longo dos ambientes, enquanto a segunda (três etapas) é uma generalização permitindo ter em conta diferenças na variância do erro ao usar o modelo AMMI ponderado (WAMMI, proposto nesta tese). A parte final da tese ilustra uma estratégia para simular e modelar GEI e QEI em características complexas como a produção/rendimento, com base numa série de parâmetros fisiológicos dependendo apenas dos dados genotípicos. Isto é realizado usando um modelo eco-fisiológico de crescimento de colheitas com sete parâmetros dependentes de QTLs.

Palavras chave: Interação entre genótipo e ambiente; Interação entre QTL e ambiente; modelos AMMI; redução da dimensão; redução da dimensão ponderada; modelos de crescimento de colheitas.

Abstract

Genotype-by-environment interaction (GEI) is frequent in multi-environment trials, and represents differential responses of genotypes across environments. With the development of molecular markers and mapping techniques, researchers can go one step further and analyse the whole genome to detect specific locations of genes which influence a quantitative trait such as yield. These locations are called quantitative trait locus (QTL), and when these QTLs have different expression across environments we talk about QTL-by-environment interactions (QEI), which is the base of GEI. Good understandings of these interactions enable researchers to select better genotypes across different environmental conditions and, consequently, to improve crops in developed and developing countries. In this thesis I intend to present new strategies to improve detection and better understanding of QTLs, especially those exhibiting QEI in the context of multi-environment trials, by using and providing open source software.

The first part of this thesis presents a comparison between two of the most used methods to analyse and to structure GEI: the joint regression analysis (JRA) and the additive main effects and multiplicative interaction (AMMI) model. This comparison is made in terms of "robustness" with different incidence rates of missing values, and in terms of dominant/winner genotypes. In the following chapters two- and three-stages approaches are presented in which the AMMI model is used to gain accuracy in the phenotypic data, and their scores used to order the environments to find ecological or biological patterns. The first approach (two stages) is appropriated when the error variance is constant across environments, whereas the second (three stages) is more general and accounts for differences in the error variances by using the proposed weighted AMMI model (WAMMI). The final part of the thesis illustrates a strategy to simulate and to model GEI and QEI in complex traits, with the example of yield, based on a number of physiological parameters purely genotype dependent. This is done by using an eco-physiological genotype-to-phenotype model with seven parameters defined with a simple QTL basis.

Keywords: Genotype-by-environment interaction; QTL-by-environment interaction; AMMI models; Low-rank approximations; Weighted low-rank approximations; Eco-physiological crop growth models.

Table of contents

List of Figures

List of Tables

List of Abbreviations

AMMI – additive main effects and multiplicative interaction

ANOVA – analysis of variance

AQ – AMMI analysis followed by QTL scans

CxC – 'Cayuga' × 'Caledonia'

CIM – composite interval mapping

df – degrees of freedom

DH – doubled haploid

EM – expectation-maximization

FDMC – fruit dry matter content

FTF – fraction to fruits

GEI – genotype-by-environment interactions

GGE – genotype main effect plus genotype by environment interaction

GLI – genotype-by-location interactions

G-P – genotype-to-phenotype

IPC – interaction principal component

JRA – joint regression analysis

LOD – logarithm of odds

LUE – light use efficiency

MET – multi-environment trial

MS – mean square

PCA – principal components analysis

PHS – preharvest sprouting

QEI – QTL-by-environnement interaction

QTL – quantitative trait locus

RMSPD – root mean square predictive difference

S×M – 'Steptoe' × 'Morex'

S/N – signal to noise

SREG – sites regression

SS – sum of squares

SVD – singular value decomposition

WAMMI – weighted additive main effects and multiplicative interaction

WSVD – weighted singular value decomposition

Chapter 1

1. General Introduction

1.1. Introduction

One of the main challenges in statistical genetics is to find superior genotypes over a wide range of agro-ecological conditions and also over a number of years. This is also a challenge for farmers, breeders and geneticists although farmers and breeders have often conflicting interests: breeders want genotype that can be sold everywhere and farmers a genotype adapted to their climate and soil management. To achieve this purpose, multi-environment trials (METs) are conducted in which a series of genotypes is evaluated over environmental conditions and over time. The data from these MET are usually summarized in a two-way table with genotypes in the rows and environments (local/year combinations) in the columns. In the most of these two-way tables it is possible to find differences between genotypes in their phenotype (e.g. yield) stability along environments, i.e. the genotypic and environmental effects are not simply additive and genotype-by-environment interaction (GEI) is present in the data. GEI is defined by the change of genetic ranking of genotypes with the environment, e.g., a genotype that is superior at well watered conditions may yield poorly under dry conditions. The GEI can be expressed either as crossovers, when two different genotypes change in rank order of performance when evaluated in different environments, or inconsistent responses of some genotypes across environments without changes in rank order. The study and understanding of these interactions is a major challenge, in order to improve complex traits (e.g. yield) across environmental gradients.

With the development of molecular markers and mapping techniques, researchers can go one step further and analyse the whole genome to detect specific locations of genes which influence a quantitative trait. These locations are called quantitative trait locus (QTL) and when these QTLs have different expression across environments we talk about QTL-by-environment interactions (QEI), which is the base of GEI. A good understanding of these interactions allows researchers to select better genotypes across different environmental gradients and, consequently, to improve crops for developed and, in particular, for developing countries, based on theirs climate and soil characteristics.

One more step further can be achieved when using computer simulations to "replace" the field experiments. Many studies have been made and many papers written about topics such as "eco-physiological models", "crop growth models" (Spitters, 1990, van Ittersum et al., 2003) and "genotype-to-phenotype (G-P) models" (Chenu et al., 2009). These models allow the use of genetic and environmental characteristics to simulate the behaviour of each genotype in each environmental set-up along the growing season (Rodrigues et al., 2012a, Cooper et al., 2009, Bertin et al., 2010, Letort et al., 2008).

Figure 1.1 shows the number of publications per year, from 1990 to 2011 (October), which included GEI, QEI, QTL or G-P models in the title or abstract plus keywords. The published research about G-P models had a sharp peak in1996 and then decreased to about 20% in 2003, where started to increase almost linearly until now. The number of publication on GEI between 1996 and 2002 didn't change much but, since 2003 it increased so that in 2010 were published 3.5 more papers than in 2002 (210 in 2002; 737 in 2010). The number of publications about QTLs has increased linearly from 1993 until 2008 but since then it seems to be stagnated. With the increase in number of publications about QTLs and about GEI, it would be expectable a relatively high increase of research on QEI. However, there is little research on this topic, only about 1% of all publications about QTLs also focus on QEI (Figure 1.2). Therefore I would expect a sharp growth in the number of publications about QEI soon.

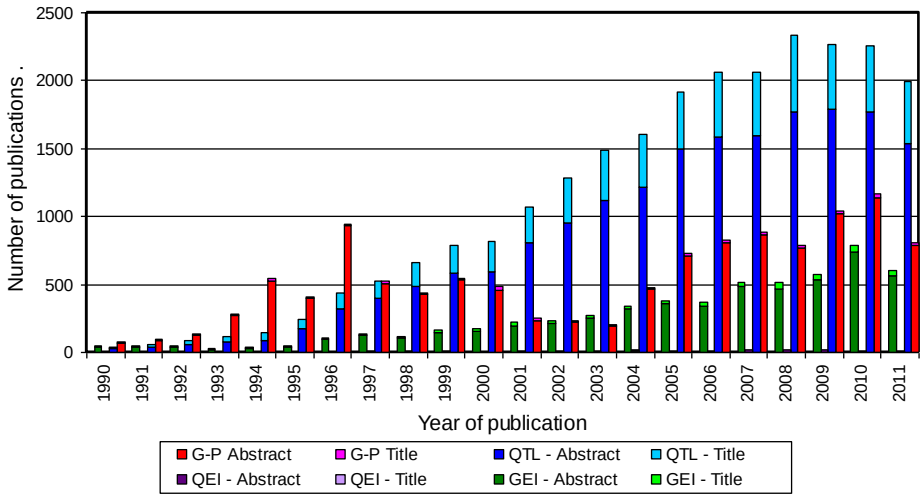

Figure 1.1. Number of publications about GEI, QEI, QTL and G-P models, per year. The information was obtained from the Scopus database in the period 1990-2011. The count for G-P model includes the text "crop growth model", "genotype-to-phenotype model" and "physiological model". These results are very similar to the ones in the ISI Web of Science database.

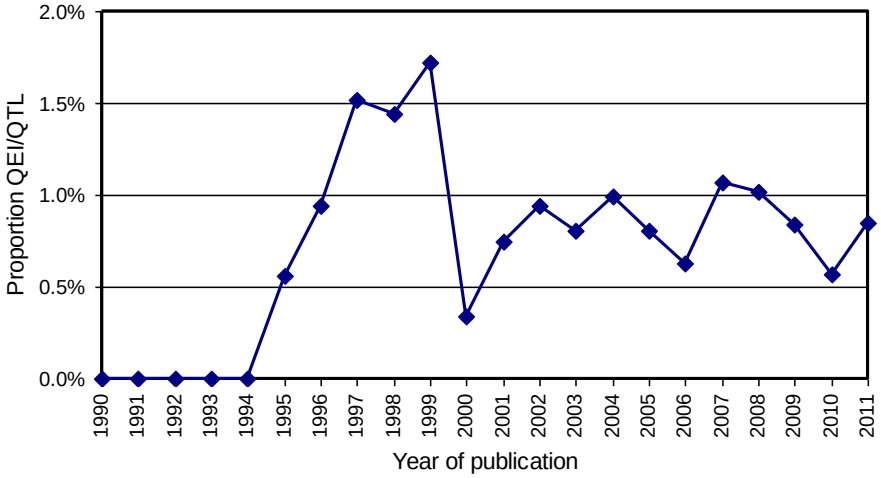

Figure 1.2. Proportion of publications about QEI within the number of publications about QTLs, per year. The information was obtained from the Scopus database in the period 1990-2011.

1.2. Genotype-by-environment interactions – the statistical analysis of two–way tables

To better understand the GEI and QEI, and to make predictions for different environments and/or different years, a wide range of statistical methods have been used. They have been applied to the output of extensive experiments and plant breeding programs conducted under different environmental conditions (or locations) and over several years (van Eeuwijk et al., 2005, Malosetti et al., 2010, Aastveit and Mejza, 1992, Kang and Gauch, 1996).

Figure 1.3 shows the behaviour of the research on GEI by statistical technique used, along time. As in Figure 1.1 we can observe that the amount of research on GEI together with QTL analyses has been almost constant since 2005. Research on regression based techniques continues to increase within GEI analysis and is the most common statistical tool used since 1990. The particular case of factorial regression models represents less than 10% of the total research on regression for GEI. Research articles which use graphical techniques such as biplots (Gabriel, 1971) or genotype main effect plus genotype-by-environment interaction (GGE) biplots (Yan and Kang, 2002) had a sharp increase, especially since 2004. There is also a clear increase for research on singular value decomposition techniques such as principal component analysis (PCA) and additive main effects and multiplicative interaction (AMMI) models (Gauch, 1992). It explains also the steep increase of biplots because PCA and AMMI models also use these graphical representations in their outputs.

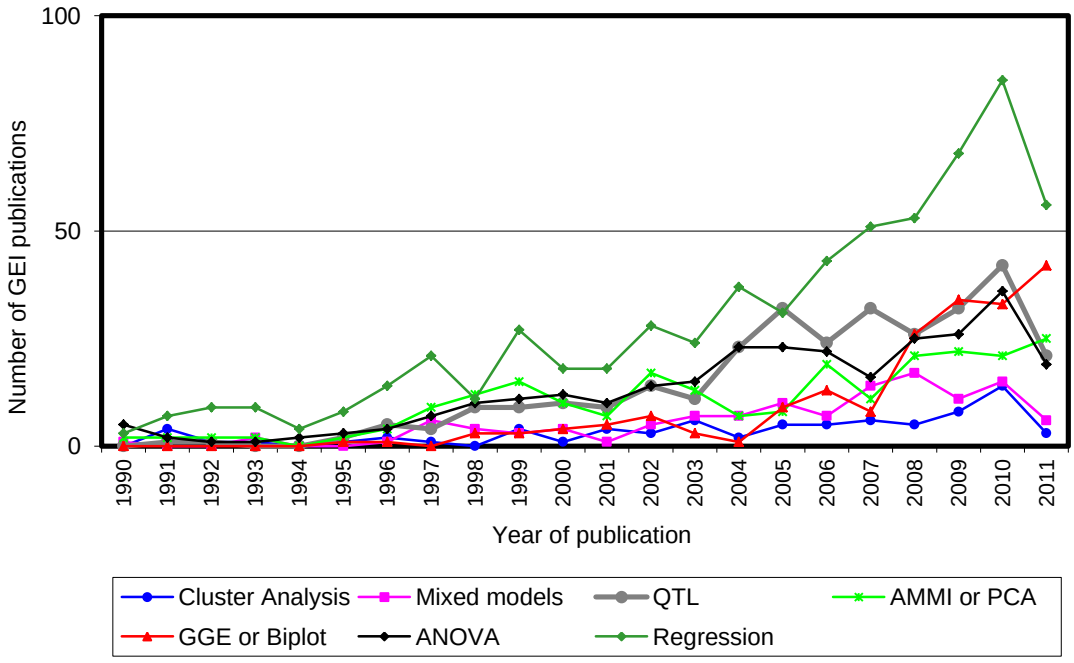

Figure 1.3. Number of publications on research about GEI, per statistical method, per year. The information was obtained from the Scopus database in the period 1990-2011, searching for "genotype environment interaction" together with each of the statistical methods considered.

1.2.1. Statistical models based on regression and singular value decomposition

The simplest model to describe phenotypic observations along environments is the additive model without interaction term. In this case, the expected phenotypic response for genotype $i,\ i = 1, \dots, I,$ in environment $j,\ j = 1, \dots, J,$ equals the grand mean plus the genotype and environment main effects (both expressed as deviations from the grand mean), that is

$$y_{i,j} = \mu + G_i + E_j + \varepsilon_{i,j}. \tag{1.1}$$

The additive model is the base of all the models with interaction, but it is only applicable when there is no GEI in the two-way table with genotypes in the rows and environments in the columns, that is, when the phenotypic response across environments is a set of parallel lines. If there is interaction between genotypes and environments, model (1.1) can be written to account for GEI, that is

$$y_{i,j} = \mu + G_i + E_j + (G.E)_{i,j} + \varepsilon_{i,j}, \tag{1.2}$$

where $(G.E)_{i,j}$ represents the GEI term for genotype *i* and environment j. The full interaction model (1.2) has as many parameters to be estimated as genotype-by-environment combinations, which is associated with tests less precise because the lack of degrees of freedom, and represents a less parsimonious model. An alternative extension of the additive model (1.1) was first proposed by Finlay and Wilkinson (1963), where the phenotypic responses across environments are regressed on the phenotypic mean over environments (a measure of productivity or biological quality in the absence of other environmental characterizations). The GEI is expressed by the *I* slopes β_i and the model can be written as

$$y_{i,j} = \mu + G_i + E_j + \beta_i E_{j.} + \varepsilon_{i,j}. \tag{1.3}$$

Another regression based model was presented by Gusmão (1985) where the (physical) block information is used to correct for spatial effects. In this way the phenotypic responses per block are regressed across environments resulting in $i \times b$ regressions, where b is the number of blocks.

A further alternative to the full interaction model is the additive main effects and multiplicative interaction (AMMI) model (Gollob, 1968, Mandel, 1969, Bradu and Gabriel, 1978, Gauch, 1988, Gauch, 1992), which is more flexible than the Finlay and Wilkinson regression because can partition the interaction in $N = \min(I - 1,\ J - 1)$ terms. It combines the analysis of variance (ANOVA) and principal component analysis (PCA), with ANOVA performed first and then PCA (i.e. the singular value decomposition) applied to the resultant matrix of GEI (Gauch, 1992). The model can be written as

$$\begin{aligned} y_{i,j} &= \mu + G_i + E_j + \sum_{n=1}^{N} \lambda_n \gamma_{i,n} \delta_{j,n} + \varepsilon_{i,j} \\ &= \mu + G_i + E_j + \textstyle\sum_{n=1}^{N} a_{i,n} b_{j,n} + \varepsilon_{i,j}, \end{aligned} \tag{1.4}$$

where λ_n is the singular value for interaction principal component (IPC) n, $\gamma_{i,n}$ is the left singular vector for genotype i in component n, $\delta_{j,n}$ is the right singular vector for environment j in component n, $\varepsilon_{i,j}$ is the residual for genotype *i* in environment j, and N is the number of retained components. A similar alternative is the GGE model (Yan and Kang, 2002) which applies the PCA to the two-way table without the environmental main effects, i.e.

$$y_{i,j} = \mu + E_j + \textstyle\sum_{n=1}^{N} a_{i,n} b_{j,n} + \varepsilon_{i,j}, \tag{1.5}$$

with $a_{i,n}$ and $b_{j,n}$ genotypic and environmental parameters (scores) for the bilinear term n. Both the AMMI and GGE models are more useful when using graphical representations such as biplots (Gabriel, 1971).

1.2.2. The inclusion of environmental and genotypic information in the model

When specific environmental (and/or genotypic) information is available (e.g. rainfall, radiation, temperature, marker information), the advisable linear-bilinear model to be used is the biadditive factorial regression model, also termed as reduced rank factorial regression (Denis, 1988, van Eeuwijk et al., 1996, van Eeuwijk, 1995) because it allows the inclusion of this extra information in the model. Considering the simple case in which the interaction is due to two environmental variables Z_{1j} and Z_{2j}, the model can be written as

$$y_{i,j} = \mu + G_i + E_j + \beta_{1,i}Z_{1,j} + \beta_{2,i}Z_{2,j} + \varepsilon_{i,j,} \quad (1.6)$$

where $\beta_{1,i}$ and $\beta_{2,i}$ are the genotypic sensitivities to the two environmental variables, respectively. This model is an extension of the Finlay-Wilkinson regression (1.3) in which the interaction is written based on several real environmental variables. This allows a physiological interpretation of the GEI in terms of real environmental information. The generalization for the case when H environmental covariates $Z_{1,j}, \dots, Z_{H,j}$ are available is straightforward:

$$y_{i,j} = \mu + G_i + E_j + \sum_{h=1}^{H} \beta_{h,i}Z_{h,j} + \varepsilon_{i,j}. \quad (1.7)$$

A similar expression can be obtained when, besides the H environmental covariates, we also have information about K genotypic covariates (e.g. physiological parameters or marker information). This generalization for H environmental covariates $Z_{1,j}, \dots, Z_{H,j}$, and K genotypic covariates $X_{1,i}, \dots, X_{K,i}$, can be written as:

$$y_{i,j} = \mu + G_i + E_j + \sum_{h=1}^{H} \beta_{h,i}Z_{h,j} + \sum_{k=1}^{K} X_{k,i}\tau_{k,j} + \sum_{k=1}^{K}\sum_{h=1}^{H} \varphi_{k,h}X_{k,i}Z_{h,j} + \varepsilon_{i,j}. \quad (1.8)$$

These regression coefficients are not genotype or environment dependent. The coefficients $\beta_{h,i}$ are genotypic sensitivities to the environmental covariables $Z_{h,j}$, and the $\tau_{k,j}$ denote environmental weighting constants with respect to the genotypic covariable $X_{k,i}$ (Baril et al., 1995). The parameters $\varphi_{k,h}$ represent coefficients with respect to cross-products of genotypic covariables, $X_{k,i}$, and environmental covariables $Z_{h,j}$. Further generalizations are possible depending on the research interests (van Eeuwijk et al., 1996, Romagosa et al., 2009).

1.2.3. Taking into account the variance structure of the data

A more elaborated approach to understand GEI is the mixed model framework (Galwey, 2006, Verbeke and Molenberghs, 2009). This methodology, combines the modelling of the mean and the variance, and provides a powerful tool to analyse GEI. The main advantage of these models is the availability of modelling the heterogeneity of variance across environments and correlations between environments. Unlike the models presented before where all terms, except the residual, are fixed, the mixed linear model (Searle, 1971) provides a framework where the fixed effects can be combined with several random terms. Residual maximum likelihood (REML) (Patterson and Thompson, 1971, Searle et al., 1992) is used to estimate

variances and random parameters. A mixed model for a two-way table indexed by genotypes and environments is

$$y_{i,j} = \mu + G_i + E_j + (G.E)_{i,j} + \varepsilon_{i,j} \quad (1.9)$$

where the model parameters are defined as before. Typically, E_j is fixed and G_i, $(G.E)_{i,j}$ and $\varepsilon_{i,j}$ are random, following a normal distribution with zero mean and a variance specific to the term (Boer et al., 2007, Malosetti et al., 2004).

1.2.4. QTL-by-environment interactions

When dealing with QEI instead of GEI, the described fixed and mixed models can be easily adapted. For example the QTL model with interaction can be written as

$$y_{i,j} = \mu + QTL_i + G_i^* + E_j + (QTL.E)_{i,j} + (G.E)_{i,j}^* + \varepsilon_{i,j}, \quad (1.10)$$

where QTL_i is the QTL main effect, $(QTL.E)_{i,j}$ is the QEI, G_i^* is the genotypic residual, and $(G.E)_{i,j}^*$ is the residual from the interaction. More details on these models and the how to include genetic information such as marker information can be found in van Eeuwijk et al. (2005) and Romagosa et al. (2009).

A major point of interest is whether QEI can be detect for the phenotypic trait of interest and to see whether we could interpret this QEI in terms of underlying QTLs for physiological parameters or molecular markers. The QTL model that we are interested uses explicit marker derived information to describe the GEI in terms of QTLs in their dependence on the environments (i.e. the QEI). The inclusion of this marker information, genetic predictors, allows to test whether the phenotypic trait (e.g. yield) is affected by the DNA at a particular genome position, and whether this effect depends on the environment. A mixed linear model definition following Boer et al. (2007) is

$$\begin{aligned} y_{i,j} &= [\mu + E_j] + [G_i + (G.E)_{i,j}] \\ &= [\mu_j] + [\textstyle\sum_{k=1}^{K} x_{k,i}\alpha_{k,j} + \theta_{i,j}] \end{aligned} \quad (1.11)$$

where μ_j is the intercept for each environment, $x_{k,i}$ is derived from marker genotype information for genotype i, $\alpha_{k,j}$ the QTL allele substitution effect for environment j, K is the total number of QTL underlying $y_{i,j}$ (e.g. yield), and $\theta_{i,j}$ follows a multivariate normal distribution with zero mean vector and a given variance-covariance (VCOV) matrix. The choice of the best VCOV structure can be done by following the procedure described in Malosetti et al. (2004) and Boer et al. (2007).

1.2.5. Eco-physiological genotype-to-phenotype models

All the models described so far are intended to analyse the data after collected in multi-environment trials. That procedure of collecting data is expensive, time-consuming and has limitations regarding the number of genotypes, traits and environmental conditions considered. Simulation tools such as genotype-to-phenotype models have proved to be useful in a better understanding of GEI and QEI (van Eeuwijk et al., 2010).

A physiologically inspired alternative approach for collecting field data is based on physiological crop growth simulation models. Crop growth models represent a class of genotype-to-phenotype (G-P) models

with a prior biological structure (Spitters, 1990, van Ittersum et al., 2003) that can be used to help understanding GEI and QEI (Tardieu, 2003, van Eeuwijk et al., 2005, Letort et al., 2008, Chenu et al., 2009, Cooper et al., 2009, Bertin et al., 2010, van Eeuwijk et al., 2010). These models allow the simulation along the growing season (i.e. every day) of the trait of interest (e.g. yield) and need, as input: (i) genotypic information of the crop at hand, i.e. the genetic map with the position of the markers in the chromosomes and marker information; (ii) information about the physiological parameters of the model for each of the considered genotypes; and (iii) environmental characterizations of the study (i.e. weather, soil, etc.). These models allow the inclusion of genetic information such as previously found QTLs for the trait and/or QTLs for the physiological parameters, which will result in a more parsimonious and meaningful model. A particularly strong point of crop growth models in comparison to more statistical G-P models is that they contain explicit representations of development over time which may be useful in describing GEI (Chenu et al., 2009). Otherwise the "time" would be an extra dimension on the phenotypic observations and harder to collect.

1.3. Objectives and outline of the thesis

Despite the wide range of available references and techniques (as described before) to explore and better understand GEI and QEI, not all of them are available to all breeders and researchers. In some cases, because the statistical methods are too complex to be computationally implemented and applied by non-statisticians. In other cases because, although these complex techniques are already well implemented in statistical packages, the software is commercial and too expensive for developing countries where the statistical improvements are slow to arrive.

One of the goals of this thesis is to propose strategies to improve the detection and understanding of QTLs, especially those exhibiting QEI in the context of METs, using open source software (e.g. QTL Cartographer, Wang et al., 2007; MATMODEL, Gauch, 2007; and R/qtl, Broman and Sen, 2009). One of the strategies described here consist in the two stages AQ analysis, that is, the application of a parsimonious AMMI model (Gauch, 1992) to the phenotypic data in order to gain accuracy, and then use those AMMI predicted values to obtain the QTL scans (Gauch et al., 2011). The possibility of ordering the environments by AMMI scores allows the analysis of patterns with ecological or biological interpretation. Other strategy, a three stages approach, is able to account for differences in error variance across environments. This will be done by using the weighted AMMI model (WAMMI, proposed in this thesis), instead of the standard AMMI model, and to obtain the QTL scans (Gauch et al., 2011) based on the WAMMI scores (Rodrigues et al., 2012b).

A second objective of this thesis is to illustrate a strategy for modelling GEI and QEI in complex traits (e.g. yield), that departs from dissection of a target complex trait in a number of component traits, where each of the component traits is purely genotype dependent. An eco-physiological genotype-to-phenotype model with seven parameters, simulated for a back cross population of pepper (*Capsicum annuum* L.), is considered. The model parameters, i.e. yield components, are defined with a simple QTL basis where the

QTLs are assumed to be in different chromosomes. We show that the QTL associated to the most important parameters, for the trait in study, can be detected in the exact same place where they were allocated during the simulation. These QTL detections were made using only the final phenotypic data and the genetic map with the marker information.

1.3.1. Outline of the thesis

This thesis consists of five papers to be found in Chapters 2–6. Three categories can be distinguished:

- in Chapters 2 and 3 the application and comparison of methods prevails;
- in Chapters 4 and 6 new methodology is presented;
- in Chapter 5 a simulation model is discussed and its outcome analysed with existing and new methodology.

In **Chapter 2** (*A comparison between Joint Regression Analysis and the Additive Main Effects and Multiplicative interaction model: the robustness with increasing amounts of missing data*) the main properties of joint regression analysis (JRA), a model based on the Finlay-Wilkinson regression to analyse multi-environment trials, and of the additive main effects and multiplicative interaction (AMMI) model, are presented. This study compares JRA and the AMMI model with particular focus on robustness with increasing amounts of missing values completely at random.

An application is presented which uses a data set from a breeding program of durum wheat (*Triticum turgidum* L., Durum Group) conducted in Portugal. The two models result in similar dominant cultivars (in JRA) and winner of mega-environments (in AMMI) for the same environments. However, JRA had more stable results with the increase in the incidence rates of missing values.

Chapter 3 (*A comparison between joint regression analysis and the AMMI model: a case study with barley*) compares JRA and AMMI models and evaluates the agreement between the winners of mega-environments obtained from the AMMI analysis and the genotypes in the upper contour of the JRA. An iterative algorithm is used to obtain the environmental indexes for JRA, and standard multiple comparison procedures are adapted for genotype comparison and selection. This study includes three data sets from a spring barley (*Hordeum vulgare* L.) breeding program carried out between 2004 and 2006 in Czech Republic. The results from both techniques are integrated in order to advice plant breeders, farmers and agronomists for better genotype selection and prediction for different years and/or different environments.

In **Chapter 4** (*Two New Strategies for Detecting and Understanding QTL-by-Environment Interactions*) two new strategies for detecting QTLs and understanding QEI are presented. The first is to use a parsimonious AMMI model to gain accuracy for the phenotypic data used in QTL scans, thereby improving QTL detection. The second is to order the environments by AMMI parameters that summarize GEI information

in order to reveal consistent patterns and systematic trends that often have an evident ecological or biological interpretation. These two strategies together are illustrated with two examples: preharvest sprouting scores of a biparental wheat (*Triticum aestivum* L.) population from 14 environments spread over five years, and yield for a doubled haploid barley (*Hordeum vulgare* L.) population tested in 16 environments.

Chapter 5 (*A complex trait with unstable QTLs can follow from component traits with stable QTLs: an illustration by a simulation study in pepper*) illustrates a strategy for modeling of GEI and QEI in complex traits that departs from dissection of a target complex trait in a number of component traits, where each of the component traits is purely genotype dependent. An eco-physiological genotype-to-phenotype model converts the set of genotype specific component traits into the complex target trait by integrating the components with environmental inputs over the duration of the growing season. We developed a seven component eco-physiological model for yield in pepper that simulated for a back cross population yield and yield components, where the yield components were given a simple QTL basis. We demonstrate the viability of our modeling approach for complex traits by a case study in sweet pepper (*Capsicum annuum* L.). We show how credible patterns of GEI and QEI for yield can be simulated from genotype specific yield components with a simple QTL basis.

Chapter 6 (*Weighted AMMI to study genotype-by-environment interaction and QTL-by-environment interaction*) introduces a generalization of AMMI model that accounts for heterogeneity of error variance across environments, the weighted AMMI, or WAMMI. WAMMI is useful for studying GEI as well as QEI. For QEI, we perform an initial analysis by WAMMI, and take the predicted values from this analysis as starting point for QTL analyses per environment. We look at the performance of this strategy in relation to QTL scans on the actual data and AMMI predicted values. We also make a comparison with a full mixed model approach to QTL mapping for multiple-environments. We used two data sets for making comparisons: (i) data from a simulated pepper (*Capsicum annuum*) back cross population using a crop growth model to relate genotypes to phenotypes; and (ii) a doubled-haploid barley (*Hordeum vulgare* L.) population. Our results demonstrate that the QTL scans of the WAMMI predicted values outperform the QTL scans for the actual data and for the AMMI predicted values, being very similar to the QTL mixed model approach, with respect to the number of QTLs detected.

Chapter 7 (*Discussion*) summarizes the results from the preceding chapters and presents a short discussion about the usefulness of the eco-physiological genotype-to-phenotype models when compared with greenhouse experiments.

Chapter 2

2. A comparison between joint regression analysis and the additive main effects and multiplicative interaction model: the robustness with increasing amounts of missing data

Abstract

This chapter joins the main properties of joint regression analysis (JRA), a model based on the Finlay-Wilkinson regression to analyse multi-environment trials, and of the additive main effects and multiplicative interaction (AMMI) model. The study compares JRA and AMMI with particular focus on robustness with increasing amounts of randomly selected missing data. The application is made using a data set from a breeding program of durum wheat (*Triticum turgidum* L., Durum Group) conducted in Portugal. The results of the two models result in similar dominant cultivars (JRA) and winner of mega-environments (AMMI) for the same environments. However, JRA had more stable results with the increase in the incidence rates of missing values.

Published as: Rodrigues, P.C., Pereira, D.G. and Mexia, J.M. (2011). A comparison between joint regression analysis and the additive main effects and multiplicative interaction model: the robustness with increasing amounts of missing data. *Scientia Agricola* 68: 679–686.

2.1. Introduction

Joint regression analysis (JRA) has been widely used in crop sciences, to structure and understand genotype-by-environment interaction (GEI) (Eberhart and Russell, 1966, Finlay and Wilkinson, 1963, Gusmão, 1985, Mooers, 1921, Pereira and Mexia, 2008, Yates and Cochran, 1938, Zheng et al., 2009), and in genetics, to analyse quantitative trait loci (QTL) -by-environment interaction (Emebiri and Moody, 2006, Korol et al., 1998).

In this chapter we are mainly interested in the approach proposed by Gusmão (1985) in which the precision in analysing series of randomized block experiments was highly increased, by considering environmental indexes for individual blocks instead of only one environmental index per environment. In the literature some variants of JRA are also denoted as SREG (Sites Regression) model (Cornelius et al., 1992, Crossa et al., 2002, Setimela et al., 2007).

Williams (1952), Gollob (1968), Mandel (1971), Bradu and Gabriel (1978) and Gauch (1988) have made an important contribution to the development of additive main effects and multiplicative interaction (AMMI) models. These models have been widely used to analyse multi-environment trials (METs) because of their flexibility in allowing the use of several multiplicative terms to explain the GEI.

One of the difficulties in choosing the right tool to analyse METs arises when there are missing values in the two-way table of genotypes and environments. These missing values can be either systematic (Calinski et al., 1992, Denis and Baril, 1992), or selected completely at random in the two-way table.

This chapter brings together the main features of JRA and AMMI models, and compares them for analysing a durum wheat (*Triticum turgidum* L., Durum Group) trial with particular focus on robustness with increasing amounts of random missing data, either missing replications or missing cells (more likely when the proportion of missing values is high). The aim here is not to compare the method's ability to estimate missing values in comparison to real data (Alarcón et al., 2010, Bergamo et al., 2008) but to compare the overall stability when increasing the incidence rate of missing values. An emphasis is made in the comparison between (i) the upper contour of JRA and the mega-environments of the AMMI model; and (ii) the stability of the dominant/winner genotypes across environments. To obtain the results for the JRA we developed an R code, and the MATMODEL software (Gauch and Furnas, 1991) was used to fit the AMMI models.

2.2. Materials and methods

2.2.1. Joint regression analysis

JRA has proven to be an important model for analysing and interpreting the GEI of two-way classified tables and continues to be largely used as a complement of traditional statistical analysis in genetics, plant breeding, and agronomy, for determining yield stability of different genotypes or agronomic treatments across environments (Crossa, 1990). JRA may also be used for the analysis of series of experiments in genotype comparison and selection. This technique is based on the adjustment of a

linear regression, per genotype, of the yield on a synthetic variable measuring productivity, the environmental index.

JRA, when applied to two-way tables obtained from METs, aims to determine the stability of the genotypes or agronomic treatments over a wide range of environmental conditions and to interpret the interaction (non-additivity). Let $y_{i,j}$ be a continuous response variable (usually yield) corresponding to a row factor $i,\ i = 1, \dots, I$ (usually the genotypes), and a column factor $j,\ j = 1, \dots, J$ (usually the environments). The model used for the analysis of METs can be defined as

$$y_{i,j} = \mu + G_i + E_j + (GE)_{i,j} + \varepsilon_{i,j}, \tag{2.1}$$

where μ is the grand mean, G_i and E_j are the genotype and environment main effects, $(GE)_{i,j}$ is the interaction and $\varepsilon_{i,j}$ is the residual. A sub-model of (2.1), aiming at estimating some stability parameters for making comparisons between varieties is given by JRA, and allows us to partitioning the GEI into two parts of interest, i.e.

$$(GE)_{i,j} = b_i E_j + \delta_{i,j}, \tag{2.2}$$

where b_i is a linear regression coefficient for the i-th genotype and $\delta_{i,j}$ a deviation (unexplained GEI) (Freeman, 1973). The JRA model can then be written as

$$\begin{aligned} y_{i,j} &= [\mu + G_i] + [E_j + b_i E_j] + \varepsilon_{i,j} \\ &= [G_i^*] + [b_i^* E_j] + \varepsilon_{i,j}^*, \end{aligned} \tag{2.3}$$

where $\varepsilon_{i,j}$ comprises both the unexplained GEI and the experimental error (Shukla, 1972). We assume fixed genotypic and environmental effects and random residual term.

The model (2.3) used in the present chapter does not take into account the block effects since it uses the blocks as environments, following Gusmão (1985). If an experiment is designed with randomized blocks and the treatments correspond to the J genotypes to be compared, for each block in each design, the environmental index is measured by the average yield. For each of the J genotypes, a linear regression of yield on environmental indexes is adjusted.

2.2.2. L_2 environmental indexes

For convenience, let us consider the joint regression model of the second equation in (2.3), where $G_i^* = \mu + G_i, b_i^* = 1 + b$, $E_j, j = 1, \dots, b$, is the environmental index corresponding to blocks instead of environments, b the number of blocks, $y_{i,j}$ is a continuous response (e.g. yield) for cultivar/genotype i in block j if present, and the pairs $(G_i^*, b_i^*),\ i = 1, \dots, I,$, are the regression coefficients, for the I genotypes.

To obtain the estimates for the regression coefficients and the environmental indexes, the goal function to be minimized should be

$$S\left(\mathbf{G}^{*J}, \mathbf{b}^{*J}, \mathbf{E}^{b}\right) = \sum_{i=1}^{b}\sum_{j=1}^{J} p_{ij}\left(y_{i,j} - G_i^* - b_i^* E_j\right)^2. \tag{2.4}$$

Usually the weight $p_{i,j}$ is 1 [0] when genotype i is present [absent] in block j. These weights may differ from block to block to express differences in representativeness of the blocks and thus we take $p_{i,j} = p_j$

when the *i*-th genotype is present. The main problem in such modeling is how to estimate the parameters. However, the lately proposed so called zigzag algorithm (Pereira and Mexia, 2010) is very efficient in finding the estimates of $\left(G_i^*, b_i^*\right),\ i=1,\ldots,I,$ and $E_j, j=1,\ldots,b$. This zigzag algorithm is an alternating least squares based algorithm (Calinski et al., 1992, Denis and Baril, 1992, Digby, 1979, Gabriel and Zamir, 1979, Gauch and Zobel, 1990). For the complete case, Pereira and Mexia (2010) presented an alternative algorithm, the double minimization algorithm, which converges to the absolute minimum of the goal function (2.4) and is an adaptation of the algorithm first presented by Fisher and Mackenzie (1923). More details on the zigzag and double minimization algorithms can be found in Pereira and Mexia (2010).

2.2.3. The zigzag algorithm

Using the zigzag algorithm the minimization of the loss function (2.4) is carried out iteratively, starting with some initial values for the environmental indexes. Since the choice of these initial values has some effect on the number of iterations it should be made carefully. For the complete case (i.e. all the genotypes are present in each environment) the average yield per block can be a good initial value (Gusmão, 1985). When incomplete blocks are used we have a very convenient situation when α-designs are used. Then as the initial values for environmental indexes one may take the average yields for the corresponding superblock. In the worst case any initial values may be taken (Pereira, 2004), but the choice of values close to the environmental index speeds up the calculation.

After choosing the starting values for environmental indexes, the goal function is minimized with respect to the regression coefficients $(G_i^*,\ b_i^*),\ i=1,\ldots,I$. Then the G_i^* and $b_i^*, i=1,\ldots,I,$ are fixed and new environmental indexes are computed, and so on until the convergence of the algorithm. At the end of each iteration the environmental indexes are rescaled so that its range is kept unchanged. Hence, the iteration procedure is called zigzag algorithm and it may be described as follows:

(i) Calculation of the initial values for the environmental indexes $\mathbf{x}_0^b$, which ranges within the interval

$[a_0, b_0],\ a_0 = Min\{x_{01},\ldots,x_{0b}\}$ *and* $b_0 = Max\{x_{01},\ldots,x_{0b}\}$;

(ii) Minimize the function $S\left(\mathbf{G}^J, \mathbf{b}^J \middle| \mathbf{x}_0^b\right)$ and obtain $\tilde{G}\left(\mathbf{x}_0^b\right)$ and $\tilde{b}\left(\mathbf{x}_0^b\right)$;

(iii) To minimize $S\left(\mathbf{x}^b \middle| \tilde{G}^J\left(\mathbf{x}_0^b\right), \tilde{b}^J\left(\mathbf{x}_0^b\right)\right)$, minimize the functions:

$$h_i\left(x \middle| \mathbf{G}^J, \mathbf{b}^J\right) = \sum_{j=1}^{J} p_{ij}\left(Y_{ij} - \tilde{G}_j - \tilde{b}_j x_i\right)^2,\ i=1,\ldots,b,$$

to obtain the new vector $\mathbf{x}'^b_0$, of the new environmental indexes;

(iv) Standardize the vector of environmental indices to keep unchanged the range. With

$$a'_0 = Min\left\{x'_{01},\ldots,x'_{0b}\right\},\ b'_0 = Max\{x'_{01},\ldots,x'_{0b}\}$$

take

$$x_{1i} = a_0 + \frac{b_0 - a_0}{b'_0 - a'_0}\left(x'_{0i} - a'_0\right);$$

to obtain the vector $\mathbf{x}_1^b$, the new environmental indexes;

(v) Repeat the steps from (ii) to (iv) until successive sums of sums of squares of weighted residuals differ by less than a fixed amount (e.g. 10^{-9}).

2.2.4. Upper contour

When two of the regressions on genotypes intersect it means that one of the genotypes is better for higher environmental indexes while the other is preferable for lower environmental indexes. The intersection of regressions shows more than one genotype with similar performance. The upper contour of the JRA is a concave polygonal (Mexia et al., 1997), constituted by segments of the adjusted regression lines, that contains the higher adjusted yields for the environmental indexes (Figure 2.1). Each of these segments will correspond to a range of variation of the environmental indexes in which the associated genotype will have the maximum adjusted yield (Pereira and Mexia, 2008). These genotypes are called dominant and should be selected. The remaining genotypes should be compared with the dominant to check whether they are dominated on the entire range for the adjusted environmental indexes, $[c, d]$. If so, they can be safely discarded from the breeding program.

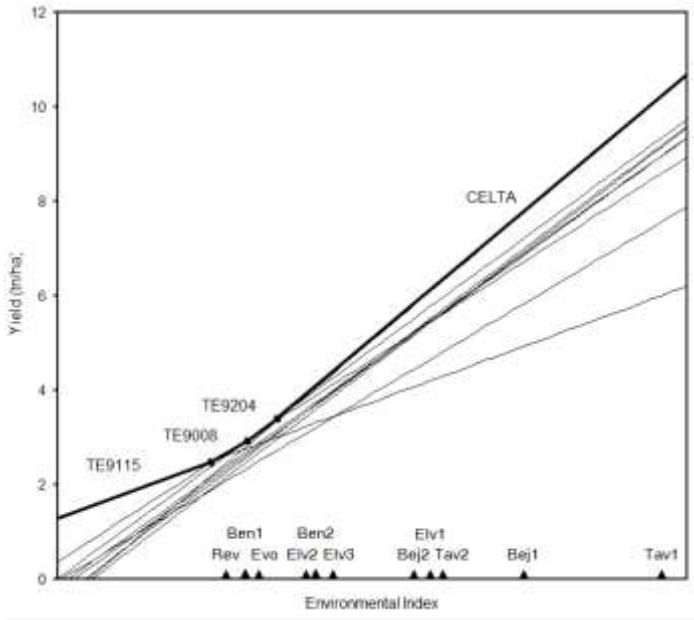

Figure 2.1. Upper contour with the four dominant genotypes in the durum wheat population. The abbreviations for the 11 environments are placed in the axis of the environmental indexes (Bej1: Beja1; Bej2: Beja2; Ben1: Benavila1; Ben2: Benavila2; Evo: Évora; Elv1: Elvas1; Elv2: Elvas2; Elv3: Elvas3; Rev: Revilheira; Tav1: Tavira1; Tav2: Tavira2).

An analogy can be made between Figure 2.1 in this chapter and Figure 2 in Gauch and Zobel (1997), where the AMMI1 nominal yields from a corn trial is depicted as a function of the environment interaction principal component (IPC) axis 1. A more detailed comparison in what concerns the winner genotypes across the environments is presented latter in this chapter.

2.2.5. Genotype comparison and selection

Let L be the number of dominant genotypes with dominant ranges $(c_{i'} = \tilde{\theta}_{i'}, \mathrm{d}_{i'} = \tilde{\theta}_{i'+1}),\ i' = 1,\ldots,L-1$. The entire range for the environmental indexes will be $\left(c = \tilde{\theta}_1\ ,\ \mathrm{d} = \tilde{\theta}_L\right)$. To have interaction between genotypes i and i', and environments there are two possible cases for different slopes, $\tilde{\beta}_i < \tilde{\beta}_{i'}$ and $\tilde{\beta}_i > \tilde{\beta}_{i'}$. After establishing the upper contour, non-dominant genotypes should be compared with the dominant ones. This comparison should be made on the left [right] extreme of the dominance range if the non-dominated genotypes have lower [greater] slope than the dominant one. So, when $\tilde{\beta}_i < \tilde{\beta}_{i'}\ \left[\tilde{\beta}_i > \tilde{\beta}_{i'}\right]$ we are led to compare the adjusted values $\tilde{\alpha}_{\mathrm{i}} + \tilde{\beta}_i x$ and $\tilde{\alpha}_{i'} + \tilde{\beta}_{i'} x$ at the environmental index $\tilde{\theta}_i\ \left[\tilde{\theta}_{i+1}\right]$. These comparisons between slopes may be made using one of the statistical tests: (i) one-sided *t* tests without correction for multiple testing; (ii) Scheffé multiple comparison tests (Scheffé, 1959); (iii) Bonferroni multiple comparison method (Seber and Lee, 2003); (iv) Tukey multiple comparison method (complete case); and (v) Control of False Discovery Rate which is robust against erroneous rejections (Benjamini and Hochberg, 1995). More details of these tests can be found in Pereira and Mexia (2008).

2.2.6. AMMI models

The core idea of the AMMI models is: (i) first apply the additive analysis of the variance model (ANOVA) to a two-way table (in the present case with genotypes and environments); and (ii) secondly apply the multiplicative principal component analysis (PCA) model to the residual from the additive model (in this case to the interaction) (Gauch, 1992). The AMMI model with N multiplicative terms can be written as

$$\mathrm{y}_{\mathrm{i,j}} = \mu + \alpha_{\mathrm{i}} + \beta_{\mathrm{j}} + \sum_{\mathrm{n}=1}^{\mathrm{N}} \lambda_n \gamma_{n,i} \delta_{n,j} + \varepsilon_{\mathrm{i,j}}, \tag{2.5}$$

where $y_{i,j}$ is the yield of genotype i in environment i; μ the grand mean; α_i the genotype mean deviations (the genotype means minus the grand mean); β_j the environment mean deviations; λ_n the singular value for the PCA axis n; $\gamma_{n,i}$ and $\delta_{n,j}$ are the genotype and environment PCA scores for PCA axis n; N is the number of PCA axes retained by the model; and $\varepsilon_{i,j}$ is the residual. If the experiment is replicated, an error term $\epsilon_{i,j,r}$, which is the difference between the $y_{i,j}$ mean and the single observation for replicate r, should be added.

The main purposes of the AMMI models were pointed out by Crossa (1990): (i) model diagnosis (Bradu and Gabriel, 1978); (ii) to clarify GEI (Crossa et al., 1990, Zobel et al., 1988); and (iii) to improve the accuracy of yield estimates (Crossa et al., 1990, Zobel et al., 1988).

2.2.7. Durum wheat yield data

All the properties and comparisons presented in this chapter are illustrated with a data set resulting from a breeding program in Portugal, carried out by the Portuguese National Plant Breeding Station

(ENMP, Elvas) in the years of 1992/1993 and 1993/1994. It contains the yield from nine genotypes (CELTA; HELVIO; TE9006; TE9007; TE9008; TE9110; TE9115; TE9204; and TROVADOR) of durum wheat (*Triticum turgidum* L., Durum Group), measured in 11 environments (Benavila1; Revilheira; Évora; Elvas1; Beja1; Tavira1; Elvas2; Tavira2; Elvas3; Benavila2 and Beja2), and performed in complete randomized blocks with four replicates. These environments were obtained in two years, the first 6 in the first and the second 5 in the second year. Only the locations Tavira, Benavila and Beja were the same in both years. All the locations in this data set are in south Portugal, Tavira being at the sea side (Algarve) while the remaining in the inland (Alentejo). More details about this data set can be found in Pereira and Mexia (2010).

2.2.8. Simulation of missing values

Since the plants may be destroyed by animals, floods or during the harvest, and the yield measurements may be erroneously performed and inadequately introduced in the data base, missing values are common in agricultural experiments. When dealing with missing values researchers should decide between: (i) find a good tool to estimate the missing values (Alarcón et al., 2010, Bergamo et al., 2008), or (ii) chose a robust technique against missing observations to perform the analysis. In the present study we will be interested in the second approach, namely to compare the robustness of JRA and AMMI with the increasing of missing data. Our interest here is to study the case where the missing values were selected "completely" at random, instead of having systematic patterns (Calinski et al., 1992, Denis and Baril, 1992). Our simulation procedure can be summarized in the following steps:

(i) Choose the incidence rate of missing values α (e.g. $\alpha = 5, 10, 25, 50, 75\%$);

(ii) Remove, "completely" at random, $\alpha\%$ of the two-way table with genotypes and environments, leaving at least one observation in each environment and in each genotype;

(iii) **a.** Use the zigzag algorithm (Pereira and Mexia, 2010) to compute the regression coefficients and the L_2 environmental indexes for JRA by minimizing the loss function (2.4); Results such as those shown in Figure 2.1 and in Table 2.3 can be obtained using the appropriated multiple comparison tests mentioned above. **b.** Use the MATMODEL software (Gauch and Furnas, 1991) to estimate the missing values. Results such as those shown in Table 2.3 can be obtained by this software.

(iv) Repeat (ii) and (iii) n times for each incidence rate of missing values. The number of interactions n should be chosen based on the size of the original two-way table. In this particular case we used $n = 100$.

For higher incidence rates of missing values it is more likely that not only replications are missing, but cells (means). In this case an Expectation-Maximization (EM) algorithm provides an effective general strategy for obtaining maximum likelihood estimates (Gauch, 1992). This procedure has been adapted for AMMI and is called EM-AMMI (Gauch and Zobel, 1990), and is implemented in the MATMODEL software (Gauch and Furnas, 1991).

2.3. Results and discussion

2.3.1. A comparison between the algorithms and the alternative methods

This subsection presents a comparison between the two algorithms mentioned in the above section - (i) zigzag algorithm (Pereira and Mexia, 2010) and (ii) double minimization algorithm (Pereira and Mexia, 2010); and the two methods based in the joint regression model - (iii) the regression analysis of the mean yield of individual genotypes on the overall mean of the trial (Finlay and Wilkinson, 1963), and (iv) the regression analysis of the genotype mean yield on block mean, proposed by Gusmão (1985). This comparison is illustrated with a numerical example using the durum wheat yield population. Estimates of intercept, slope and the coefficients of determination obtained from the Finlay and Wilkinson (1963) and Gusmão (1985) methods, and the zigzag and double minimization algorithms are presented in Table 2.1.

Table 2.1. Adjusted regression coefficients and coefficients of determination, as evaluated by the two procedures and two algorithms.

	Finlay and Wilkinson (1963)			Gusmão (1985)			Zigzag and Double Minimization		
Genotype	Intercept	Slope	R^2	Intercept	Slope	R^2	Intercept	Slope	R^2
CELTA	-0.518	1.239	0.893	-0.472	1.229	0.907	-0.544	1.245	0.918
TE9007	-0.542	1.121	0.907	-0.492	1.110	0.918	-0.544	1.121	0.924
TE9006	-0.300	1.086	0.815	-0.361	1.100	0.863	-0.416	1.112	0.870
TE9204	0.077	1.067	0.861	0.058	1.071	0.895	0.016	1.080	0.899
HELVIO	-0.130	1.051	0.902	-0.112	1.047	0.924	-0.244	1.065	0.894
TROVADOR	-0.140	1.042	0.841	-0.206	1.056	0.892	-0.154	1.056	0.928
TE9008	0.375	0.951	0.883	0.403	0.945	0.900	0.376	0.951	0.899
TE9110	-0.089	0.892	0.773	-0.051	0.884	0.783	-0.037	0.880	0.767
TE9115	1.268	0.551	0.510	1.232	0.559	0.542	1.297	0.545	0.507

To compare these four procedures it is important to analyze the slopes and coefficients of determination. They produced almost the same results regarding the ordering of the genotypes per slope (only the Gusmão's method gave a small difference). The coefficients of determination are mainly similar, the zigzag and Double Minimization algorithms being lower than Gusmão (1985) only for three environments (HELVIO, TE9110 and TE9115). Moreover, the zigzag and double minimization have completely agreed and may be seen as the most suited for regression analysis of complete randomized blocks because of their convergence to the minimum of the loss function (2.4).

Another comparison can be made regarding the sums of the sums of squares of residuals for the two procedures and two algorithms (Table 2.2). Here the advantage of the zigzag and double minimization algorithms over the two other procedures is evident since the algorithms induce lower sums of the sums of squares of residuals. This result is true for all the examples and the mathematical proof can be found in Pereira and Mexia (2010). If we compute the pairwise Pearson correlations between the environmental indexes for the four alternatives in Table 2.2, we conclude that all the obtained environmental indexes are highly correlated (minimum of 0.984). In particular, the results obtained using the zigzag and double minimization algorithms have a coefficient of correlation of 1.000 since they completely agree with each other, and they are slightly better than the Finlay and Wilkinson (1963) and Gusmão (1985) approaches. In

the case of a comparison using α-designs or incomplete blocks (instead of the randomized complete block design) some advantage within the two algorithms could be presented better (Pereira and Mexia, 2010).

Table 2.2. Sums of the sums of squares of residuals, as evaluated by the two procedures and two algorithms.

Finlay and Wilkinson	Gusmão	Zigzag and Double Minimization
249.5	207.3	205.5

2.3.2. Genotype comparison and selection

The results for some of the multiple comparison tests mentioned above can be found in Table 2.3. The graphical representation of the dominant genotypes, together with the ranges of dominance (i.e. the lower and upper bound for the interval where the each genotype is dominant) and environments where that dominance occurs, is depicted in Figure 2.1. The bounds of the environmental indexes 2.21 and 8.84 (Table 2.3, complete data) are kept unchanged by the zigzag algorithm and correspond to the lowest and highest mean yield of all the blocks.

2.3.3. AMMI preliminary analyses

Table 2.4 gives the ANOVA for AMMI4. The genotypes, environments and GEI account for 4.1%, 86.4%, and 9.5% of the treatment sum of squares (SS). The noise in the GEI may be estimated by the interaction df times the error MS, namely 40.80, which by difference from the total of 141.74 (total GEI SS) implies a GEI signal SS of 100.94, or 71.21% (Gauch, 1992). Figure 2.2 shows the numbers of indirect replications for the AMMI model family from AMMI0 to AMMI8. The models are less parsimonious, or more complex, moving to the right. AMMI2 achieves the highest number of indirect replications of 1.66 (i.e. 1 replication gives 1.66 more information when considering the parsimonious AMMI2 model). To the left of this figure, excessively simple models underfit the real signal, whereas to the right, excessively complex models overfit the spurious noise. This relationship between accuracy and parsimony has been named as Ockham's hill (Gauch, 2006, MacKay, 1992).

Since the signal is much simpler than the noise, the signal is extracted selectively in early model parameters whereas noise is extracted selectively in late model parameters. A parsimonious model, which captures the most of the signal and discards most of the noise, can be chosen by stopping at the right point (Gauch, 1992). From Table 2.4 it is possible to obtain the SS of the GEI signal of 100.94 ("total GEI SS" minus "noise in GEI") and the SS for the first two PCs together of 115.05 (77.04 for IPC1 and 38.01 for IPC2), which means that these two PCs are mostly signal whereas the remaining are mostly noise. The *F* tests in Table 2.4 also suggested retaining the first two PCs. For comparison with AMMI, the Finlay-Wilkinson linear regressions on the environmental means capture a SS of 43.63, which is about 56.6% of the GEI SS captured by IPC1.

Table 2.3. Dominant and number of significantly dominated genotypes for JRA, environments where the genotypes were dominant (JRA) and where the genotypes were winners (AMMI). The results are for the complete data set and the incidence rates of missing values, and based on one run (out of 100) of the simulation described above. Abbreviations for the environments: Bej1: Beja1; Bej2: Beja2; Ben1: Benavila1; Ben2: Benavila2; Evo: Évora; Elv1: Elvas1; Elv2: Elvas2; Elv3: Elvas3; Rev: Revilheira; Tav1: Tavira1; Tav2: Tavira2.

	JRA		JRA				JRA	AMMI
	Dominant or Winner genotype	Range of dominance	Number of significantly dominated genotypes					
			t test*	t test**	Scheffé*	Bonferroni*	Environments	Environments
Complete data	TE9115	[2.21; 2.27]	3	0	0	0		Ben2
	TE9008	[2.27; 2.80]	2	0	0	0	Rev, Ben1, Evo	Rev
	TE9204	[2.80; 3.40]	3	1	0	1		
	CELTA	[3.40; 8.84]	4	2	0	2	Elv1, Bej1, Tav1, Elv2, Tav2, Elv3, Ben2, Bej2	Ben1, Evo, Elv1, Bej1, Tav1, Elv2, Tav2, Elv3, Bej2
5% of missing values	TE9115	[2.21; 2.38]	2	0	0	0		Rev, Ben2
	TE9008	[2.38; 2.60]	2	0	0	0	Rev	
	TE9204	[2.60; 3.48]	3	1	0	0	Ben1, Evo	
	CELTA	[3.48; 8.88]	4	2	1	2	Elv1, Bej1, Tav1, Elv2, Tav2, Elv3, Ben2, Bej2	Ben1, Evo, Elv1, Bej1, Tav1, Elv2, Tav2, Elv3, Bej2
10% of missing values	TE9008	[2.22; 3.17]	4	1	0	1	Ben1, Rev, Evo	
	TE9204	[3.17; 3.64]	5	1	1	1	Elv2, Ben2	Ben1, Rev, Bej2
	CELTA	[3.64; 9.47]	5	2	1	2	Elv1, Bej1, Tav1, Tav2, Elv3, Bej2	Evo, Elv1, Bej1, Tav1, Elv2, Tav2, Elv3, Ben2
25% of missing values	TE9115	[2.09; 2.10]	5	3	0	2		
	TE9008	[2.10; 3.17]	5	3	0	2	Ben1, Rev, Evo	
	TE9204	[3.17; 3.75]	6	4	2	3	Elv2, Elv3, Ben2	Ben1, Rev, Evo, Bej1, Elv3, Ben2, Bej2
	CELTA	[3.75; 8.77]	6	5	2	4	Elv1, Bej1, Tav1, Tav2, Bej2	Elv1, Tav1, Elv2, Tav2
50% of missing values	TE9115	[2.07; 2.09]	5	4	1	2		
	TE9008	[2.09; 3.16]	2	2	1	2	Ben1, Rev, Evo	
	TE9204	[3.16; 3.85]	3	3	2	3	Elv2, Elv3, Ben2	Ben1, Rev, Evo, Bej1, Ben2, Bej2
	CELTA	[3.85; 9.21]	3	3	2	3	Elv1, Bej1, Tav1, Tav2, Bej2	Elv1, Tav1, Elv2, Tav2, Elv3
75% of missing values	TE9115							Elv3
	TE9204	[1.52; 3.47]	8	8	8	8	Ben1, Rev, Evo, Ben2	
	CELTA	[3.47; 9.10]	8	8	8	8	Elv1, Bej1, Tav1, Elv2, Tav2, Elv3, Bej2	Ben1, Rev, Evo, Elv1, Bej1, Tav1, Elv2, Tav2, Ben2, Bej2

*0.05; **0.01

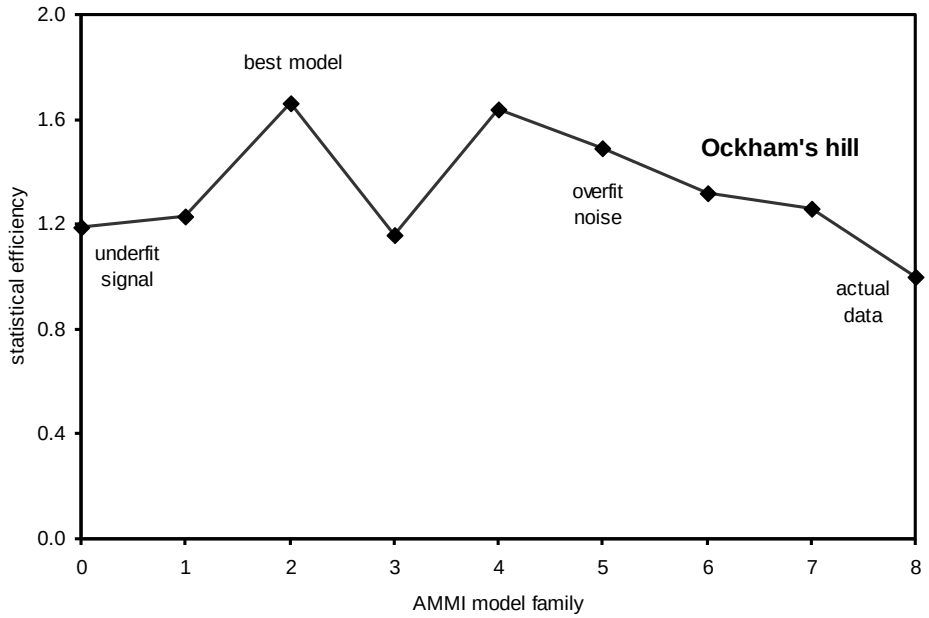

Figure 2.2. Ockham's hill for accuracy of the yield estimates for the durum wheat experiment. The abscissa shows AMMI models of increasing complexity from AMMI0 to AMMI8, and the ordinate shows the number of indirect replications determined by jackknife resampling (e.g. the parsimonious AMMI2 model extract 1.66 time more information than the full AMMI8 model).

Table 2.4. AMMI4 analysis of variance. The grand mean is 4.502 t ha^{-1}.

Source	df	SS	MS	*p*-value*
Total	395	1648.74	4.174	
TRT	98	1497.37	15.279	< 0.001
GEN	8	61.35	7.669	< 0.001
ENV	10	1294.27	129.427	< 0.001
GEI	80	141.74	1.772	< 0.001
IPC 1	17	77.04	4.532	< 0.001
IPC 2	15	38.01	2.534	< 0.001
IPC 3	13	10.79	0.830	0.076
IPC 4	11	10.15	0.923	0.052
Residual	24	5.76	0.240	0.985
Error	297	151.37	0.510	

*Based on F tests. df = degrees of freedom, SS = sum of squares, MS = mean square, TRT = treatments, GEN = genotypes, ENV = environments, GEI = genotype-by-environment interaction, IPC = interaction principal component.

Figure 2.3 depicts the AMMI1 biplot for the durum wheat experiment. The choice of the AMMI1 biplot instead of AMMI2 was made to allow the comparison with Figure 2.1. The abscissa shows the main effects and the ordinate shows the IPC1 scores. The 9 genotypes are represented in bold font and the 11 environments in normal font. The first IPC captures 54.73% (77.04/141.74) of the GEI sum of squares. But, since this GEI is only 71.23% (100.94/141.74) signal, this graph captures the most of GEI signal and a small amount of noise (Gauch, 1992). With this biplot it is easier to understand the association between genotypes and environments where they perform better regarding grain yield.

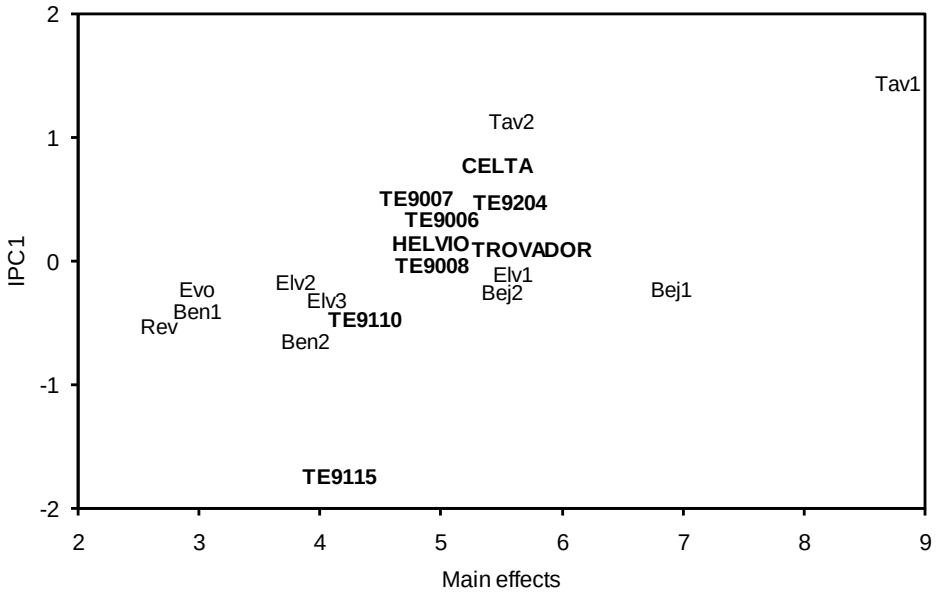

Figure 2.3. AMMI1 biplot for the durum wheat experiment. Bold font represents the codes of the genotypes and plain text the abbreviations for the environments (Bej1: Beja1; Bej2: Beja2; Ben1: Benavila1; Ben2: Benavila2; Evo: Évora; Elv1: Elvas1; Elv2: Elvas2; Elv3: Elvas3; Rev: Revilheira; Tav1: Tavira1; Tav2: Tavira2).

IPC1 makes a distinction between Tavira (Algarve, sea side) and the rest of the environments (Alentejo, inland) (Figure 2.3). When comparing with Figure 2.1, we can see that the four dominant genotypes are ordered by IPC1 scores in Figure 2.3. This provides an agreement between the environmental indexes and IPC1 scores, and connects them to a measure of yield production. The order of environments along the main effects of Figure 2.3 and environmental indexes of Figure 2.1 is the same, as expected.

2.3.4. Upper contour and mega-environments

In this subsection we intend to make a comparison between the upper contour of JRA and the AMMI mega-environments (Gauch and Zobel, 1997). Figure 2.1 shows the 11 environments placed in the axis of the environmental indexes. The first three environments, namely Rev, Ben1 and Evo, have higher yield with the genotype TE9008, and the remaining eight environments have better production with the genotype CELTA. Following the same analysis using the AMMI mega-environments as Gauch and Zobel (1997), based on AMMI1 estimates, we may conclude that this data set has three winners: (i) CELTA wins in nine environments; (ii) TE9008 wins in the environment Rev; and (iii) TE9115 wins in the environment Ben2. However the main conclusion is taken by both analyses: CELTA is the universal winner (Table 2.3).

2.3.5. Stability with missing values

Pereira et al. (2007) concluded that JRA is an extremely robust technique against missing observations in what concerns genotype comparison and selection. They used a series of 17 experiments of α−designs of winter rye genotypes, in the years of 1997 and 1998, and considered proportions of missing values from 5% to 75%, with step size of 5% generated randomly in triplicate. The durum wheat data set was used here to test the stability and agreement in choosing the dominant genotypes for different incidence rates of missing values, between JRA and AMMI. Table 2.3 presents the main results for different incidence rates of missing values. The missing values were chosen randomly as described before.

The analysis of Table 2.3 should be performed between methods and between incidence rates of missing values. Regarding the comparison between methods, the most similar results are for the complete data without missing values, with eight environments having higher yield for the same (dominant/winner) genotypes. The number of environments dominated/won by the same genotypes decreases when increasing the proportion of missing values. The only exception is the case with 75% of missing values, with 6 agreements between analyses, which is more likely to change each time the random procedure to remove observations, is run.

Regarding the comparison between percentages of missing values, Table 2.3 (second, eighth and ninth columns) illustrates a more stable and robust performance of JRA, since the dominant genotypes are kept unchanged for an incidence of missing values until 50%. While for JRA there are six environments (Rev, Elv1, Bej1, Tav1, Tav2 and Bej2) which are dominated by the same genotypes in all the cases (with exception of the extreme 75% incidence rate of missing values), for the AMMI analysis it only happens in 4 environments (all of them are won by CELTA). Moreover for the AMMI model the genotype TE9008 and TE9115 only win in one of the five cases (incidence rates), while for the JRA the dominant genotypes are more stable.

Although the dominant genotypes have little change with the incidence rate of missing values it seems clear that CELTA is the strongest genotype regarding the yield production. It is always dominant for higher environmental indexes and always wins one mega-environment. With 75% of missing values (297 out of 396 observations) the JRA yet identifies two of the dominant genotypes presented in the upper contour of Figure 2.1, while AMMI identifies a "small" mega-environment Elv3 and a larger environment with the remaining ten mega-environments (Table 2.3).

We carried out 100 simulations as described before, and Table 2.3 shows the results for one of them chosen randomly. The 100 data sets for each proportion of missing values resulted in the identification of, at least, one dominant/winner genotype coincident to the complete data set when considering 75% of missing values. For 50% of missing values or less, JRA always identified TE9008 and CELTA as dominant genotypes, whereas TE9204 (not dominant/winner in the complete data set) and CELTA almost always win one AMMI mega-environment. A detailed summary of the 100 runs is presented in Table 2.5.

Table 2.5. Proportion of runs in which dominant genotypes (JRA) and winners of mega-environments (AMMI) are common to the results of the original data.

Proportion of missing values	Dominant or Winner genotype	JRA	AMMI
%		---------- % ----------	
5%	TE9115	7	28
	TE9008	100	47
	TE9204	78	93
	CELTA	100	100
10%	TE9115	8	14
	TE9008	100	71
	TE9204	56	98
	CELTA	100	100
25%	TE9115	12	9
	TE9008	100	62
	TE9204	72	100
	CELTA	100	100
50%	TE9115	21	15
	TE9008	100	36
	TE9204	43	94
	CELTA	100	100
75%	TE9115	3	34
	TE9008	19	29
	TE9204	84	41
	CELTA	98	100

2.4. Conclusion

The aim was not to compute estimates of missing values and compare them with the original data, but to compare the final results (i.e. dominant/winner genotypes and environments where they were dominant/winner) between JRA and AMMI and between the complete data and incomplete data sets with different incidence rates of missing values. The main conclusions were the similarity between the dominant genotypes in JRA and the winners of the mega-environments in the AMMI analysis; and a more stable performance of JRA for higher proportions of missing values. The results from JRA tend to be more significant than those from AMMI models in these kind of trials, because the genotypes in the program have proved to have strong adaptability. Further simulation studies should be done to access these results. However the literature favors AMMI models over JRA because it automatically captures more GEI.

Chapter 3

3. A comparison between joint regression analysis and the AMMI model: a case study with barley

Abstract

Joint Regression Analysis (JRA) and Additive Main effects and Multiplicative Interaction (AMMI) models are compared in order to (i) access the ability of describing genotype-by-environment interaction effects and (ii) evaluate the agreement between the winners of mega-environments obtained from the AMMI analysis and the genotypes in the upper contour of the JRA. An iterative algorithm is used to obtain the environmental indexes for JRA, and standard multiple comparison procedures are adapted for genotype comparison and selection. This study includes three data sets from a spring barley (*Hordeum vulgare* L.) breeding program carried out between 2004 and 2006 in Czech Republic. The results from both techniques are integrated in order to advice plant breeders, farmers and agronomists for better genotype selection and prediction for new years and/or new environments.

Published as: Pereira, D.G.*, Rodrigues, P.C.*, Mejza, S. and Mexia, J.T. (2011). A comparison between joint regression analysis and the AMMI model: a case study with barley. *Journal of Statistical Computation and Simulation* 82: 193-207. DOI: 10.1080/00949655.2011.615839. The original paper can be found online in: http://www.tandfonline.com/doi/abs/10.1080/00949655.2011.615839.

*These authors contributed equally to this work.

3.1. Introduction

The change of genetic ranking of genotypes with the environment (local/year combinations) is known as genotype-by-environment interaction (GEI) (Kang and Gauch, 1996). This interaction can be due to contrasting drought stress levels, winter low temperature stress, abiotic stresses, growing cycle duration, availability of nutrients, etc. The GEI can be expressed either as crossovers, when two different genotypes change in rank order of performance when evaluated in different environments, or as inconsistent responses of some genotypes across environments without changes in rank order. The study and understanding of these interactions are a major challenge for breeders and agronomic researchers, in order to improve complex traits (e.g. yield) across environmental conditions.

Two of the most widely used techniques to structure and understand GEI are the Joint Regression Analysis (JRA) (Finlay and Wilkinson, 1963) and the Additive Main Effects and Multiplicative Interaction (AMMI) models (Gauch, 1992).

JRA may be used for the analysis of a series of experiments concerning genotype comparison and selection. After selecting the variable of interest (e.g. yield), the joint regression model adjusts a linear regression per genotype across all the environments on a synthetic variable measuring productivity, the environmental index. Many variants of JRA were developed along the time. The one we are interested in this chapter was proposed by Gusmão (1985), who showed that the precision in analysing a series of randomized block experiments was highly increased by considering environment indexes for individual blocks instead of only one environmental index per environment. Mexia et al. (1999) proposed an original numerical algorithm (zigzag algorithm) which leads to the best linear unbiased estimators of the joint regression parameters. They introduced the L_2 environmental indexes obtained by minimizing the sum of sums of squares of residuals in order of both the coefficients of the regressions and to the environmental indexes. An upper contour can then be defined by the adjusted regression lines which can be used to carry out genotype selection (Mexia et al., 1997), and genotype comparison when well-articulated with appropriated multiple comparison procedures. The genotypes whose regression lines partake of the upper contour are called dominant, while the remaining are compared with them using multiple comparison tests.

The AMMI model is the most well known and most widely used linear-bilinear model (Gauch, 1988, Gauch, 1992). It first applies the additive analysis of variance (ANOVA) model to a two-way table, and then the multiplicative principal component analysis (PCA) model to the residual from the ANOVA, that is, to the interaction. A remarkable achievement for the utility and success of AMMI models is its ability to build mega-environments (Gauch and Zobel, 1997), that is, groups of environments with a similar response to the variable of interest (e.g. yield) and to deal with biplot graphs that are very useful to delineate mega-environments. A useful role of mega-environments is the possibility of a more reliable prediction for new years and new environments with similar environmental conditions as in a given mega-environment based on only one year of wide testing.

In this chapter, we aim to integrate the results from JRA and AMMI to better structure and understand the GEI. A comparison is made between the techniques using a data set from a multi-environment breeding program of spring barley (*Hordeum vulgare L.*) carried out between 2004 and 2006 in Brno, Czech Republic.

The main objectives of this study were: (i) to present multiple comparison tests for genotype comparison and selection in JRA; (ii) to use JRA and AMMI to access genotype performance, comparing and analysing the results from the JRA's upper contour and AMMI's mega-environments; (iii) to infer whether the conclusion of these statistical analyses were in agreement with the decisions made by the local management team of the plant breeding program, for example, when deciding to add or remove genotypes from the breeding program; and (iv) to integrate the results from JRA and AMMI to structure and understand the GEI in order to advice breeders for better genotype selection and prediction for new years and/or new environments.

3.2. Materials and methods

3.2.1. Joint regression analysis

For convenience, let us consider the data arranged in a two-way table with b rows and J columns. Suppose $y_{i,j}$ is a continuous response variable (e.g. yield) for genotype j in block i if present. The joint regression model is:

$$y_{i,j} = \alpha_j + \beta_j x_i + \varepsilon_{i,j} \quad (i = 1,...,b;\ j = 1,...,J), \tag{3.1}$$

with α_j and β_j being the regression coefficients for the J genotypes and x_i being the block environmental indexes. These environmental indexes represent the averages over block/superblock and can be considered as a (spatial) measure of productivity.

The goal function to be minimized will be

$$S\left(\alpha^J, \beta^J, x^b\right) = \sum_{i=1}^{b}\sum_{j=1}^{J} p_{i,j}\left(y_{i,j} - \alpha_j - \beta_j x_i\right)^2. \tag{3.2}$$

Usually, the weight $p_{i,j}$ is 1 [0] when genotype j is present [absent] from block i. When the genotype is present we take $p_{i,j} = p_i$. These weights may differ from block to block to express differences in the representativeness of the blocks. The main problem in such modelling is as to how to estimate the parameters. One can observe that the so-called zigzag algorithm (Mexia et al., 1999) is very efficient in finding the estimates of α_j and β_j and x_i when compared with other algorithms (Pereira and Mexia, 2009). Although it has not been established that the zigzag algorithm converges to the absolute minimum of the goal function (3.2) in the complete case, the results are very similar to those of the double minimization algorithm which converges to the absolute minimum (Pereira and Mexia, 2009).

The minimization of the loss function: Using the zigzag algorithm, the minimization of the loss function (3.2) is carried out iteratively, starting with some initial values for the environmental indexes. For

the complete case (i.e. all the genotypes are present in each environment), the average yield per block can be a good initial value (Gusmão, 1985). When incomplete blocks are used, we have a very convenient situation when α-designs are used. Then, as the initial values for environmental indexes, one may take the average yields for the corresponding superblock. In the worst case, any initial values may be taken, since the computation time does not increase much.

After choosing the initial values for environmental indexes, the goal function is minimized with respect to the regression coefficients α_j and $\beta_j, j = 1, \ldots J$. Then, the α_j and β_j, are fixed and new environmental indexes are computed. The process is repeated until the convergence of the algorithm. After each iteration, the environmental indexes are rescaled so that the range of environmental indexes is kept unchanged. Hence, the iteration procedure is called zigzag algorithm (Mexia et al., 1999, Pereira and Mexia, 2010, Rodrigues et al., 2011).

Upper contour and genotype comparison: When the joint regression model (3.1) is adjusted for the *J* genotypes, we obtain the upper contour defined by the topmost adjusted linear regressions, which is a convex polygonal (Mexia et al., 1997) whose nodes

$$\tilde{\theta}_{j,j'} = \frac{\tilde{\alpha}_j - \tilde{\alpha}_{j'}}{\tilde{\beta}_{j'} - \tilde{\beta}_j},\ \mathrm{j} \neq \mathrm{j}';\ j, j' = 1, \ldots, J, \tag{3.3}$$

occur where two of the adjusted regression lines for the genotypes j and j' intersect (Figure S3.1). These nodes limit sub-ranges, which correspond to genotypes with maximum yields for the values of the environmental index in the corresponding sub-range. The genotypes in the upper contour are called dominant and should be selected. The remaining genotypes should be compared with the dominant ones in order to access whether they are significantly "dominated" within the entire range of the environmental indexes $\left[\theta_{\min}\ ;\ \theta_{\max}\right]$ (Pereira and Mexia, 2008). This comparison should be made on the extremes of the genotype dominance range.

With j' being the dominant genotype in the range $\left[\theta^{j'}_{\min}\ ;\ \theta^{j'}_{\max}\right]$, we will have either the case $\beta_{j'} > \beta_j$ or $\beta_{j'} < \beta_j$. When $\beta_{j'} > \beta_j$ $\left[\beta_{j'} < \beta_j\right]$, the minimum of $\left(\tilde{\alpha}_{\mathrm{j}'} + \tilde{\beta}_{j'} x\right) - \left(\tilde{\alpha}_{\mathrm{j}} + \tilde{\beta}_j x\right)$, for $\theta^{j'}_{\min} \le x \le \theta^{j'}_{\max}$, is attained at $x = \theta^{j'}_{\min}$ $\left[\mathrm{x} = \theta^{j'}_{\max}\right]$. Thus, in comparing genotypes j and j', if $\beta_{j'} > \beta_j$, we are led to test

$$H^{j,j'}_{\circ}:\ \alpha_{j'} + \beta_{j'}\theta^{j'}_{\min} = \alpha_j + \beta_j\theta^{j'}_{\min} \tag{3.4}$$

against

$$H^{j,j'}_{1}:\ \alpha_{j'} + \beta_{j'}\theta^{j'}_{\min} > \alpha_j + \beta_j\theta^{j'}_{\min};$$

and, when $\beta_{j'} < \beta_j$, the hypotheses to be tested are

$$H^{j,j'}_{\circ}:\ \alpha_{j'} + \beta_{j'}\theta^{j'}_{\max} = \alpha_j + \beta_j\theta^{j'}_{\max} \tag{3.5}$$

against

$$H^{j,j'}_{1}:\ \alpha_{j'} + \beta_{j'}\theta^{j'}_{\max} > \alpha_j + \beta_j\theta^{j'}_{\max}.$$

If the first [last] genotype is dominant, there are no genotypes with higher [smaller] slope and the only hypotheses to be tested are the ones described previously when $\beta_{j'} > \beta_j$ $[\beta_{j'} < \beta_j]$.

With $\boldsymbol{x}^b$ the vector of adjusted L_2 environmental indexes, let $\boldsymbol{x}_j$ be the sub-matrix of $\boldsymbol{X} = [\boldsymbol{1}^b : \boldsymbol{x}^b]$ whose rows correspond to the blocks that contain the genotype j and $\boldsymbol{D}_j$ be the diagonal matrix of the weights, p_i, for those blocks, $j, j = 1, \dots, J$. Then we have

$$\begin{bmatrix}\tilde{\alpha}_j \\ \tilde{\beta}_j\end{bmatrix} = \left(\boldsymbol{X}^T \boldsymbol{D}_j \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{D}_j \boldsymbol{Y}_j^b,\ j = 1, \dots, J, \tag{3.6}$$

with $\boldsymbol{Y}_j^b$ being the yield vector for genotype j.

Considering $S \sim \sigma^2\ \chi_g^2$ to indicate that S is the product by σ^2 of a central chi-square distributed variable with g degrees of freedom, we will assume that

$$\begin{bmatrix}\tilde{\alpha}_j \\ \tilde{\beta}_j\end{bmatrix} \sim N\left(\begin{bmatrix}\alpha_j \\ \beta_j\end{bmatrix};\ \sigma^2 \boldsymbol{W}_j\right),\ j = 1, \dots, J, \tag{3.7}$$

is independent of $S \sim \sigma^2\ \chi_g^2$, where $\boldsymbol{W}_j = \left(\boldsymbol{X}_j^T \boldsymbol{D}_j \boldsymbol{X}_j\right)^{-1}, j = 1, \dots, J,$ and $g = \sum_{j=1}^{J} b_j - 2J$.

Putting

$$k_\ell(\theta_\circ) = \begin{bmatrix}1 & \theta_\circ\end{bmatrix}\ \boldsymbol{W}_\ell \begin{bmatrix}1 \\ \theta_\circ\end{bmatrix},\ \ell = 1,\dots,J, \tag{3.8}$$

with $\theta_0 = \theta_{\min}^{j'}\ \left[\theta_{\max}^{j'}\right]$ the environmental indexes for which the comparisons should be carried out, when $\beta_{j'} > \beta_j\ \left[\beta_{j'} < \beta_j\right]$, we can use the t-statistic

$$t_{j,j'}(\theta_o) = \frac{\tilde{\alpha}_{j'} + \tilde{\beta}_{j'}\theta_0 - \tilde{\alpha}_j + \tilde{\beta}_j\theta_0}{\sqrt{\frac{S}{g}\left(k_j(\theta_o) + k_{j'}(\theta_o)\right)}},\ j \neq j', \tag{3.9}$$

to test the hypotheses (3.4) and (3.5).

Since $\tilde{\alpha}_\ell + \tilde{\beta}_\ell\theta_0$, $l = j, j'$, follows a normal distribution with mean values $\alpha_\ell + \beta_\ell\theta_0$, and variances $\sigma^2 k_\ell(\theta_\circ)$, independent between themselves and of $S \sim \sigma^2\chi_g^2$, we can prove that when $H_{\circ,j,j'}(\theta_\circ)$ holds, $t_{j,j'}(\theta_\circ)$ follows a central t-distribution with g degrees of freedom. Thus, we can use one-tailed t tests.

If we intend to achieve a higher level of robustness, multiple comparison methods such as Scheffé or Bonferroni should be applied. When using the Scheffé method, representing by $f_{1-\alpha,r,g}$ the $1-\alpha$ quantile of the central F-distribution, with r and g degrees of freedom, $\alpha_1 + \beta_1\theta_{\max}^{j'}, \dots, \alpha_{j'-1} + \beta_{j'-1}\theta_{\max}^{j'}$, such that

$$\left|\tilde{\alpha}_j + \tilde{\beta}_j\theta_{max}^{j'} - \left(\tilde{\alpha}_{j'} + \tilde{\beta}_{j'}\theta_{max}^{j'}\right)\right| > \sqrt{\frac{S}{g}(j'-1)\left(k_j\left(\theta_{max}^{j'}\right) + k_{j'}\left(\theta_{max}^{j'}\right)\right)f_{1-\alpha,j'-1,g}},\ j = 1,\dots,j'-1, \tag{3.10}$$

are jointly significantly lower than $\alpha_{j'} + \beta_{j'}\theta_{\max}^{j'}$ at the significance level α, and we may conclude that $\alpha_1 + \beta_1\theta_{\max}^{j'}, \dots, \alpha_{j'-1} + \beta_{j'-1}\theta_{\max}^{j'}$, which hold condition (3.10), are significantly dominated genotypes.

On the other hand, $\alpha_{j'+1} + \beta_{j'+1}\theta_{\min}^{j'}, \dots, \alpha_J + \beta_J\theta_{\min}^{j'}$, for which

$$\left|\tilde{\alpha}_j + \tilde{\beta}_j\theta^{j'}_{min} - \left(\tilde{\alpha}_{j'} + \tilde{\beta}_{j'}\theta^{j'}_{min}\right)\right| > \sqrt{\frac{S}{g}\left(J - j'\right)\left(k_j\left(\theta^{j'}_{min}\right) + k_{j'}\left(\theta^{j'}_{min}\right)\right)f_{1-\alpha,J-j',g}}\,,\ j = j'+1,\ldots,J, \tag{3.11}$$

are also jointly significantly lower than $\alpha_{j'} + \beta_{j'}\theta^{j'}_{\min}$ at the significance level α, and we may conclude that $\alpha_{j'+1} + \beta_{j'+1}\theta^{j'}_{\min},\ldots,\alpha_J + \beta_J\theta^{j'}_{\min}$ which hold condition (3.11) are significantly dominated genotypes. One should note that all significant differences hold at the same joint significant level α.

When we use the Bonferroni multiple comparison method, $\alpha_1 + \beta_1\theta^{j'}_{\max},\ldots,\alpha_{j'-1} + \beta_{j'-1}\theta^{j'}_{\max}$, such that

$$\left|\tilde{\alpha}_j + \tilde{\beta}_j\theta^{j'}_{max} - \left(\tilde{\alpha}_{j'} + \tilde{\beta}_{j'}\theta^{j'}_{max}\right)\right| > t_{1-\alpha/2(j'-1)g}\sqrt{\frac{S}{g}\left(k_j\left(\theta^{j'}_{max}\right) + k_{j'}\left(\theta^{j'}_{max}\right)\right)},\ j = 1,\ldots,j'-1, \tag{3.12}$$

are jointly significantly lower than $\alpha_{j'} + \beta_{j'}\theta^{j'}_{\max}$ at the significance level α, and $\alpha_{j'+1} + \beta_{j'+1}\theta^{j'}_{\min},\ldots,\alpha_J + \beta_J\theta^{j'}_{\min}$ for which

$$\left|\tilde{\alpha}_j + \tilde{\beta}_j\theta^{j'}_{min} - \left(\tilde{\alpha}_{j'} + \tilde{\beta}_{j'}\theta^{j'}_{min}\right)\right| > t_{1-\alpha/2(j-j')g}\sqrt{\frac{S}{g}\left(k_j\left(\theta^{j'}_{min}\right) + k_{j'}\left(\theta^{j'}_{min}\right)\right)},\ j = j'+1,\ldots,J, \tag{3.13}$$

are also jointly significantly lower than $\alpha_{j'} + \beta_{j'}\theta^{j'}_{\min}$ at the significance level α.

These multiple comparison methods (Scheffé and Bonferroni) may be used in booth complete and incomplete case. However, for the complete case, the Tukey method ((Scheffé, 1959), p. 73) may also be used. Then, $\alpha_1 + \beta_1\theta^{j'}_{\max},\ldots,\alpha_{j'-1} + \beta_{j'-1}\theta^{j'}_{\max}$, for which

$$\left|\tilde{\alpha}_j + \tilde{\beta}_j\theta^{j'}_{max} - \left(\tilde{\alpha}_{j'} + \tilde{\beta}_{j'}\theta^{j'}_{max}\right)\right| > T_{1-\alpha,j',g}\sqrt{k\left(\theta^{j'}_{max}\right)\frac{S}{g}},\ j = 1,\ldots,j'-1, \tag{3.14}$$

where $T_{1-\alpha,k,g}$ is the $1-\alpha$ quartile of the studentized range statistic with k and g degrees of freedom, are jointly significantly lower than $\alpha_{j'} + \beta_{j'}\theta^{j'}_{\max}$ at the α-level. Lastly, $\alpha_{j'+1} + \beta_{j'+1}\theta^{j'}_{\min},\ldots,\alpha_J + \beta_J\theta^{j'}_{\min}$, for which

$$\left|\tilde{\alpha}_j + \tilde{\beta}_j\theta^{j'}_{min} - \left(\tilde{\alpha}_{j'} + \tilde{\beta}_{j'}\theta^{j'}_{min}\right)\right| > T_{1-\alpha,J-j'+1,g}\sqrt{k\left(\theta^{j'}_{min}\right)\frac{S}{g}},\ j = j'+1,\ldots,J, \tag{3.15}$$

are jointly significantly lower than $\alpha_{j'} + \beta_{j'}\theta^{j'}_{\min}$ at the α-level.

We point out that in the complete case $k_l(\theta_o) = k(\theta_o)$ since the matrices X_l, $l = 1,\ldots,J$, are all equal to $\boldsymbol{X} = [\mathbf{1}^b : \boldsymbol{x}^b]$.

To measure selection effectiveness we can use the ratios

$$\begin{cases} r_1 = \dfrac{\text{Number of dominant cultivars}}{\text{Number of cultivars}} \\ r_2 = \dfrac{\text{Number of non dominated cultivars}}{\text{Number of cultivars}} \end{cases},$$

to obtain the proportion of dominant and non-dominated genotypes, respectively. JRA's power in selecting genotypes increases with the decrease of r_1. Other techniques such as control of false discovery rate can be used for these comparisons (Pereira and Mexia, 2008).

All the results of JRA, including the computation of the environmental indexes using the zigzag algorithm and all the multiple comparison tests, presented in this chapter were obtained using the R software.

3.2.2. AMMI model

The AMMI model has been widely used to analyze multi-environment trials. It combines the ANOVA and the PCA, where ANOVA is performed first to extract the main effects of the two-way table with genotypes and environments and then PCA is applied to the resultant matrix with GEI (Gauch, 1992). The AMMI model can be written as

$$y_{i,j} = \mu + G_i + E_j + \sum_{n=1}^{N} b_{i,n} z_{j,n} + \varepsilon_{i,j}, \tag{3.16}$$

where $y_{i,j}$ is the yield of genotype i in environment j, μ is the grand mean, G_i are the genotype mean deviations (genotype means minus the grand mean), E_j are the environment mean deviations, $b_{i,n}$ and $z_{j,n}$ are the genotypic and environmental parameters (scores) for the term n (i.e. the genotype and environment principal component scores for PCA axis n), N is the number of interaction principal component (IPC) axes retained by the model and $\varepsilon_{i,j}$ is the residual.

If crossovers are present in the data, it is likely that mega-environments, that is, groups of environments with similar outcome regarding the response variable (e.g. yield), can be constructed. A mega-environment can be defined as a portion (not necessarily contiguous) of a crop species' growing region with a fairly homogeneous environment that causes similar genotypes to perform best (Gauch and Zobel, 1997). The mega-environments should be built in order to maximize the differences between them and minimize the differences within them. In this way, it is possible to use the results from a given location with a higher predictive reliability for other locations and/or years under similar environmental conditions of a given mega-environment. This homogeneity within mega-environments facilitates the job of plant breeders and lowers the costs with multi-environment trials.

Several statistical strategies, using either classification or ordination methods, have been presented to group locations/environments into mega-environments (Paderewski et al., 2011). Here, we will be mainly interested in the ordination procedure provided by the AMMI model, because of its ability to deal with biplot graphs (Bradu and Gabriel, 1978) that are very useful to delineate mega-environments (Gauch and Zobel, 1997).

The strategy of creating mega-environments is useful only if the number of mega-environments is manageable. Although predictable GEI (e.g due to soils or consistent climatic differences across locations) increases the number of mega-environments, unpredictable GEI (e.g. due to climate variation between years) will decrease it (Annicchiarico et al., 2005). Even if a statistical test diagnoses AMMI3 or a higher model, practical constraints of achieving a workable number of mega-environments usually limit the model to AMMI1 or AMMI2.

The software MATMODEL version 3.0 (Gauch, 2007) was used to perform the AMMI analyses and compute the mega-environments.

3.2.3. The Data

The data set used in this chapter is from a plat breeding program of spring barley (*Hordeum vulgare L.*) experiments carried out between 2004 and 2006 by the Central Institute for Supervising and Testing in Agriculture in Brno, Czech Republic. In 2004, we had three replications for each of the 43 genotypes in 28 locations; in 2005, three replications of 41 genotypes in 26 locations; and in 2006, three replications of 42 genotypes in 22 locations.

3.3. Results

This section presents a comparison between JRA and the AMMI model using the spring barley data set. Sections 3.3.1–3.3.3 present the main results from the JRA, while subsections 3.3.4–3.3.6 present the main results from the AMMI analysis.

Table 3.1 gives the estimates, obtained using the zigzag minimization algorithm, for intercept, slope and coefficients of determination for the three years in study. The genotypes are ordered by slope, with the intermediate being omitted to avoid an extensive table.

Table 3.1. Adjusted regressions coefficients and coefficients of determination, ordered by slope in each year.

2004				2005				2006			
Genotype	$\tilde{\alpha}_j$	$\tilde{\beta}_j$	R_j^2	Genotype	$\tilde{\alpha}_j$	$\tilde{\beta}_j$	R_j^2	Genotype	$\tilde{\alpha}_j$	$\tilde{\beta}_j$	R_j^2
5076180	-1.81	1.24	0.95	5076209	-0.65	1.16	0.96	5076389	-0.83	1.21	0.90
5076212	-1.19	1.17	0.92	5076741	-0.60	1.12	0.95	5073987	-1.32	1.20	0.83
5076188	-1.33	1.14	0.87	5076690	-0.63	1.12	0.90	5075152	-1.08	1.18	0.84
5076178	-1.34	1.13	0.94	5075710	-0.41	1.12	0.94	1020194	-1.41	1.17	0.93
5075710	-0.62	1.12	0.85	5076684	-0.79	1.11	0.97	5077249	-0.69	1.12	0.96
5076182	-0.76	1.11	0.96	1020181	-0.62	1.10	0.98	1020037	-1.40	1.12	0.92
1020181	-0.32	1.08	0.94	5076389	-0.56	1.09	0.94	5076684	-0.78	1.09	0.87
…	…	…	…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…	…	…	…
5076192	1.12	0.88	0.90	1020194	0.002	0.94	0.86	5076665	0.99	0.86	0.87
1020062	0.98	0.88	0.89	1020178	0.16	0.91	0.95	5077168	0.67	0.86	0.88
1020067	0.37	0.87	0.84	5076700	0.80	0.91	0.93	5077231	0.97	0.85	0.84
5076205	1.08	0.87	0.95	5073863	0.82	0.90	0.92	5077169	1.00	0.85	0.84
1020077	1.14	0.85	0.80	5073811	0.45	0.89	0.98	5077202	1.00	0.85	0.79
5075636	1.44	0.83	0.88	1020130	0.60	0.89	0.97	5076700	0.94	0.84	0.71
1020034	0.95	0.81	0.73	5076665	1.13	0.86	0.96	5076678	1.25	0.77	0.76

3.3.1. JRA – 2004

Table 3.2 shows the dominant genotypes, range of dominance and environments where the genotypes are dominant. Figure S3.1 depicts the adjusted regression lines for the four genotypes which form the upper contour.

Table 3.2. The dominant genotypes, range of dominance, environments where the genotypes are dominant and the number of significantly dominated genotypes for 2004.

Dominant genotypes	Range of dominance	Environments	Number of dominated genotypes			
			t-test	Scheffé	Bonferroni	Tukey
5076209	[6.45 ; 10.59]	UHO,LED, STV, STV1, HRA, HE, UHO1, LED1, CHT, BR, BR1, HRA1, HE1, CHR, VYS, CHT1, CHR1, CAS, SED, SED1, VER, JAR, VYS1, CAS1, VER1, JAR1	18	0	6	3
1020191	[6.38 ; 6.45]		22	0	8	5
5076389	[6.07 ; 6.38]	LIP1	22	0	6	3
5075636	[5.99 ; 6.07]	LIP	20	1	7	3

Notes: The number of significantly dominated genotypes was obtained at the 5% significant level using the one-sided t-tests and the Scheffé, Bonferroni and Tukey multiple comparison methods.

Genotype 507209 is dominant in almost all the environments (Table 3.2 and Figure S3.1). LIP and LIP1 are the environments where the less productive genotypes are dominant (1020191, 5076389 and 5075636). Table 3.2 also presents the results for genotype comparison and selection, using the multiple comparison procedures described above. The dominant genotypes, that is, genotypes which integrate the upper contour, should be compared with the remaining in order to evaluate whether the differences are significant. The results of the multiple comparisons using (i) one-sided *t* tests, (ii) the Scheffé method, (iii) the Bonferroni method and (iv) the Tukey method are presented. Using the t-test, we obtain an efficiency ratio $r_1 = 4/43 = 0.09$, which represents an efficient genotype selection. The Scheffé method is, usually, too conservative and leads to a non-rejection of the hypothesis (3.4) and (3.5), and the Bonferroni multiple comparison method is the most advisable method because of its robustness in obtaining dominated genotypes.

3.3.2. JRA – 2005

A similar analysis was performed in 2005. Table 3.3 shows the dominant genotypes, range of dominance and environments where the genotypes are dominant, and presents the results from the multiple comparison tests. Figure S3.2 depicts the adjusted regression lines for the three genotypes which form the upper contour.

Table 3.3. The dominant genotypes, range of dominance, environments where the genotypes are dominant and the number of significantly dominated genotypes for 2005.

Dominant genotypes	Range of dominance	Environments	Number of dominated genotypes			
			t-test	Scheffé	Bonferroni	Tukey
5076209	[6.18 ; 9.18]	BR1, LED, LIP1, LED1, BR, CHT, HE, CHR, CHR1, JAR, CHT1, CAS1, PJA, CAS, JAR1, VER, PJA1, VER1	31	1	23	13
5075152	[5.60 ; 6.18]	UHO, LIP, HE1	29	0	15	13
5075636	[3.17 ; 5.60]	VYS1, VYS, HRA1, HRA, UHO1	29	0	12	10

Notes: The number of significantly dominated genotypes was obtained at the 5% significant level using the one-sided t-tests and the Scheffé, Bonferroni and Tukey multiple comparison methods.

Genotype 507209 is dominant (most productive) in environments with higher environmental indexes. UHO, LIP and HE1 are the environments where the intermediate productive genotype (5075152) is dominant, and VYS1, VYS, HRA1, HRA and UHO1 are the environments where the less productive genotype (5075636) is dominant (Figure S3.2 and Table 3.3). Using t-test, we obtained the efficiency ratio $r_1 = 3/41 = 0.07$.

3.3.3. JRA – 2006

Table 3.4 shows the dominant genotypes, range of dominance and environments where the genotypes are dominant, and presents the results from the multiple comparison tests for 2006. Figure S3.3 depicts the adjusted regression lines for the four genotypes which form the upper contour.

Table 3.4. The dominant genotypes, range of dominance, environments where the genotypes are dominant and the number of significantly dominated genotypes for 2006.

Dominant genotypes	Range of dominance	Environments	Number of dominated genotypes			
			t-test	Scheffé	Bonferroni	Tukey
5076389	[7.27 ; 8.63]	VER, VER1	30	0	11	7
5075710	[4.88 ; 7.27]	UHO, HRA, CHR, HE, HE1, LED1, LED, BR, BR1, CHR1, CAS	20	0	6	4
5076209	[4.18 ; 4.88]	STV1, CAS1, HRA1, LIP1	30	3	20	14
5077153	[3.62 ; 4.18]	CHT, STV, LIP, UHO1, CHT1	30	0	15	8

Notes: The number of significantly dominated genotypes was obtained at the 5% significant level using the one-sided t-tests and the Scheffé, Bonferroni and Tukey multiple comparison methods.

Genotype 5076389 is dominant only in the environments VER and VER1 on the rightmost range of environmental indexes. Genotype 5075710 is the winner in most of the environments. STV1, CAS1, HRA1 and LIP1 are the environments where the intermediate productive genotype (5076209) is dominant, and CHT, STV, LIP, UHO1 and CHT1 are the environments where the less productive genotype (5077153) is dominant. Using t-test, we obtain the efficiency ratio $r_1 = 4/42 = 0.1$.

Overall, all the efficiency ratios r_1 of the studied years were relatively low. This represents a good performance of JRA to recommend new genotypes. Moreover, the use of this approach would speed up the process of genotype selection. We propose that in the future a significantly dominated genotype in two consecutive years should be eliminated and the rightmost and intermediate dominant genotypes should be selected.

3.3.4. AMMI analysis – 2004

Following Gauch (1992), AMMI analysis was applied to the two-way table with three replications for each of the 43 genotypes in the 28 locations. Table 3.5 shows the ANOVA for the AMMI5 model. The proportions of the treatments' sum of squares (SS) due to genotypes, environments and GEI account for 4.1%, 77.8% and 10.2%, respectively. So, the interaction is as important as 2.5 times the genotype main

effects. Since the mean square (MS) error is 0.208 and the interaction has 1134 degrees of freedom, the noise in the interaction may be estimated as 235.87 ((Gauch, 1992), p. 147). Subtracting it from the interaction SS of 655.50, the GEI signal can be estimated as 419.63, or 64.02%. We can then conclude that the proportion of the interaction, after removing the noise (i.e. 419.65), is more than 1.5 times the genotype main effect (i.e. 263.60). The model diagnosis was also made by using the cross-validation suggested by Gauch (1992) and the AMMI5 was the most accurate for this data set. On comparison, although JRA provides a more parsimonious model, it captures a SS of 47.24, which represents only 32.2% of the GEI explained by IPC1.

Table 3.5. Results of the ANOVA for the AMMI5 model in 2004.

Source	df	SS	MS
Total	3611	6398.8	1.772
Treatments	1203	5898.7	4.903
Genotypes	42	263.6	6.276
Environments	27	4979.6	184.430
G x E	1134	655.5	0.578
IPC 1	68	146.8	2.159
IPC 2	66	82.8	1.254
IPC 3	64	64.2	1.004
IPC 4	62	55.3	0.891
IPC 5	60	50.0	0.834
Residual	814	256.5	0.315
Error	2408	500.0	0.208

Notes: This analysis is for the yield in spring barley for 2004. The grand mean is 8.367 t ha^{-1}.

Using the software MATMODEL (Gauch, 2007), it is possible to identify the mega-environments first defined by Gauch and Zobel (1997). Figure 3.1 depicts the AMMI1 biplot with two mega-environments. This model was chosen because the SS of the IPC decline rapidly after the IPC1. If we compare the results from AMMI1 model (Figure 3.1) with those from the full AMMI model, we go from 2 to 14 mega-environments. This big difference is due to the amount of noise in the full AMMI model, which makes things more complicated than they really are. Ordinarily, any small mega-environment, with few members or little advantage over other near winners, is ignored and its members are reassigned to a nearby larger mega-environment.

3.3.5. AMMI analysis – 2005

Table 6 gives the results of the ANOVA for the AMMI5 model. Genotypes, environments and GEI account for 2.0%, 93.0% and 5.0% of the total SS, respectively. Althoug the GEI is very small, it is responsible for a factor of more than 2.5 times the genotype main effects. In this case, almost all of the variation is due to environmental changes between the locations. Cross-validation was used for model diagnosis, and we obtained the AMMI3 as the most accurate. For sake of simplicity, Figure 3.2 presents the AMMI2 biplot. IPC1 and IPC2 are responsible for explaining 21.2% and 12.2% of the interaction SS.

On comparison with AMMI, the JRA was found to be responsible for a SS of 31.81, which represents only 41.5% of the GEI captured by IPC1.

As in 2004, the number of mega-environments is reduced from 11 to 6 when moving from the full AMMI model to the AMMI2 with two IPCs.

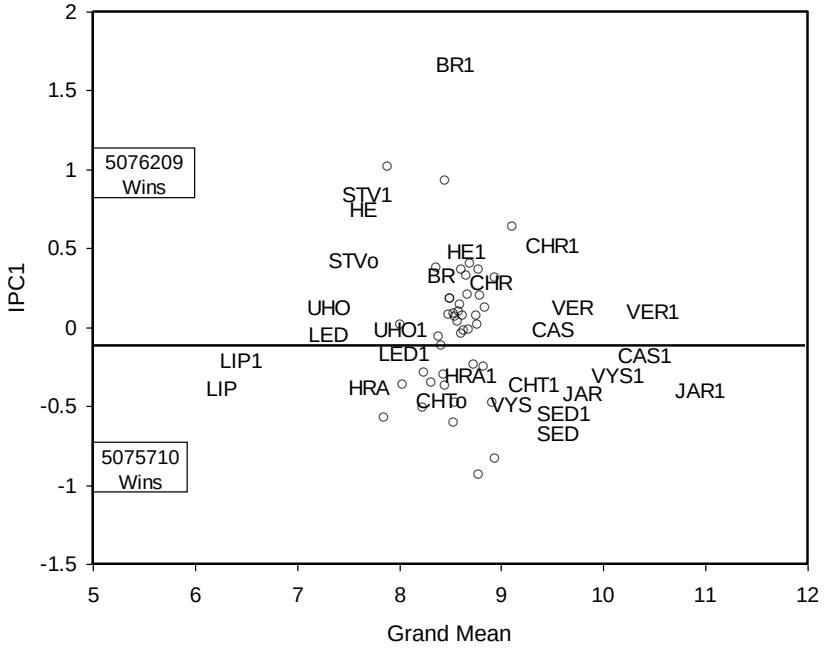

Figure 3.1. AMMI1 biplot for 2004. The abscissa represents the grand mean and the ordinate represents the IPC1 scores. The two mega-environments obtained are presented together with the winner genotypes.

Table 3.6. Results of the ANOVA for the AMMI5 model in 2005.

Source	df	SS	MS
Total	3194	7534.0	2.359
Treatments	1065	7195.1	6.756
Genotypes	40	141.6	3.539
Environments	25	6692.3	267.691
G x E	1000	361.2	0.361
IPC 1	64	76.7	1.199
IPC 2	62	44.0	0.709
IPC 3	60	37.2	0.620
IPC 4	58	26.7	0.460
IPC 5	56	24.4	0.435
Residual	700	152.3	0.218
Error	2129	339.0	0.159

Notes: This analysis is for the yield in spring barley for 2005. The grand mean is 8.367 t ha^{-1}. The noise in GEI may be estimated by the interaction degrees of freedom times the error MS, namely 159.00, which by difference from the total of 361.20 implies a GEI signal of 202.20, or 55.98%.

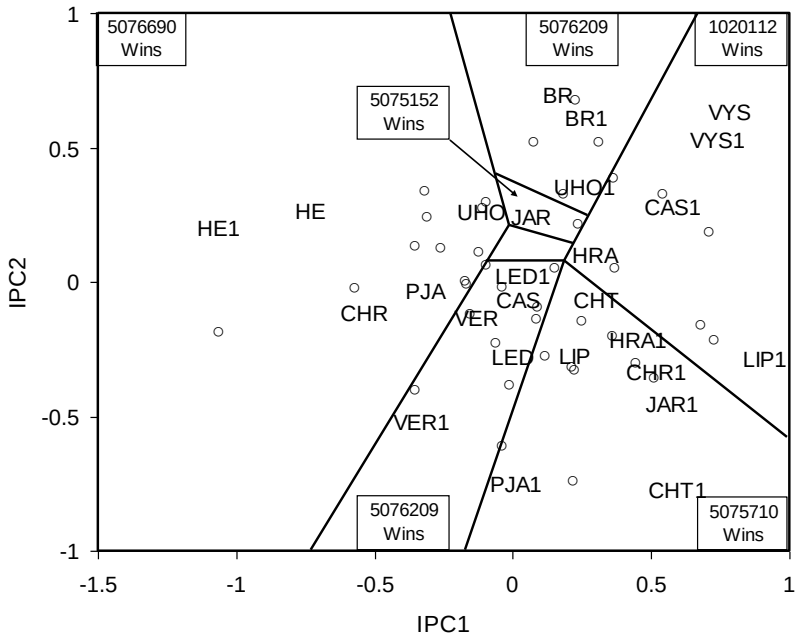

Figure 3.2. AMMI2 biplot for 2005. The abscissa represents the IPC1 and the ordinate represents the IPC2 scores. The six mega-environments obtained are presented together with the winner genotypes.

3.3.6. AMMI analysis – 2006

Table 3.7 gives the results of the ANOVA for the AMMI5 model. Genotypes, environments and GEI account for 4.9%, 76.2% and 11.9% of the total SS, respectively. Although the GEI is not very high when compared with the environmental main effects, it is still very significant being more than twice the genotype main effects. By cross-validation, the AMMI2 model was found to be the most accurate, with the IPC1 and IPC2 being responsible for capturing 37.7% and 19.4% of the GEI, respectively. This proportion together, that is, 57.1% of the GEI, is represented in the AMMI2 biplot in Figure 3.3. For comparison, the JRA captures a SS of 53.87 (i.e. 18.6% of the IPC1 SS), so the AMMI analysis is also more effective for the 2006 data set.

Again the simple and parsimonious AMMI2 model corresponds to 6 mega-environments, while the full AMMI model with all the noise results in 14 mega-environments.

3.3.7. Comparison between JRA and AMMI model

When summarizing all the data sets together, we can observe that genotype 5076209 is dominant in the three studied years (in 2 years is the rightmost range of productivity and in the other is suitable for use in middle-fertility environments) and wins a mega-environment in all the years as well. Genotype 5076389 is dominant in two of the studied years and is the winner in one mega-environment in 2006. Genotype 5075710 is dominant only in 2004 but wins a mega-environment in the three years under study.

Genotype 5076209, the most important in this study, is now a variety. It can be characterized as mid-early non-malting variety, mid-high plant that is mid-resistant to lodging giving very large grain and high to very high yield of grain.

Table 3.7. Results of the ANOVA for the AMMI5 model for 2006.

Source	df	SS	MS
Total	2771	6469.9	2.335
Treatments	923	6018.9	6.521
Genotypes	41	320.0	7.804
Environments	21	4929.0	234.715
G x E	861	769.9	0.894
IPC 1	61	290.0	4.755
IPC 2	59	149.4	2.533
IPC 3	57	63.3	1.110
IPC 4	55	50.5	0.919
IPC 5	53	42.4	0.800
Residual	576	174.2	0.302
Error	1848	451.0	0.244

Notes: This analysis is for the yield in spring barley for 2006. The grand mean is 8.367 t ha^{-1}. The noise in GEI may be estimated by the interaction degrees of freedom times the error MS, namely 210.08, which by difference from the total of 769.9 implies a GEI signal of 559.82, or 72.71%.

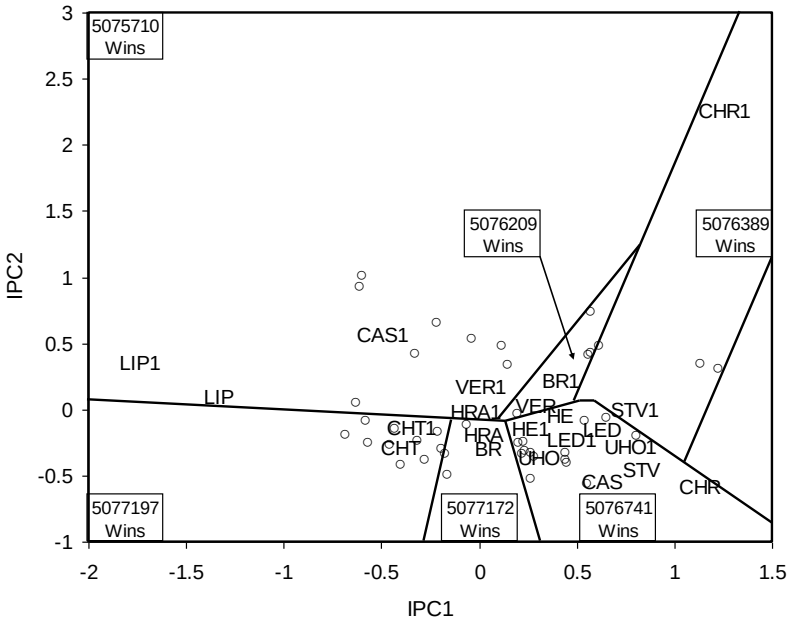

Figure 3.3. AMMI2 biplot for 2006. The abscissa represents the IPC1 and the ordinate represents the IPC2 scores. The six mega-environments obtained are presented together with the winner genotypes.

3.4. Discussion

Several comparisons have been made between the Finlay and Wilkinson regression (1963) and the AMMI models (Gauch, 1992) (e.g. Annicchiarico (1997b)), but none have made use of some of the key features of JRA referred in this chapter, namely the use of the zigzag algorithm to estimate the regression

coefficients and environmental indexes, and the application of multiple comparison procedures to test whether a dominant genotype is significantly better than the remaining ones.

When comparing the model accuracy between JRA and the AMMI model, we found the JRA to capture only between 18.6% and 41.5% of the respective AMMI IPC1 SS. This is one of the reasons why literature usually favors the AMMI model over the JRA. Moreover, the GEI in the AMMI model is analyzed with the singular value decomposition, which is the least-squares solution for the fitting of the data. This means that the AMMI1 model will always explain as much or (usually) more of the GEI SS than the JRA. However, when the IPC1 scores for environments are highly correlated with the environment indexes, the JRA captures nearly the same GEI SS as the AMMI1 model. This leads to the advantage of the JRA for being a more parsimonious model and to a clearer association with environmental characteristics. When comparing the models JRA and AMMI for predictability based on measures proposed by Brancourt-Hulmel et al. (1997) and Annicchiarico (2002), the AMMI model performs better for all years (Table 3.8).

Table 3.8. Model comparison for predict ability for yield in spring barley for 2004, 2005 and 2006.

Year	Brancourt-Hulmel et al. (1997)		Annicchiarico (2002)	
	JRA	AMMI	JRA	AMMI
2004	1.125	2.159	0.012	0.023
2005	0.795	0.958	0.008	0.020
2006	1.314	3.662	0.016	0.103

To assess the repeatability of the JRA and the AMMI model over time, we used the data of the year i, i = 2004 and 2005, to model and obtain the recommended top yielding genotypes per location and method (Figures 3.1–3.3 and Figures S3.1–S3.3). Then, the validation was made with the actual yield in the following year $i + 1$. A similar empirical model comparison concerning repeatability was made by Annicchiarico et al. (2006), where 2 years were used for modelling and 1 year for validation. They selected the AMMI model with one IPC and concluded that genotypic parameters (mean yield and IPC1) were highly repeatable and site parameters (mean yield and IPC1) were moderate to fairly low repeatable over time mainly due to within-site variation in annual rainfall. In our case, from the 28 locations tested in 2004, 24 were repeated in 2005. For the JRA, the dominant genotypes in 2004 yielded within the best 25% tested in 2005 for 13 locations and above the median for 17 locations. For the AMMI model, the winners of mega-environments in 2004 yielded within the best 25% tested in 2005 for 18 locations and above the median for 21 locations. From the 26 locations tested in 2005, 20 were repeated in 2006. For the JRA, the dominant genotypes in 2005 yielded within the best 25% tested in 2005 for 11 locations and above the median for 14 locations (of the 17 locations which had a 2005 dominant genotype tested in 2006). For the AMMI model, the winners of mega-environments in 2005 yielded within the best 25% tested in 2006 for 7 locations and above the median for 12 locations (of the 16 locations which had a 2005 dominant genotype

tested in 2006). The worse performance for the AMMI model is due to genotype 1020112, which had a bad performed in 2006 after being winner of one mega-environment in 2005.

The only alternative to Annicchiarico et al. (2006) to access repeatability for AMMI models is the direct validation of the AMMI predictors presented by Ebdon and Gauch (2011). They fitted a parsimonious AMMI5 model for turfgrass trials with 103 genotypes, 24 locations and 3 years (1997-1999), The model was validated with 10 genotypes planted in six locations (2005-2007 averages), concluding that the use of a parsimonious AMMI model can improve predictions across years and locations.

When modelling site-specific genotypic responses according to duration, considering only one year is a major limitation because it is not possible to separate the non-repeatable interaction between genotypes and locations (Annicchiarico, 2009).

In this example, the results of the Finlay and Wilkinson (1963) regression and Gusmão (1985)'s JRA approach were in agreement because we use complete data. However, when dealing with α-designs or incomplete blocks, the JRA performs better (Pereira and Mexia, 2009) and, in some cases, gives results very similar to the AMMI model (Rodrigues et al., 2011).

3.5. Supplementary material

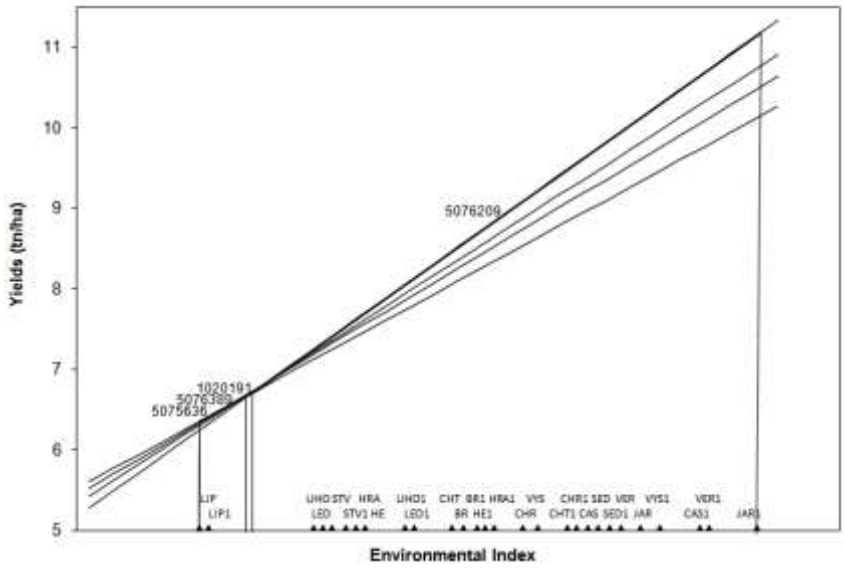

Figure S3.1. Adjusted regressions using L_2 environmental indexes, for 2004. The abscissa depicts the position of all the environments under study along the environmental indexes.

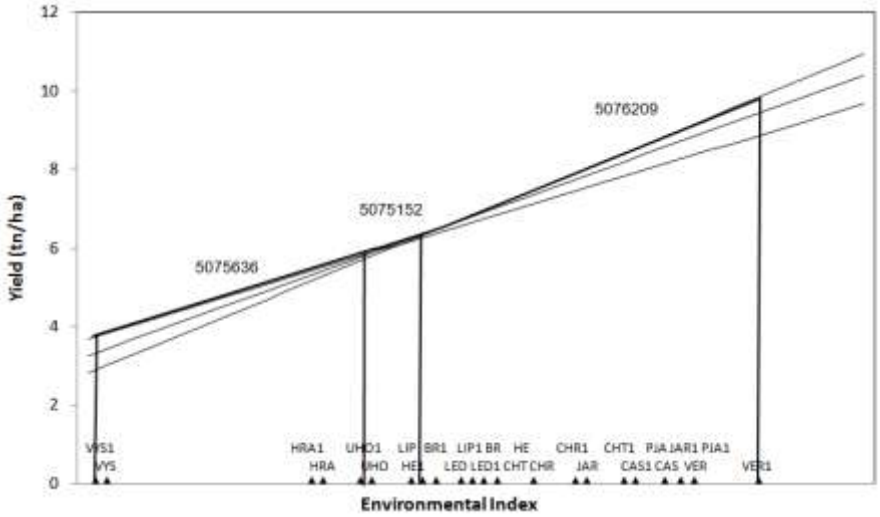

Figure S3.2. Adjusted regressions using L_2 environmental indexes, for 2005. The abscissa depicts the position of all the environments under study along the environmental indexes.

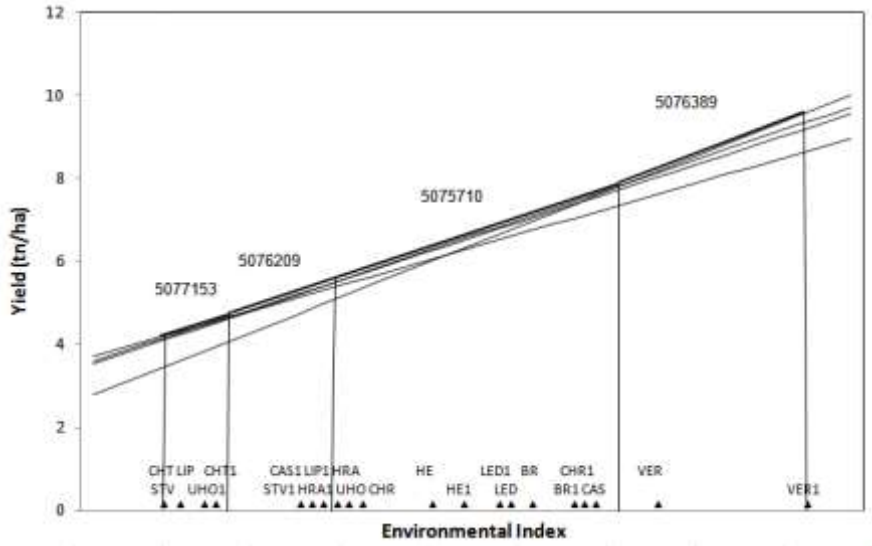

Figure S3.3. Adjusted regressions using L_2 environmental indexes, for 2006. The abscissa depicts the position of all the environments under study along the environmental indexes.

Chapter 4

4. Two new strategies for detecting and understanding QTL-by-environment interactions

Abstract

Two new strategies are proposed to improve the detection and understanding of quantitative trait loci (QTL), especially those exhibiting QTL-by-environment interactions (QEI), in the context of experiments conducted in multiple environments. First, a parsimonious Additive Main effects and Multiplicative Interaction (AMMI) model is applied to the phenotypic data in order to gain accuracy and thereby to increase the logarithm of odds (LOD) scores for QTL detections. Second, the environments are ordered by AMMI parameters that summarize genotype-by-environment interaction information in order to reveal consistent patterns and systematic trends that often have an evident ecological or biological interpretation. The combination of greater accuracy for the phenotypic data and systematic trends for the environments provides for more consistent and understandable QTL results. These new strategies are illustrated with two examples: preharvest sprouting scores of a biparental wheat (*Triticum aestivum* L.) population from 14 environments spread over five years, and yield for a doubled haploid barley (*Hordeum vulgare* L.) population tested in 16 environments. AMMI parameters can also provide successful predictions of entire QTL scans for new environments. The statistical methods developed here are of great generality, applicable across microbial and plant populations grown in multiple environments, and may be adapted to animal and human genetic studies.

Published as: Gauch, H.G., Rodrigues, P.C., Munkvold, J.D., Heffner, E.L. and Sorrells, M. (2011). Two New Strategies for Detecting and Understanding QTL by Environment Interactions. *Crop Science* 51: 96–113. The original paper can be found online in: https://www.crops.org/publications/cs/tocs/51/1.

4.1. Introduction

QTL scans are often conducted in multiple environments in order to increase generality and reliability, but frequently the outcome is inconsistent QTL detection. Inconsistent results raise questions about both validity and utility of these QTLs, especially those that are only marginally significant or infrequent. Inconsistent QTLs can emerge from a mixture of two causes: 1) from false positives and false negatives due to inadequate population size, or imperfect statistical models and noisy phenotypic data; and 2) from actual QTL-by-environment interactions (QEI) due to a given allele increasing a phenotypic trait in only some environments while having no detectable effect or even a significant negative effect in other environments.

Unfortunately, it is often difficult to decide whether a given inconsistent QTL results from spurious noise or actual QEI. Better methods for discriminating between these two possibilities would increase the value of QTL studies conducted across multiple environments. This chapter recalls two strategies for improving QTL scans that have already been published. First, a useful strategy has been refining statistical models to increase the power and reliability of QTL detections. This includes the development of composite interval mapping (CIM; Zeng, 1994) and more recently various best linear unbiased predictor (BLUP) and Bayesian procedures (Heffner et al., 2009, Zhang et al., 2005), along with producing convenient QTL software.

Second, another strategy specifically aimed at QEI detection in QTL experiments with multiple environments, is to apply the Additive Main effects and Multiplicative Interaction (AMMI) model to phenotypic data in order to gain strength from other environments (Jiang and Zeng, 1995) and derive interaction principal components (IPC) that summarize the genotype-by-environment interactions (GEI). This compresses the GEI matrix into IPC vectors that can serve as interaction traits for QEI scans (Romagosa et al., 1996). Recall that QTL scans require three kinds of input data: marker data for each genotype, a chromosome (or linkage group) map, and phenotypic data for each genotype. This last item comprises a vector whose length is the number of genotypes G, whereas the GEI information comprises a matrix of dimensions G and E. Consequently, it is useful to compress the information from a matrix into a vector, and applying principal components analysis to the interaction matrix produces a least-squares solution facilitating the interpretation of QTL and QEI scans. This approach has seen little adoption to date, as Piepho (2000) observed.

This chapter presents two new strategies for detecting QTLs and understanding QEI. One is to use a parsimonious AMMI model to gain accuracy for the phenotypic data used in QTL scans, thereby improving QTL results. The other is to use IPC environment scores to order the environments in a manner that reveals consistent patterns and systematic trends that may have an evident ecological or biological interpretation. These two new strategies for detecting and understanding QEI employ statistical methods that are of great generality, and are illustrated here in a biparental wheat (*Triticum aestivum* L.) population and a doubled haploid (DH) barley (*Hordeum vulgare* L.) population. They are fully applicable across microbial and plant populations grown in multiple environments, and may be adapted to animal and human genetic studies.

4.2. Materials and methods

4.2.1. Genotypic and phenotypic data

The first dataset used in this study concerns preharvest sprouting (PHS) in wheat, with the experimental methods described in detail by Munkvold et al. (2009). The visual rating of PHS (Figure S4.1) used a scale from 0 for no evidence of sprouting to 10 for extensive sprouting throughout the spike (Anderson et al., 1993). A doubled haploid (DH) population was derived from a cross between the PHS resistant variety Cayuga and the PHS susceptible variety Caledonia and phenotypic data were collected on 209 genotypes. There were 205 markers mapped to 42 linkage groups. The experiment was conducted in 17 environments in the vicinity of Ithaca, NY: Caldwell, Ketola, and Snyder in 2001; Caldwell, Helfer, and Ketola in 2002; Helfer and McGowan in 2003; Helfer, Ketola, and McGowan in 2004; Helfer, Ketola, and McGowan in 2005; and Caldwell, Helfer, and Snyder in 2006. However, the Helfer 2002 data were excluded from the analysis in Munkvold et al. (2009) because of concern about uneven misting during PHS evaluation in the greenhouse, so they used 16 environments.

From the Cayuga x Caledonia (CxC) dataset, we formed two subsets, called our primary and our prediction datasets. In order to reduce the amount of missing data, we retained only 197 of the 209 DH genotypes. The primary dataset included only 11 of the 17 environments, deleting the 2001 data because it was lacking about 50 genotypes and deleting the 2006 data to provide for a prediction exercise. The prediction dataset included the primary dataset (2002–2005) and the 2006 data, for a total of 14 environments. Environments are given brief and transparent code names, such as Cal2 for Caldwell in 2002. The only previous QTL study of this population is by Munkvold et al. (2009).

The second example is the Steptoe x Morex (SxM) barley mapping population, which was the first product of the North American Barley Genome Mapping Project (Hayes et al., 1993, Hayes et al., 1996). Since it was first made available online by Hayes et al. (1993), it has become a reference data set in QTL analysis, leading to dozens of publications.

The phenotypic data used in multi-environment QEI studies can vary in two particularly important properties: (1) noise level and (2) complexity of the GEI and hence the QEI. The CxC wheat data have high noise and rather simple interactions. By contrast, the SxM barley data have low noise and complex interactions.

4.2.2. Statistical analyses

AMMI analysis was done by MATMODEL version 3.0 (Gauch, 2007). Consider a two-way factorial experiment with a phenotype PHS, measured for G genotypes in E environments, with replication.

The AMMI model combines analysis of variance (ANOVA) and principal component analysis (PCA), with ANOVA performed first and then PCA applied to the resultant table of genotype-by-environment interactions (Gauch, 1992). First, the model equation for the ANOVA portion is:

$$y_{i,j} = \mu + \alpha_i + \beta_j + \varepsilon_{i,j}, \quad (4.1)$$

where $y_{i,j}$ is the PHS score for genotype i in environment j, μ is the grand mean, α_i is the deviation from the grand mean of genotype i, β_j is the deviation from the grand mean of environment j, and $\varepsilon_{i,j}$ is the GEI for genotype i in environment j.

Second, the interaction $\varepsilon_{i,j}$ is partitioned into interaction principal components (IPC), usually stopping before reaching the full model and thus leaving a residual:

$$\varepsilon_{i,j} = \sum_{n=1}^{N} \lambda_n \gamma_{i,n} \delta_{j,n} + \theta_{i,j}, \tag{4.2}$$

where λ_n is the singular value for IPC component n, $\gamma_{i,n}$ is the eigenvector value for genotype i in component n, $\delta_{j,n}$ is the eigenvector value for environment j in component n, $\theta_{i,j}$ is the residual for genotype g in environment j, and summation is over components $n = 1$ to N with the maximum possible choice being the full model having N equal the minimum of $G - 1$ and $E - 1$.

The members of the AMMI family are distinguished by a suffix, with AMMI0 retaining no IPC and hence having only the additive portion of the model, AMMI1 retaining 1 IPC, AMMI2 retaining 2 IPCs, and so on, until the final full model, with expected values equalling the actual data, being denoted by AMMIF. The square of the singular value for component *n* is the eigenvalue, λ_n^2. For each and every IPC, the eigenvectors γ and δ are scaled as unit vectors, $\sum_i \gamma_i^2 = \sum_j \delta_j^2 = 1$. MATMODEL scales the genotype and environment IPC scores as $\lambda^{0.5}\gamma_i$ and $\lambda^{0.5}\delta_j$ so that their product approximates the interaction $\varepsilon_{i,j}$ directly. Otherwise an additional multiplication by the singular value λ would be needed to approximate interactions were the unit eigenvectors γ and δ used for IPC values instead of these scores.

Combining the ANOVA and PCA parts of AMMI, the expected values for the AMMI1 model, denoted $Y1_{i,j}$, are:

$$Y1_{i,j} = \mu + \alpha_i + \beta_j + \lambda_1 \delta_{i,1} \gamma_{j,1} \tag{4.3}$$

Likewise, the expected values $Y2_{i,j}$ for the AMMI2 model would include the first two IPCs, and so on for higher members of the AMMI model family.

Model diagnosis for the most predictively accurate member of the AMMI model family was done by the jackknife or leave-one-out procedure (Efron and Gong, 1983). The data matrix for both experiments contained the average over two replications for each genotype in each environment. Each matrix entry in turn was temporarily withheld and an expectation-maximization algorithm was used to impute the missing cell. For each matrix entry the difference between the imputed and actual value was squared and these values were summed over the matrix, and finally the square root taken to obtain the root mean squared predictive difference (RMSPD). This procedure was repeated for the AMMI0 to AMMI7 models, selecting that model with the smallest RMSPD. Incidentally, cross-validation is used more commonly in the AMMI literature, but it is better suited to experiments with at least three or four replications (Gauch, 1992).

QTL scans were conducted with QTL Cartographer 2.5 (Wang et al., 2007) using CIM. Significance of QTL detections at the 0.05 level was determined by a permutation test with 1000 permutations (Churchill and Doerge, 1994).

4.3. Results for the wheat experiment

4.3.1. Preliminary analyses

In the previous study of QTLs for PHS in the present CxC population, Munkvold et al. (2009) applied CIM to their 16 environments individually, plus the mean over environments, for a total of 17 QTL scans. Significance was judged by a permutation test at the 0.05 level using 1000 permutations. As shown in their Table 3, they found 16 QTLs for PHS, with one QTL detected in all 17 scans and 12 detected in only 1 or 2 scans. There was a total of 65 detections, or an average of about 4 detections per QTL. Table 4.1 lists the 6 of those 16 QTLs that are prominent in our findings. For convenient reference, these main QTLs are given concise code names of QTL1 to QTL6. Incidentally, as also found by Munkvold et al. (2009), in environment Caldwell 2002, the peak for QTL1 shifts from marker WMC474 at 14 centimorgans (cM) to marker GWM429 at 6 cM.

Table 4.1. Main QTLs for preharvest sprouting. The six main QTLs are given concise code names here, QTL1 to QTL6. The full QTL names from Munkvold et al. (2009) are listed for each, along with the location of the peak in cM and the closest marker. The linkage groups are specified in the final portion of the full QTL names.

Code	QTL	Peak (cM)	Closest Marker
QTL1	Qphs.cnl-2B.1	14	WMC474
QTL2	Qphs.cnl-6D.1	28	CFD37
QTL3	Qphs.cnl-2D.1	37	wPT-9997
QTL4	Qphs.cnl-3D.1	26	GPW4152
QTL5	Qphs.cnl-1B.1	15	BARC240
QTL6	Qphs.cnl-4D.1	12	RHT-DF-MR2

Figure 4.1 shows QTL scans for the 11 environments in our primary dataset listed in the same chronological and alphabet order as in Table 1 in Munkvold et al. (2009). Obviously, from a biological or ecological viewpoint, this is an arbitrary order. As with Table 3 in Munkvold et al. (2009), Figure 4.1 shows that some QTLs are common and others are rare. But, the overall impression is one of inconsistent QTL detections because there are no apparent patterns in the QTLs and no obvious relationships among the environments (other than occasional pairs of scans that are rather similar, such as the bottom two scans). Such graphs, showing QTLs for each environment separately, are presented routinely for multi-environment experiments. As Piepho (2000) observes, separate analyses by each environment circumvent the problem of dealing with QEI, but it is difficult to integrate numerous separate analyses into a systematic pattern and coherent understanding.

Table 4.2 gives the ANOVA for the AMMI3 model. Note that genotypes, environments and GEI account for 35.0%, 31.6%, and 33.4% of the treatment sum of squares (SS). In this ANOVA table, MS for blocks were not fully taken into account since it is likely that another source of error is coming from different field researchers selecting spikes.

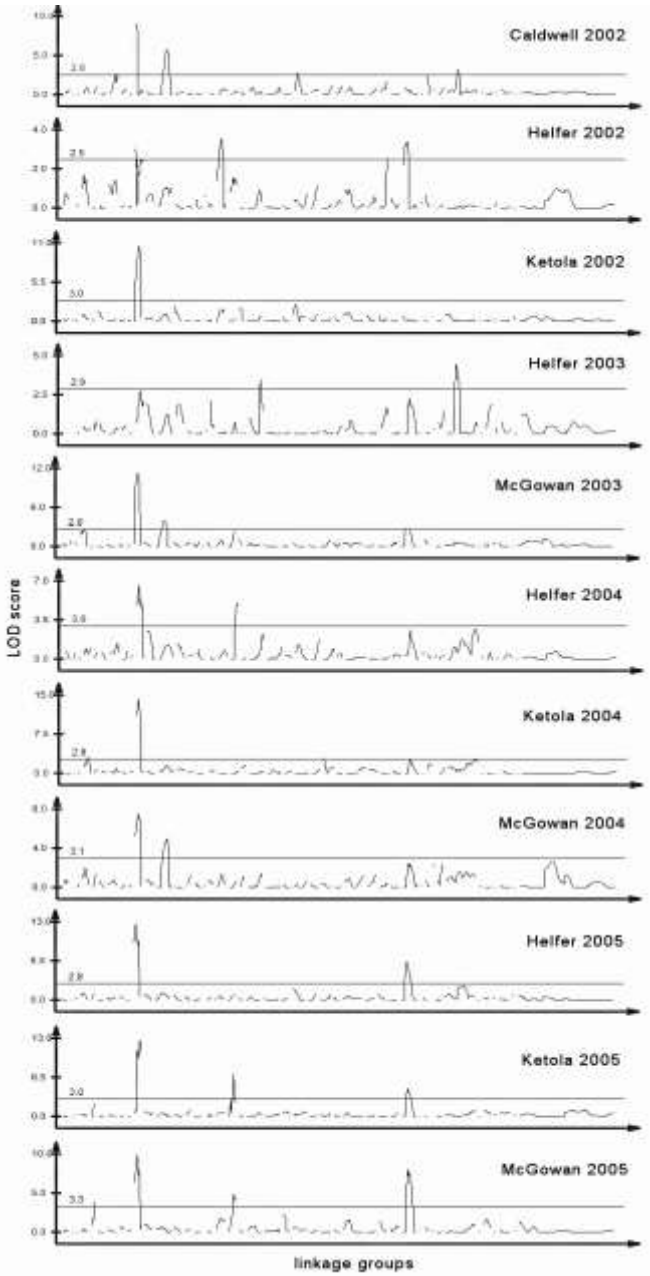

Figure 4.1. QTL scans for the 11 environments of the wheat PHS experiment simply ordered by location name and year. These scans are based on the raw data.

The amount of noise in the GEI may be estimated by the interaction degrees of freedom (df) times the error mean square (MS), namely 2195, which by difference from the GEI total of 2661 implies a GEI signal of 466, or 17.5% (Gauch, 1992: 147; Voltas, et al., 2002). The signal-to-noise (S/N) ratio for GEI is 466 / 2195 = 0.21, which is quite low. Because GEI has most of the experiment's df, most of the noise goes into GEI. By a similar calculation, the data matrix as a whole (genotype and environment main effects and GEI combined) has a much higher S/N ratio of 2.29.

Because the GEI is 17.5% signal and 82.5% noise, the most incisive perspective for crop scientists is that the variability in this dataset consists of 35.0% genotype effects, 5.8% GEI signal, and 27.6% GEI noise. Effective analysis for crop scientists requires focus on the 35.0% genotype effects and 5.8% GEI

signal while ignoring the 31.6% environment effects and discarding the 27.6% GEI noise. Incidentally, for comparison with AMMI, the linear regressions on environment means described by Finlay and Wilkinson (1963) capture a SS of 344.38, which is only 59.7% of the GEI captured by IPC1.

Table 4.2. AMMI3 analysis of variance. This analysis is for preharvest spouting scores of doubled haploid progeny from a cross between the resistant variety Cayuga and the susceptible variety Caledonia. The grand mean is 4.097. The noise in the genotype-by-environment interaction (GEI) may be estimated by the interaction df times the error MS, namely 2194.68, which by difference from the total of 2660.79 implies a GEI signal of 466.11, or 17.5%.

Source	df	SS	MS	Probability
Total	4306	10370.35	2.408	
Treatments	2166	7974.12	3.682	0.0000000
Genotypes	196	2789.94	14.234	0.0000000
Environments	10	2523.39	252.339	0.0000000
GEI	1960	2660.79	1.358	0.0011753
IPC1	205	577.31	2.816	0.0000000
IPC2	203	366.51	1.805	0.0000017
IPC3	201	321.77	1.601	0.0003115
Residual	1351	1395.20	1.033	0.8994796
Error	2140	2396.23	1.120	

Figure 4.2 shows QTL scans for the main effects (averages over the 11 environments) and IPC1 to IPC3 from AMMI analysis. Such graphs were introduced by Romagosa et al. (1996). QTL1 shows a main effect and an interaction effect on IPC2, and similarly QTL2 shows a main effect and an interaction effect on IPC1. QTL3 to QTL5 show only main effects. QTL6 shows only an interaction effect on IPC1.

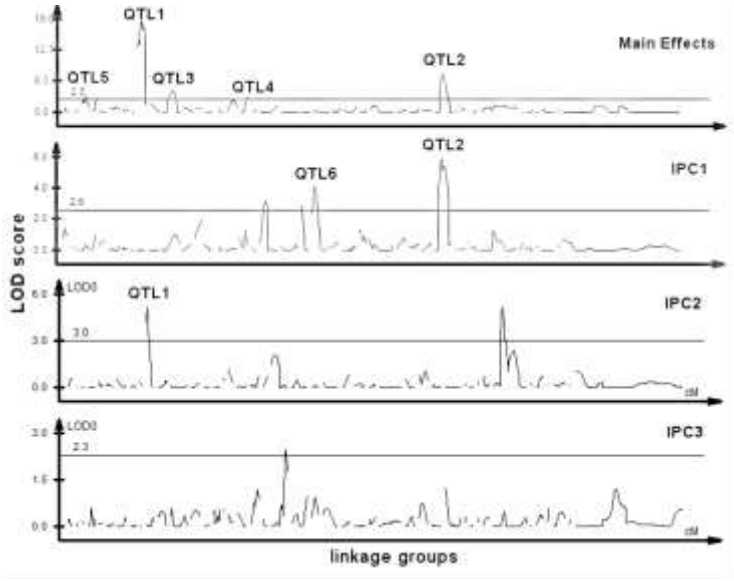

Figure 4.2. QTL scans for the main effects and IPC1 to IPC3 for the wheat PHS experiment. Peaks are identified for the six main QTLs, QTL1 to QTL6.

4.3.2. Gaining accuracy

The GEI was 82.5% noise, implying potential for improving the accuracy of the phenotypic data with a parsimonious AMMI model. Figure 4.3 shows the RMSPD from the jackknife procedure for the AMMI model family from AMMI0 to AMMI7. Moving to the right, models are more complex, or less parsimonious. AMMI1 achieves the lowest RMSPD of 1.131.

The relationship between accuracy and parsimony exemplified in Figure 4.3 has been aptly named "Ockham's hill" (MacKay, 1992, Gauch, 2006). Statistical theory provides three interrelated explanations for Ockham's hill: signal-noise selectivity, variance-bias tradeoff, and direct-indirect information (Stein, 1955; Gauch, 2002: 269–326). Plotting RMSPD values for an AMMI model family results in what may be termed an Ockham's valley, rather than an Ockham's hill, because decreasing RMSPD values indicate increasing predictive accuracy, so the best model is at the bottom of the valley. To the left of the best model, excessively simple models underfit real signal; whereas to the right, excessively complex models overfit spurious noise.

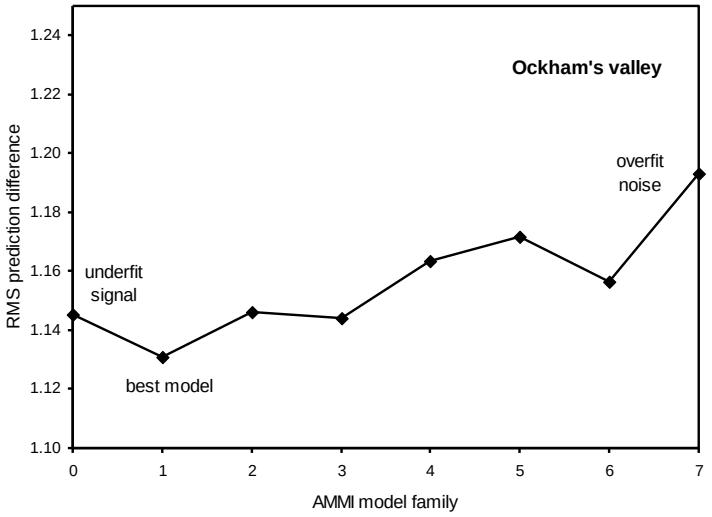

Figure 4.3. Ockham's valley for the wheat PHS experiment. The abscissa shows AMMI models of increasing complexity from AMMI0 to AMMI7, and the ordinate shows the root mean square (RMS) predictive difference determined by jackknife resampling. The most predictively accurate member of the AMMI family is AMMI1. To that model's left, excessively simple models underfit real signal, whereas to the right, excessively complex models overfit spurious noise.

Figure 4.4 shows QTL scans before and after AMMI1 refinement of the phenotypic data, using Ketola 2004 as a typical example. For QTL1, the peak LOD score is increased by AMMI1 from 11.8 to 18.4 and for QTL2 from 2.8 to 6.7. The thresholds for significance at the 0.05 level are 2.8 for the raw data and 2.3 for the AMMI1 estimates, so all four of those detections are significant. For QTL3 to QTL5, the changes are 1.4 to 5.5, 0.4 to 3.4, and 2.6 to 3.8, so only AMMI1 detects these three QTLs. This raises the suspicion that using the less accurate raw data as the "naïve estimator" led to three false negatives. For these five QTLs, the average increase in peak height due to AMMI1 pre-processing of the phenotypic data is 3.76.

The asterisk (*) in Figure 4.4 marks a QTL detected only by the raw data with a peak height of 2.9 near marker wPT-3661, which corresponds to *QPhs.cnl-5B.1* with PHS resistance coming from the Caledonia parent in Munkvold et al. (2009). In their study, which used somewhat more genotypes and more environments than our subset of the data, this QTL was detected only in the mean over their 16 environments, but not in any individual environment. In our study, this QTL was quite marginal, with a LOD score of 2.9 barely exceeding the threshold of 2.8. Hence, this QTL might well be a false positive, in which case AMMI should be credited for not detecting it.

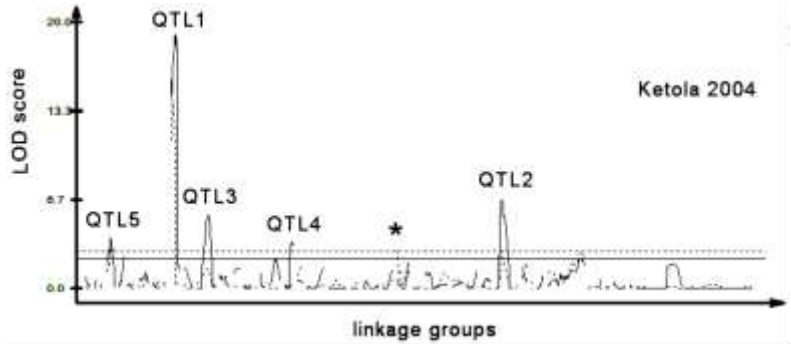

Figure 4.4. QTL scans for Ketola 2004 based on the AMMI1 estimates (solid line) and the raw data or naïve estimates (dotted line). The AMMI1 estimates support detections of 5 of the main QTLs, whereas the raw data support detections of only QTL1 and QTL2. This provides presumptive evidence that the raw data had three false negatives. The asterisk (*) marks a QTL detected only by the raw data, having a LOD score of 2.9 barely exceeding the threshold of 2.8 for the 0.5 significance level.

In review, for this environment, Ketola 2004, there is strong evidence that more accurate phenotypic data from AMMI pre-processing translated into better detection of the main QTLs, QTL1 to QTL5. There is also presumptive evidence that AMMI avoided three false negatives and one false positive encountered with the raw data. Finally, the LOD threshold for AMMI1 was 0.5 lower than for the raw data. However, generalizing over all 11 environments, the average LOD threshold for AMMI1 was 2.87 and for the actual data was 2.89.

Generalizing beyond this one environment to all 11 environments for QTL1 to QTL6, there were 11 detections for QTL1, 8 for QTL2, 7 each for QTL3–QTL5, and 1 for QTL6, for a total of 41 QTL detections by AMMI1 or actual data or both. In 20 cases, both AMMI1 and actual data detected the QTL; but in the other 21 cases, only AMMI1 detected it (none of these main QTLs were detected by only the actual data). Hence, the AMMI1 pre-processed PHS data with its greater accuracy resulted in twice as many detections of the six main QTLs. In 37 cases, AMMI1 achieved the larger LOD score, whereas in the other 4 cases the actual data achieved the larger LOD score (by a narrow margin). The average LOD score over 41 cases for the actual data was 3.90 and for AMMI1 was 8.26, for an average difference of 4.36 higher for AMMI1. For perspective, the average threshold for statistical significance at the 0.05 level was a LOD score of 2.9. The largest difference concerned QTL1 in Helfer 2002 with LOD scores of 18.1 and 1.6 for the AMMI1 and naïve estimators of PHS, respectively.

In addition, there were six QTL detections at locations other than QTL1 to QTL6 by the actual data, but none by AMMI1. Four detections involved rare QTLs documented by Munkvold et al. (2009),

namely *QPhs.cnl-4A.1* in Helfer 2003, *QPhs.cnl-5B.1* in Ketola 2004, and *QPhs.cnl-7D.2* in Caldwell 2002 and Helfer 2003. Two detections were previously undocumented, a QTL near marker wPT-5887 in Caldwell 2002 and one near E35M49161L in Helfer 2002. The average LOD for these six detections was only 3.4. Munkvold et al. (2009) list 65 QTL detections at the 0.05 significance level, so it is likely that several false positives were listed. These six detections by the less-accurate raw data merit some suspicion.

For the 20 detections of the main QTLs by both AMMI1 and actual data, the confidence intervals of the peaks were measured at 1 LOD below the peak. The average width for AMMI1 was 11.3 cM and for the raw data was 13.6 cM; so, there may be slight but unimpressive evidence that AMMI narrowed the confidence interval.

4.3.3. Understanding GEI

The AMMI1 biplot (Figure 4.5) illustrates the data's structure with the abscissa representing differences in main effects (broad adaptations) and the ordinate representing differences in interaction effects (narrow adaptations). For instance, Hel4 and Hel5 differ mostly by main effects, Ket5 and McG3 differ mostly by interaction effects, Ket5 and Cal2 differ in both respects, and Hel3 and Ket4 are similar in both respects. Each of the 197 genotypes are marked in this biplot by an integer from 0 to 3 indicating how many of the three largest-effect PHS susceptibility alleles (QTL1–QTL3) are present. There is an evident trend, with few of these susceptibility alleles in PHS resistant genotypes to the left and many of these susceptibility alleles in PHS sensitive genotypes to the right.

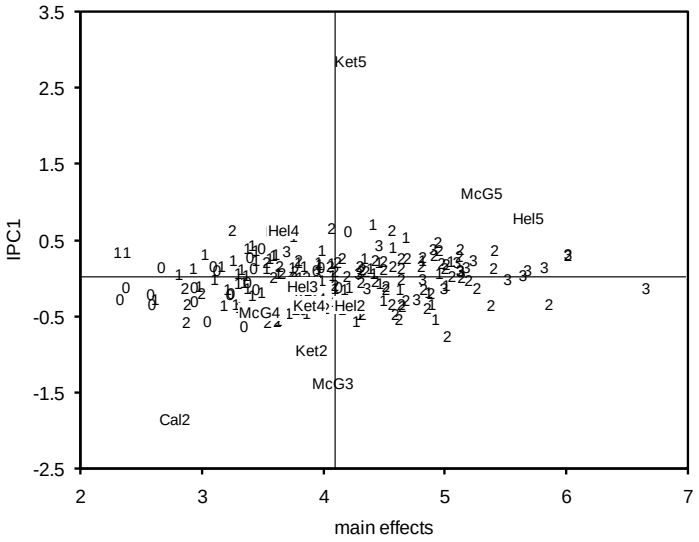

Figure 4.5. The AMMI1 biplot for the wheat PHS experiment. The abscissa shows main effects, namely genotype means (over environments) and environment means (over genotypes), and the ordinate shows IPC1 scores. The vertical line indicates the grand mean (a PHS score of 4.10) and the horizontal line indicates an IPC1 score of zero. Environment markers give the concise code names. Genotype markers are integers 0 to 3 indicating how many Caledonia alleles for QTL1 through QTL3 a genotype had.

Figure 4.6 shows the QTL scans based on the AMMI1 estimates of PHS scores and arranged in order by environment IPC1 scores. Unlike Figure 4.1, systematic trends are now quite evident. QTL3 has high

peak LOD scores at the top of this sequence and gradually declines moving down these scans. The smaller QTL5 and QTL6 also show that trend. In contrast, QTL2 has high peaks at the bottom of this sequence and gradually decreases moving up these scans. The smaller QTL4 has a similar trend. Finally, QTL1 is detected throughout these 11 scans. However, the scans used different scales so the LOD scores on the ordinates must be noted, showing that QTL1 has its highest peaks in the middle of this sequence. Remarkably, a single ordering of these scans, based on IPC1 scores capturing GEI information, brings all six QTLs into a single coherent, systematic pattern. Incidentally, besides these six QTLs, there is also a seventh trend, a small, non-significant peak in the unlinked markers near the right end that appears in several scans at the top but disappears at the bottom.

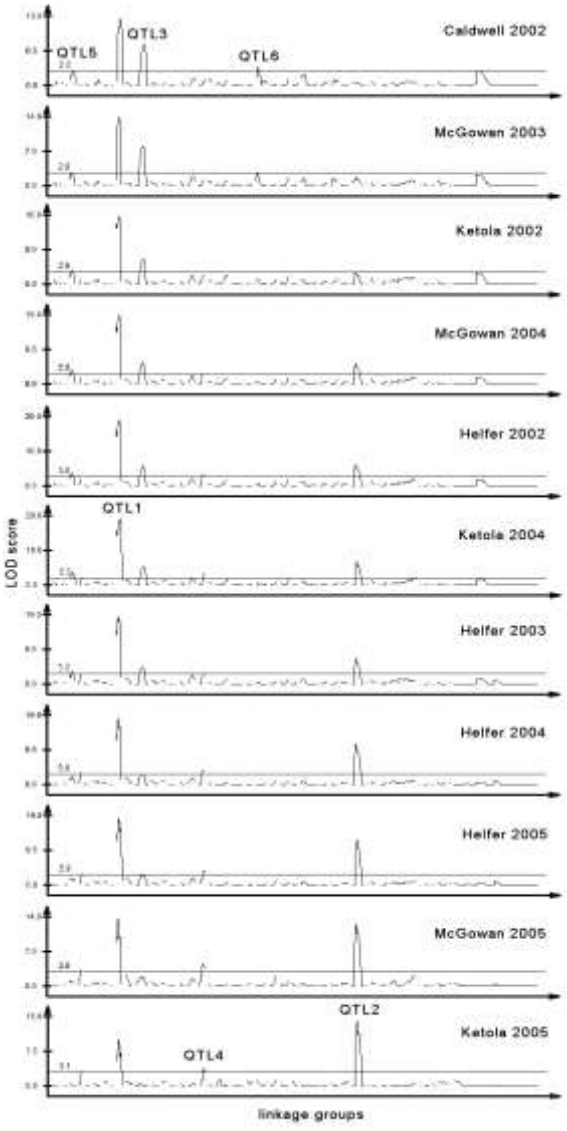

Figure 4.6. QTL scans for the 11 environments of the wheat PHS experiment, with the environments ordered by the environment IPC1 scores. These scans are based on the AMMI1 estimates. The combination of increased accuracy and systematic trend makes this figure much more informative than the starting point, the Figure 4.1. QTL3, QTL5, and QTL6 are expressed most strongly in environments shown at the top, whereas QTL2 and QTL4 have the opposite response and QTL1 has a quadratic response peaking near the middle.

When these scans were placed in the order of IPC2 or higher components, there were no evident trends, comparable to the arbitrary ordering in Figure 4.1. When these scans were placed in the order of environment means, the appearance was intermediate between an effective and a random ordering. Even though from Table 4.2 the environment means have a SS of 2523, which is several times larger than the IPC1 SS of 577, it is the IPC1 environment scores that produce an effective ordering of the QTL scans for these 11 environments. However, in the special case of linear regressions capturing nearly as much of the GEI as does IPC1, the genotype means and IPC1 scores are highly correlated, so arranging scans by genotype means will be very similar to arranging scans by genotype IPC1 scores.

The combination of greater accuracy and systematic trend provides some leverage in dealing with false positives and false negatives, that is, with inconsistent QTL detections. The systematic trend in Figure 4.6 adds credibility to the 41 QTL detections using the more accurate AMMI1 estimates of the PHS scores. For instance, QTL2 is detected in Helfer 2004 by AMMI1, but not by the raw data. In Figure 4.1, there is no basis whatsoever for arbitrating this discrepancy. But in Figure 4.6, there is every reason to accept this QTL detection as valid and to judge that the analysis of the raw data lead to a false negative. Regarding the six suspicious QTL detections by only the raw data, the strongest case is for *QPhs.cnl-7D.2* because it was detected in two environments, Caldwell 2002 and Helfer 2003. That case would gain credibility if these two environments were ecological neighbours. But in Figure 4.6, these two environments are far apart, so suspicion is warranted pending further experimental results. By discarding GEI noise and thereby gaining accuracy, AMMI pre-processing of phenotypic data provides for more reliable, consistent QTL detections. The difference between a random ordering of environments in Figure 4.1 and a systematic trend in Figure 4.6 permits the latter to communicate data patterns much more effectively.

Figure 4.7 provides an overall summary of QTL results. The abscissa shows the environment IPC scores, with the order for the 11 environments from left to right being the same as for the scans in Figure 4.6 from top to bottom. The ordinate shows the LOD scores. QTL2 and QTL4 increase to the right, whereas QTL3, QTL5, and QTL6 increase to the left. Linear regressions fit those five QTLs well, but not QTL1 ($R^2 = 0.0376$). Instead, QTL1 shows a quadratic response, peaking in the middle ($R^2 = 0.8154$).

Note that environments with similar IPC1 scores also have similar LOD scores for all 6 QTLs. Indeed, McG4, Hel2, Ket4, and Hel3 are virtually replicates. Although Hel2 was excluded from the analysis by Munkvold et al. (2009) because of experimental problems, it was not unusual in either its IPC1 score or its QTL scan, so it was included here. The IPC1 scores for the Hel location were relatively consistent over years but varied widely for the Ket and McG locations.

With the responses shown in Figure 4.7 in mind, some further remarks may be added about Figure 4.5 that showed the number of QTL1 to QTL3 present for each genotype in an AMMI1 biplot. If these QTLs are tracked individually, rather than collectively as in Figure 4.5, each showed a clear pattern. QTL1 had a large main effect, so it increased from left to right. The 197 genotypes were grouped into quintiles along the abscissa (or the ordinate), with 39 or 40 genotypes in each of these five groups. The percentages of the genotypes having QTL1 across the quintiles from left to right were 18%, 34%, 51%, 61%, and 92%. QTL2 had large interactions that were positive in Ket5, so it increased from bottom to top with

21%, 26%, 67%, 62%, and 69%. QTL3 had interactions of opposite polarity that were positive in Cal2, so it increased from top to bottom, though this smaller QTL had less dramatic results of 63%, 50%, 50%, 32%, and 45%. Hence, genotypes with or else without a given QTL showed clear segregation in the AMMI1 biplot for the three largest QTLs. However, the smallest QTLs, QTL4 to QTL6, did not show evident patterns.

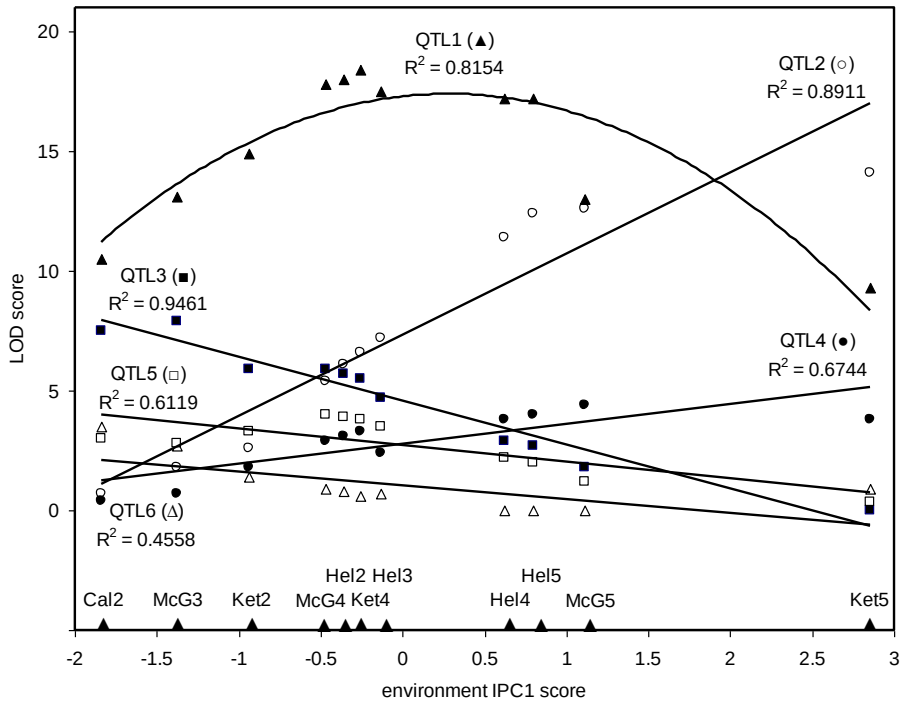

Figure 4.7. QTL expression as a function of environment IPC1 scores for the wheat PHS experiment. The abscissa shows the environment IPC1 score and the ordinate shows the LOD score. Results are shown for the six main QTLs, QTL1 to QTL6. QTL1 shows a quadratic response, peaking in the middle. QTL2 and QTL4 are expressed most strongly in environments like Ket5 at the right, whereas QTL3, QTL5, and QTL6 are expressed most strongly in environments like Cal2 at the left.

Although QTL1 to QTL6 interact with environment, there are no crossover interactions. That is, none of these QTLs have the allele from one parent increasing PHS in some environments and the allele from the other parent increasing PHS in other environments. Hence, PHS can be reduced across the entire range of growing conditions sampled here by a single genotype that includes the Cayuga allele at these QTL. It so happens that for all six of these QTLs, the allele for PHS sensitivity comes from the Caledonia parent. Obviously, the situation would have been more complex with crossover QEI, requiring different genotypes for different environments in order to optimize genotypes everywhere (Zhu et al., 1999, Voltas et al., 2002, Annicchiarico et al., 2005, Annicchiarico et al., 2009).

4.3.4. Predicting QTL scans

The tight relationship between IPC1 and the QTLs, which is evident in Figures 4.6 and 4.7, raises a question about whether the IPC1 scores are predictive of QTL scans. That possibility was examined by adding the 3 environments of the 2006 data for the CxC wheat experiment, for a new total of 14 environments. When adding (or removing) data, it is always possible that AMMI parameters may change radically, although this is less likely when the added (or removed) environments are generally like the others. In this case, the AMMI parameters for 11 and 14 environments were very similar, as confirmed by the correlation between the IPC1 scores for the 11 environments held in common being 0.9943. Accordingly, the predicted scan for each of the 3 new environments was simply the scan for the old environment having the closest IPC1 score.

In all three cases, the predicted scans are virtually indistinguishable from the corresponding observed scans already shown in Figure 4.6. Remarkably, for each and every new environment, its IPC1 score, which is a single number based on PHS data only, is highly predictive of its entire QTL scan.

4.3.5. Improving QTL detections

The choice of the AMMI1 model was based on maximizing the predictive accuracy of the estimates for the phenotypic data, as shown by the Ockham's valley in Figure 4.3. However, this model criterion in Figure 4.3 was indirect relative to the goal of optimizing QTL detection.

The criterion for quality QTL detections adopted here is the average LOD score over the 41 QTL detections for QTL1–QTL6. The choice of these particular 41 detections might be seen to favour AMMI1, perhaps virtually automatically, because the AMMI1 model was used to discover this particular roster of detections. However, the CxC experiment concerns field data, not simulated data, so the list of true QTLs in each of the 11 environments is not available. Accordingly, this roster of 41 detections is highly instructive, even if not completely definitive.

Figure 4.8 shows Ockham's hill for QTL detections. The abscissa shows the AMMI model applied to the phenotypic data. Results are shown for AMMI0–AMMI7 (but not AMMI8 and AMMI9 because the MATMODEL software used for AMMI analysis has a maximum of 7 IPCs). Also shown is the full model, AMMI10, which uses the naïve estimator equalling the actual data. The ordinate shows the average LOD score and there are three lines for three different sets of QTL detections. Beginning with the middle line in Figure 4.8, which shows the average LOD score for all 41 QTL detections, the best model with the strongest QTL detections is AMMI1. This happens to agree with AMMI1 also being the best model in Figure 4.3 concerning predictive accuracy for the phenotypic data. Recall that the average LOD score for AMMI1 was 8.26 and for the actual data was 3.90. The middle line in Figure 4.8 shows those results and also the values for additional AMMI models, including an average of 8.22 for AMMI0 and 7.31 for AMMI2.

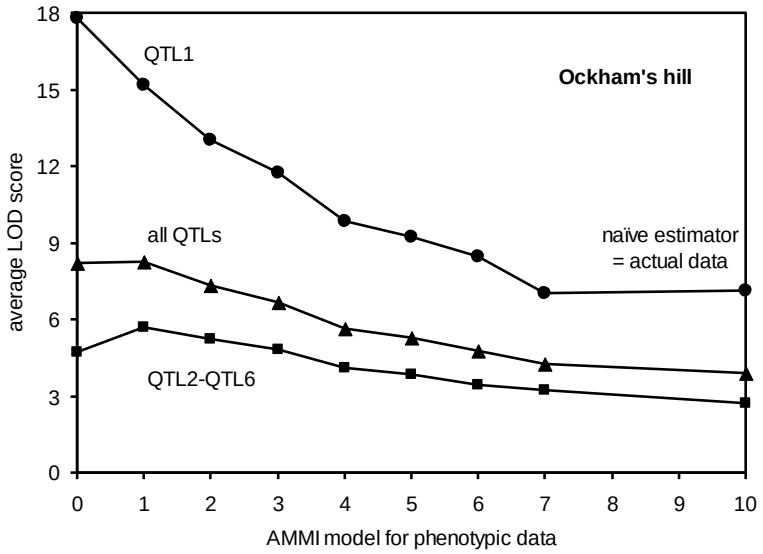

Figure 4.8. Ockham's hill for QTL detections for the wheat PHS experiment. The abscissa shows AMMI models of increasing complexity from AMMI0 to the full model, AMMI10. The ordinate shows the average LOD score and there are three lines for different sets of QTL detections: QTL1 with 11 detections, all QTLs with 41 detections, and QTL2–QTL6 with 30 detections. For detecting QTLs for main effects (namely QTL1), AMMI0 is the best member of this model family. But for detecting QEI effects (namely QTL2–QTL6), AMMI1 is best. The weakest QTL detections result from using the naïve estimator, the actual data, which includes a GEI that is 82.5% noise.

The main discrepancy between Ockham's valley for predictive accuracy in Figure 4.3 and Ockham's hill for all 41 QTL detections in Figure 4.8 is that the performance for AMMI0 is very close to AMMI1 for the latter. This discrepancy can be explained by disaggregating the QTL detections into two groups: the 11 detections for QTL1 shown by the top line in Figure 4.8, which mostly involve main effects, and the other 30 detections for QTL2–QTL6 shown by the bottom line, which mostly involve GEI. The top line for the average LOD score of the 11 detections for QTL1 shows AMMI0 as the best model, achieving an average of 17.79 that exceeds the 15.18 of AMMI1. Therefore, when attention was focused on main effects, AMMI0 performed best by discarding not only the GEI noise, but also the GEI signal prior to CIM. On the other hand, the bottom line for the average LOD score of the 30 detections for QTL2–QTL6 shows AMMI1 as the best model, achieving an average of 5.72 that exceeds both the 4.71 of AMMI0 and the 5.21 of AMMI2. Therefore, when attention was focused on GEI effects, AMMI1 performed best by discarding only GEI noise prior to CIM. Consequently, the best member of the AMMI model family may differ for QTLs associated with main and GEI effects. The nearly equivalent performance of AMMI0 and AMMI1 for all 41 QTL detections results from averaging over 11 detections for main effects (QTL1) with AMMI0 superior and 30 detections for GEI effects (QTL2–QTL6) with AMMI1 superior.

A striking feature of Figure 4.8 is the marked inferiority of the full model or actual data as compared to (several) parsimonious models. Indeed, Figures 4.3, 4.4, and 4.8 (and the comparison between Figures 4.1 and 4.6) show that in many ways, there are substantial penalties for conducting QTL scans with phenotypic data based on the naïve estimator.

4.4. Results for the barley experiment

4.4.1. Previous studies

Only one previous study has applied AMMI analysis to the SxM yield data, namely Romagosa et al. (1996), where four regions (QTL1–QTL4) of the barley genome were associated with differential genotypic expression for grain yield across environments. But several additional publications also examined QEI for the SxM yield data.

To further explore QEI using the SxM population, Zhu et al. (1999) crossed two selected DH lines (SM73 and SM145) from the original SxM population in order to accumulate favourable QTL alleles for grain yield. They used multiple regression and interval mapping procedures to explore phenotype and genotype relationships and concluded that all of their QTLs exhibited significant QEI. A QTL in chromosome 2 showed crossover QEI, that is, contrasting favourable alleles in different environments and/or the same environment in different years. However, they did not present any order or relation among the environments that could illustrate this feature.

Romagosa et al. (1996) used a second set of 92 DH lines derived from the SxM barley cross and planted them in 1995 and 1996 at Washington (WA95 and WA96) and in 1996 at Idaho (ID96) for verification of the QTL results in Romagosa et al. (1996). They confirmed the QTL detections of the previous work, namely QTL1 to QTL4. Their main conclusions were: (i) QTL1 on chromosome 3 is a consistent locus for determining yield across sites with the Steptoe allele being favourable; (ii) QTL3 on chromosome 6 also had a consistent but more limited effect on the yield across environments, with the Morex allele being favourable; and (iii) QTL2 in chromosome 2 and QTL4 in chromosome 4 were less consistent with expression affected by the environments. However, they "could not identify unique agro-climatic patterns of adaptation in these three sites."

Peighambari et al. (2005) used 72 DH lines from the SxM cross and planted them in Iran. They found QTLs for yield components only on chromosomes 1 and 5.

Piepho (2000) proposed a mixed-model method to detect QTLs with significant main effects across environments and to characterize the stability of those effects. He used the SxM barley population.

Malosetti et al. (2004) introduced a modelling framework for studying QEI using regression models in a mixed model context. They restricted their analysis of the SxM population to chromosome 2. The main conclusions of a preliminary analysis were: (i) a maximum for the QTL expression at 41.2 cM; (ii) the expression consisted exclusively of QEI, as no significant QTL main effect was present at this chromosome position; and (iii) the Steptoe allele had positive effects on yield in ID92, MTd91, and MTi91, whereas the Morex allele had positive effects on yield in MAN92, and SKs92.

Lacaze et al. (2009) used a subset of the SxM yield data with seven environments to study phenotypic plasticity, that is, the variation in phenotypic traits caused by environmental differences. They detected eight QTLs: two for main effects in chromosome 6, and six for QEI in chromosomes 1, 2, 3, 4 and 7.

4.4.2. Preliminary analyses

Following Romagosa et al. (1996), CIM was applied to all 16 environments, the main effects and PC1 to PC4, for a total of 21 QTL scans. The thresholds were obtained through a permutation test at the 0.05 level using 1000 permutations, and 5 QTLs were detected. These included the four detected by Romagosa et al. (1996): QTL1 on chromosome 3 in the interval between markers ABG399-BCD828; QTL2 on chromosome 2 in the interval ABC156A-ABG358; QTL3 on chromosome 6 in the interval CDO497-BCD340E; and QTL4 on chromosome 7 in the interval ABC324-ABC302. Also, we detected a fifth QTL on chromosome 2 near the marker ABC167B.

Table 4.3 gives the ANOVA for the AMMI7 model. As Romagosa et al. (1996) also comment, the proportions of the treatment SS due to genotypes, environments and GEI account for 7%, 70%, and 23%, so interaction is important, three times the magnitude of the genotype main effect. Since the error mean square is 0.423 and the interaction has 2235 df, the noise in the interaction may be estimated as 944.61 or 34.1% (Gauch, 1992, Voltas et al., 2002). From the interaction total SS of 2772.84, this leaves an estimated GEI signal of 1828.23 or 65.9%. Hence, just the real portion of the interaction, after discounting for noise, is still more than twice the magnitude of the genotype main effect. IPC1 captures a SS of 583.19, IPC2 captures 519.26, and thereafter these values decline rapidly. For comparison, the Finlay and Wilkinson (1963) genotype linear regressions on environmental means capture a SS of only 221.94, so AMMI analysis is considerably more effective for this dataset. These sizable GEI interactions imply that QEI effects are important.

Table 4.3. AMMI7 analysis of variance. This analysis is for yield of doubled haploid progeny from a cross between the barley varieties Steptoe and Morex. The grand mean is 5.28 MT/ha. The noise in the genotype-by-environment interaction (GEI) may be estimated by the interaction df times the error MS, namely 944.61, which by difference from the total of 2772.84 implies a GEI signal of 1828.23, or 65.93%.

Source	df	SS	MS	Probability
Total	3848	12555.35	3.263	
Treatments	2399	11942.94	4.978	0.0000000
Genotypes	149	814.09	5.464	0.0000000
Environments	15	8356.01	557.067	0.0000000
GEI	2235	2772.84	1.241	0.0000000
IPC1	163	583.19	3.578	0.0000000
IPC2	161	519.26	3.225	0.0000000
IPC3	159	384.50	2.418	0.0000000
IPC4	157	271.85	1.732	0.0000000
IPC5	155	243.59	1.572	0.0000000
IPC6	153	140.27	0.917	0.0000000
IPC7	151	134.17	0.889	0.0000000
Residual	1136	496.01	0.437	0.3034518
Error	1449	612.41	0.423	

4.4.3. Gaining accuracy

As before for the CxC wheat data, the first strategy here for detecting and understanding QEI is to increase the accuracy of the phenotypic data with a parsimonious AMMI model. The new phenotypic traits from the chosen AMMI model improve QTL detections and scans.

The differences with the first example are striking: the noise is much smaller (34.1% instead of 82.5%) and the interaction is much more complicated. Indeed, not until reaching the AMMI5 model does the sum of the eigenvalues add up to the estimated GEI signal.

In order to choose the best AMMI model for the SxM barley experiment, the jackknife procedure was used to estimate RMSPD. This barley experiment has 16 environments (and a larger number of 150 genotypes), so the entire AMMI family has 16 members with 0 to 15 IPC components. However, unlike before in Figure 4.3, this Ockham's hill exhibits no sharp peak. Instead, it is rather flat between AMMI3 and AMMI7, with the most predictively accurate model being AMMI5. However, because AMMI5 is so complex and the incremental improvement over AMMI3 is so slight, we decided to use AMMI3 in order to have a more parsimonious model. Table 4.3 shows that F tests suggest that 7 or more IPCs are statistically significant, but again these tests overestimate the number of IPCs.

Since we have chosen the AMMI3 model but QTL4 in Romagosa et al. (1996) was detected by IPC4, in this study we decided to focus our attention on their QTL1–QTL3 and our new QTL detection on chromosome 2. Henceforth, these four are called QTLa, QTLb, QTLc and QTLd, respectively.

There were 37 QTL detections of QTLa–QTLd for the 16 environments, including all detections by AMMI3 only, or actual data only, or both. In 23 cases, both AMMI3 and raw data detected the QTL; in 9 cases, only AMMI3 detected it; and in 5 cases, only the raw data. In 24 cases, AMMI3 achieved a larger LOD score; whereas in the remaining 13 cases, the raw data achieved the higher score. The average LOD score over the 37 QTL detections for the actual data was 6.92 and for AMMI3 was 9.38, for an average difference of 2.46 higher for AMMI3. There was no significant difference between the AMMI3 and raw data's thresholds. The overall average for statistical significance at the 0.05 level was a LOD score of 2.75.

As in the CxC wheat population, higher peaks were obtained for QTL scans when a parsimonious AMMI estimator was used instead of the naïve estimator to obtain more accurate phenotypic data. AMMI pre-processing of the phenotypic data improved QTL detections substantially, even though the accuracy gain for SxM was more modest than for CxC.

4.4.4. Understanding GEI

The interaction in the AMMI3 model for the SxM barley yield data is much more complex than in the AMMI1 model for the CxC wheat PHS data. And yet, given the limitations of two-dimensional paper, here the visualization of the GEI for the SxM data is further restricted to the AMMI2 model.

Figure 4.9 shows the AMMI2 biplot for the SxM barley yield data. The code names for the 16 environments were taken from Romagosa et al. (1996). This biplot depicts the first two IPCs capturing 39.76% of the GEI (21.03% for IPC1 and 18.73% for IPC2). But given that this GEI is only 66% signal, the more relevant observation is that this biplot captures about 60% of the GEI signal. For comparison, the Finlay-Wilkinson genotype regressions capture only 38.1% as much GEI as does IPC1.

This biplot identifies environment OR91 as an outlier. By contrast, the other 15 environments exhibit a clear trend.

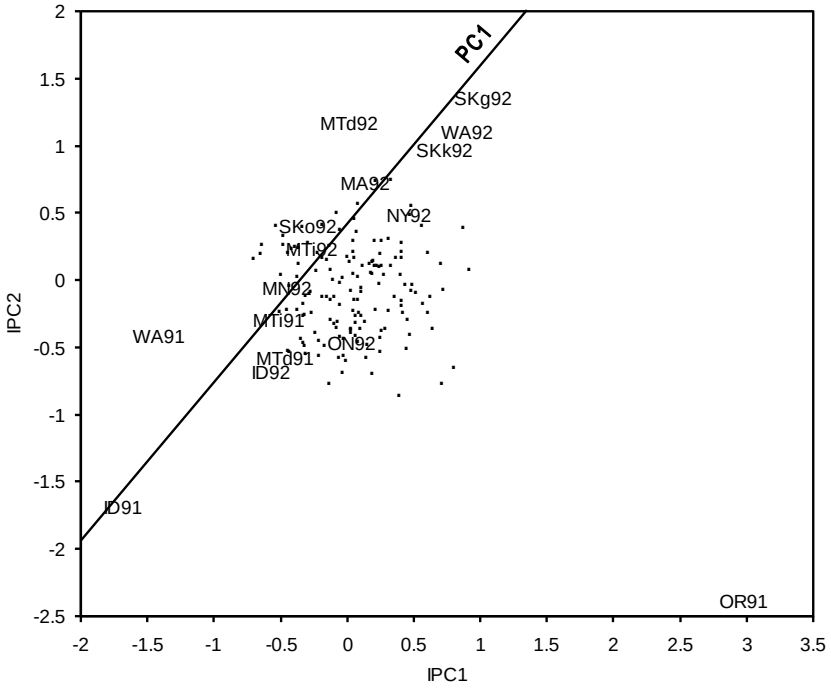

Figure 4.9. The AMMI2 biplot for the barley yield experiment. The abscissa shows the IPC1 scores and the ordinate shows the IPC2 scores. The 16 environments are marked by their code names and the 150 genotypes by dots. The first and second IPC capture 21.03% and 18.73% of the GEI, for a total of 39.76%. But since this GEI is only 65.93% signal, this graph captures approximately 60% of the GEI signal. Environment OR91 is an outlier, so the first principal component PC1 was fitted to the remaining 15 environments to obtain a systematic trend.

To determine a consensus ordering of the 15 environments within this two-dimensional biplot, PCA was used to fit a least-squares line (PC1) and then the points were projected perpendicularly onto this line. This resulted in a contrast between environments ID91 and SKg92, projected onto opposite extremes along PC1.

Incidentally, PCA is appropriate here because both sets of scores in Figure 4.9 have a similar magnitude of errors (Gauch, 1992:72–74). PCA minimizes the sum of squared residuals of its regression axis, thereby treating errors in the directions of both axes the same. By contrast, the regression of Y on X assumes no errors in X, and the regression of X on Y assumes no errors in Y, and neither of these is a plausible assumption in the present case. Letting X be the IPC1 environment score and Y be the IPC2 environment score, the PC1 drawn in Figure 4.9 captured 78.76% of the variance and its equation is

$$Y = 0.3890 + 1.1653X \tag{4.4}$$

Figure 4.10 shows the QTL scans arranged by the systematic trend provided by PC1 in Figure 4.9, using the AMMI3 estimates for the yields, with the outlier environment OR91 shown separately at the bottom. To make this graph with 16 scans manageable on a single page, some similar scans were grouped together. The trend obtained here cannot be as good as in the CxC wheat example because of the more complex GEI interactions. Nevertheless, the systematic trend using PC1 is impressive since the AMMI analysis only used the yield data.

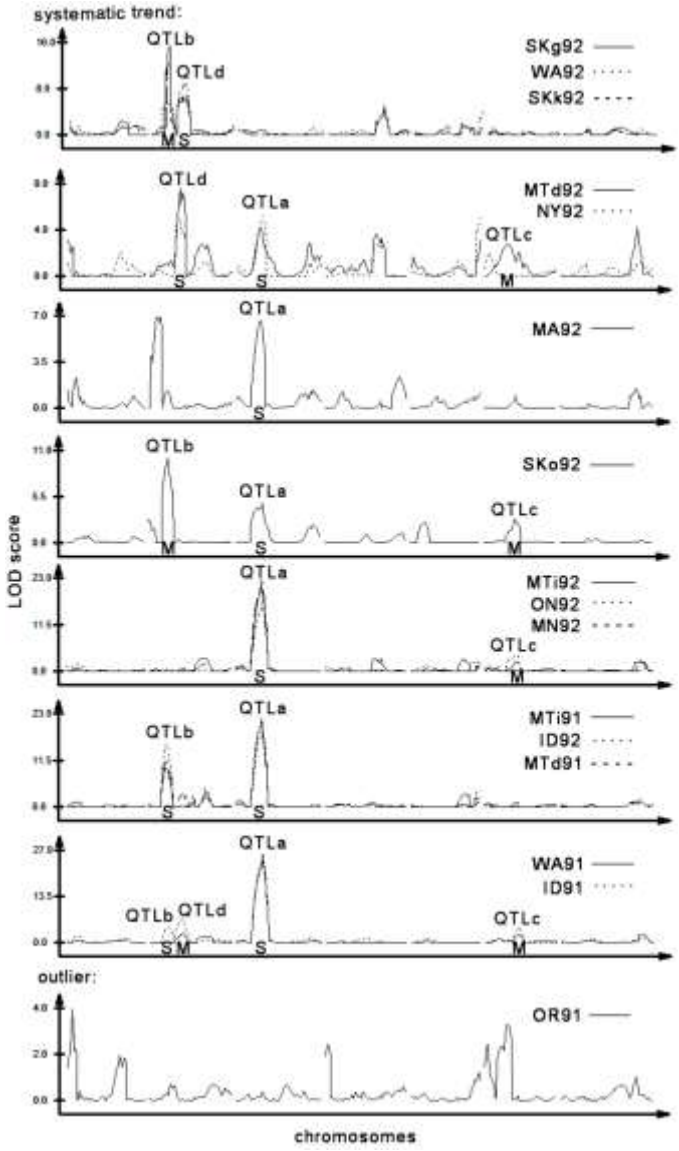

Figure 4.10. QTL scans for the 16 environments of the barley yield experiment, with the environments ordered by their PC1 scores in the AMMI2 biplot for those 15 environments showing a systematic trend, whereas environment OR91 is an outlier. In some cases, two or three environments with similar QTL scans are grouped to make this graph manageable despite its rather large number (16) of scans. Detections are noted for the 4 QTLs of primary interest, denoted QTLa–QTLd. Detections are marked "S" if barley variety Steptoe contributes the allele for higher yield, or else "M" if Morex. Note that QTLb and QTLd exhibit crossover QEI.

It would be difficult, or perhaps even impossible, to choose a better order for revealing patterns in QTL expressions. If these 16 scans are ordered by environment main effects, the patterns are weak, and likewise for environment IPC1 or IPC2 scores. Another alternative is to use a dendrogram, as in Figure 3 in Romagosa et al. (1996). But that dendrogram puts Sko92 and SKg92 adjacent despite their very different QTL scans. Likewise, ID91 and ON92 are adjacent, but ON92 has only two QTLs whereas ID91 has four. Numerous hierarchical classification algorithms exist, as well as many different similarity measures and metrics, and different choices could lead to quite different dendrograms. Also, clustering software often turns each bifurcation in an arbitrary manner, so the order shown is just a random choice

among many geometrically equivalent alternatives. For instance, cluster 1 in Romagosa et al. (1996) has 10 bifurcations and hence 2^{10} or 1024 possible orderings of the environments and cluster 2 has 2^4 or 16 possible orderings, for a total of 16384 possible orderings. One exception is a classification method called TWINSPAN that deliberately turns each bifurcation to place similar entities close together in the final ordering (Hill et al., 1975).

The parent causing each QTL in Figure 4.10 is indicated by an "S" for Steptoe or "M" for Morex. On the one hand, QTLa is always due to Steptoe and QTLc to Morex, as Romagosa et al. (1999) also concluded. On the other hand, QTLb and QTLd are sometimes due to Steptoe and sometimes Morex. However, unlike Zhu et al. (1999) and Romagosa et al. (1999), in Figure 4.10 a systematic pattern is evident. QTLb is due to Morex in the top of Figure 4.10 but due to Steptoe in the bottom. The opposite order is evident in QTLd. It may also be mentioned that the panel with the environments MTi91, ID92 and MTd91 agrees with Malosetti et al. (2004). Figure 4.10 shows a QTL on chromosome 2 slightly to the left of QTLb (and near marker ABG008) that Malosetti et al. (2004) also detected, but it appeared only in environment MA92, so it is not discussed further.

Figure 4.11 depicts an overall summary of QTL results for the SxM barley experiment. The abscissa shows the environment PC1 scores obtained in Figure 4.9 for the 15 environments in a systematic trend (but omitting the outlier, OR91). The order in Figure 4.11 from the left to the right is the same as in Figure 4.10 from the bottom to the top. The ordinate shows the signed LOD score, distinguishing the detections caused by Morex with positive LOD scores and by Steptoe with negative LOD scores. QTLa and QTLb increase to the right, whereas QTLc and QTLd increase to the left. Linear regressions were fitted to these 4 QTLs.

One may notice that these fits (R^2) are not quite as good as for the CxC wheat data. However, the QEI here is much more complex. Consequently, no single ordering or dimension can do as well as with the simpler CxC wheat case. Nevertheless, although less tidy, it is still an impressive and helpful result.

The coefficient of determination, R^2, in linear regression is simply the square of the sample correlation coefficient between the environment PC1 scores and the signed LOD scores. Accordingly, the test statistic for the correlation coefficient (and hence also for the coefficient of determination) used was

$$t = R\sqrt{\frac{n-2}{1-R^2}}, \tag{4.5}$$

where the observed value is compared with Student's t-distribution with n-2 degrees of freedom, where n is the sample size, namely 15. The p-values for QTLa–QTLd were <0.0001, 0.0027, 0.0022, and 0.0004, respectively, which means that all of these linear regressions are highly significant. Quadratic fits were also tried, but they were never an improvement.

The most interesting results in Figure 4.11 are for QTLb and QTLd, which show crossover QEI. For QTLb, the Steptoe allele increases yield in environments like ID91, whereas the Morex allele increases yield in environments like SKg92. By contrast, QTLd has a crossover QEI of opposite polarity. Given crossover QEI, it is not possible to simply pyramid QTLs in order to increase yield everywhere, but rather different mega-environments require different genotypes to optimize yield (Zhu et al., 1999). Incidentally,

some of the LOD scores for environment SKo92 are rather different from its close neighbours in Figure 4.11. The explanation is that this environment has a very distinctive and isolated location on IPC3, but components higher than IPC1 and IPC2 cannot be shown in any of our two-dimensional graphs.

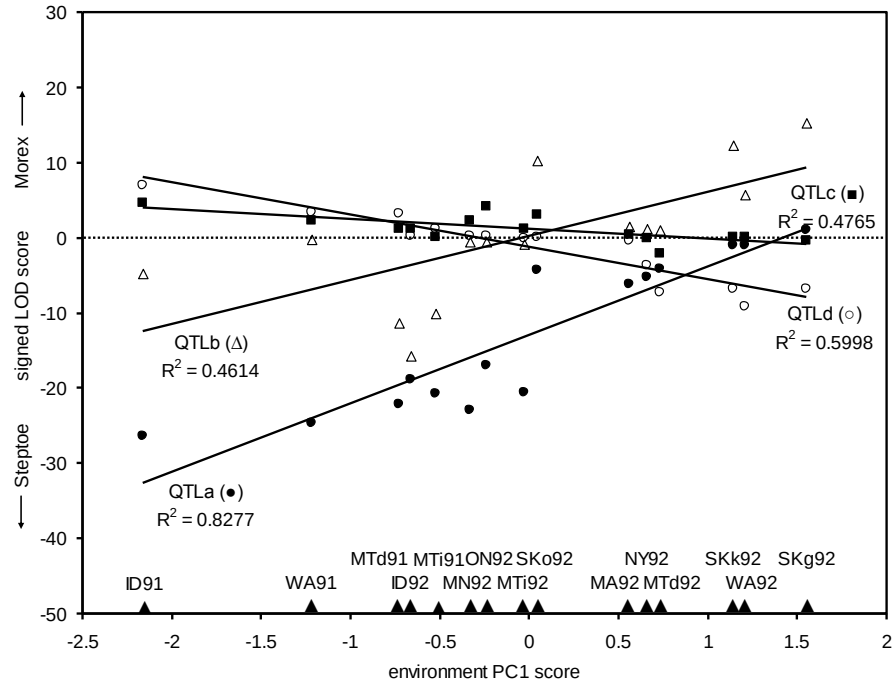

Figure 4.11. QTL expression as a function of environment PC1 scores for the barley yield experiment. The abscissa shows the environment PC1 score for those 15 environments displaying a systematic trend in Figure 4.10 (but not the outlier OR91). The ordinate shows the signed LOD score. The polarity is arbitrary, but QTL detections for higher yield due to the Morex allele were assigned positive LOD scores, and detections for Steptoe were given negative scores. The dotted horizontal line indicates a LOD score of 0. Results are shown for the four main QTLs, QTLa–QTLd, with linear fits. QTLa increases yield in environments like ID91, but has negligible expression in other environments like SKg92. QTLb shows a crossover QEI, with Steptoe contributing the allele for higher yield in ID91 but Morex in SKg92, whereas QTLd shows a crossover QEI of opposite polarity. QTLc has the smallest effects, increasing yield in environments like ID91.

4.5. Discussion

4.5.1. AQ analysis

The two new strategies for handling QEI developed here are: (1) using a parsimonious AMMI model to gain accuracy for the phenotypic data from a multi-environment QTL experiment and thereby to improve QTL scans and (2) using IPC scores to perceive systematic trends in QTL scans for individual environments and thereby to obtain more consistent and reliable results having predictive power and sometimes also ecological interpretability. This combination of AMMI analysis followed by QTL scans is here termed AQ analysis.

This way of conducting the AQ analysis was possible because the genetic and error variances are very similar across environments for the CxC wheat experiment and the SxM barley experiment. If these

variances were much different between environments, a weighted AMMI analysis should be used instead of the standard one. Another solution would be to use the mixed model equivalent to AMMI, the factor-analytic model.

Even prior to AMMI analysis, the likelihood of AQ analysis being helpful for other experiments can be judged from a few criteria. First, the potential for fruitful AQ analysis depends on the number of environments, with 10 or more being ideal and 6 or 7 being adequate in general, but meagre benefit is expected for only 3 or 4 environments. Similarly, if the trait is already known to exhibit exceptionally high heritability, that is, small GEI and high measurement accuracy, then AMMI analysis may be pointless. Beyond that, three statistics readily available from ANOVA are indicative of the relevance of AQ analysis: the SS for genotype main effects, GEI signal, and GEI noise. Our experience leads to the following rough guidelines. If the GEI signal is at least a third as large as the genotype effects, then GEI is important and hence AQ analysis is likely to reveal systematic trends of interest. Furthermore, if the GEI noise is at least a third of the GEI total, then noise is problematic and hence AQ analysis is likely to gain accuracy (unless the noise level is so high that the entire GEI is buried in noise, in which case merely the ANOVA portion of AMMI is relevant with genotype main effects but no interaction effects). These basic considerations are so few and simple that they leave room for occasional surprises, with AQ analysis turning out to be more helpful or else less helpful than expected. Nevertheless, these simple considerations provide a reasonable indication of the potential benefit of AQ analysis for a given dataset.

4.5.2. Direct and indirect criteria for model choice

The criterion for selecting AMMI1 as the best model was to maximize the predictive accuracy of the estimates for the phenotypic data, as shown by the Ockham's valley in Figure 4.3. But ideally model choice should directly optimize some criterion of primary interest. For the present study, the foremost research purpose is detecting QTLs affecting both main effects and QEI, that is, both broad and narrow adaptations. Consequently, the model criterion in Figure 4.3 is somewhat indirect relative to the foremost research objective.

Intuitively, it seems plausible that optimizing the accuracy of the phenotypic data will simultaneously optimize QTL detections – so the former is a suitable, easily calculated surrogate for the latter. But CIM involves complex calculations with subtle interactions among LOD scores assigned to the various genetic markers, so trust in this surrogate is not quite complete and automatic. Accordingly, it is worthwhile to perform model diagnosis for the AMMI family using some criterion that is a direct measure of successful QTL detection. Then model choice with the indirect criterion can be checked against model choice with the direct criterion, resulting in either confirmation or surprise, as the case may be.

For the CxC wheat experiment, AQ analyses revealed a systematic trend for the six major QTLs simultaneously. This is especially remarkable because the AMMI analysis uses only PHS phenotypic data. A plausible explanation is that the QTL trends revealed in Figure 4.6 are caused by some biological factor or ecological gradient. Furthermore, there is necessarily only one sizable underlying gradient for this CxC

experiment because were there more gradients, no ordering could be so systematic (as may be observed in cases where the data show more complex interactions). Unfortunately, the location-specific weather and environmental data needed to identify the presumed causal factor were unavailable for this experiment. We can only speculate from general weather data and soil factors that the cause may be an ecological gradient from less drought stress (especially from mid-June to mid-July) in the environments at the top of Figure 4.6 to more drought stress at the bottom of Figure 4.6.

4.5.3. Interpretation of AMMI parameters

The degree to which the IPC scores for genotypes or environments have an evident biological or ecological interpretation is highly variable across experiments. Often AMMI parameters are quite interpretable. In such cases, there may be an opportunity to predict QTL scans for new environments from knowledge of their positions along a known ecological gradient, such as warm to cool temperatures. That could provide an interesting complement or alternative to the present method that can predict QTL scans from a new environment's IPC1 score.

There are some especially effective methods for interpreting AMMI parameters in terms of environmental factors and genotypic traits, especially when these parameters are repeatable across locations or years or both (Annicchiarico et al., 2006). The strategy used by Voltas et al. (2002) in the context of barley breeding was to use AMMI first to get insight into the data, and then factorial regression described the GEI found by AMMI in terms of genetic, phenotypic, and environmental information constituting putative causal factors. Similarly, although they did not use AMMI, Yin et al. (2005) used factorial regression to model GEI as a function of environmental variables affecting barley, with particular interest in extrapolating QTL information from one environment to another. To better understand broad and narrow adaptations in durum wheat, Annicchiarico et al. (2009) looked for consistent patterns among three layers of information: AMMI parameters, morphophysiological traits, and molecular markers.

4.5.4. Number of mega-environments

The number of mega-environments that plant breeders and seed suppliers can manage is restricted by practical constraints, so the high numbers of mega-environments that go with high-order AMMI models rapidly become unmanageable. It should also be noted that only that part of the GEI involved in crossover interactions implicates mega-environments (Zhu et al., 1999, Voltas et al., 2002). Although predictable GEI (due to soils or consistent climatic differences across locations or whatever) increases the number of mega-environments, unpredictable GEI (due to within-site year-to-year climatic variation) will decrease it (Voltas et al., 2002, Annicchiarico et al., 2005).

For multi-year data at several locations, a particularly useful option is analysing genotype-by-location interactions (GLI) and QTL-by-location interaction instead of GEI and QEI (with environments representing location-year combinations), because only GEI effects due to locations (or other factors known in advance, such as crop management) could be exploited by growing or selecting specifically-

adapted genotypes where possible. That would be feasible through the same techniques here applied to GEI analysis, by adopting a suitable error term for testing GLI principal components under the assumption of years as a random factor, that is, genotype-by-location by year interactions or else average within-location genotype-by-year interactions, depending on the ANOVA model (Annicchiarico, 2002:37–39). Another advantage of analysing GLI instead of GEI is the lower complexity of the selected AMMI model, which derives from discarding non-repeatable genotype-by-location interactions (Annicchiarico, 2002:5–7). Consequently, even if complex GEI and QEI are of statistical and scientific interest and are best captured by AMMI2 or a higher model, the portion of the GEI and QEI of agricultural utility may be limited to that captured by a parsimonious AMMI1 model, which can be handled by the various kinds of graphs illustrated here. The AMMI1 model identifies the one largest piece of the GEI, which may be worth exploiting for its narrow adaptations within particular mega-environments.

Fortunately, few mega-environments may suffice to optimize yield throughout a growing region. Annicchiarico et al. (2009) needed only two mega-environments for durum wheat in Algeria, and Crossa et al. (1991) needed only two mega-environments for bread wheat in a huge international trial. Figure 4.5 in Gauch and Zobel (1997) showed an Ockham's hill for yield as a function of the number of mega-environments (defined by the number of genotypes winning in at least one environment, which increases with higher-order AMMI models).

4.5.5. Future prospects

The combination in AQ analysis of noise reduction and systematic trends improves QTL detections. In this study, there is evidence of substantial reductions in *both* false positives *and* false negatives. Nevertheless, in an actual field experiment, the true locations of QTLs cannot be known perfectly, and therefore false positives and negatives cannot be identified unambiguously. Although simulation experiments have other limitations, they can complement field experiments because all QTLs are known precisely by construction and hence all false positives and negatives can be diagnosed correctly. In this study, the understanding of how individual QTL interact with the environment informs the plant breeder as to their utility for improving the trait in different environments (Figures 4.7 and 4.11). Also, this analysis allows the breeder to select locations that optimize detection of specific QTL effects such as the Hel location for QTL1 in Figure 4.7. That location consistently maximized the LOD score for QTL1 over years. This can become especially important for evaluating large recombinant populations for fine mapping and gene cloning because it is impractical to grow them in multiple locations.

Further understanding of the advantages from the present AQ analysis, as well as from variations on the methods illustrated here (such as substituting Bayesian for CIM detection of QTLs; Zhang, et al., 2005), awaits application of AQ analysis to a larger number of field experiments together with complementary insights from simulation experiments. Also, statistical theory might elucidate some key relationships, such as the relationship between the level of noise in phenotypic data and the frequencies of false positives and false negatives in QTL detections. AMMI gains accuracy by fitting a parsimonious

model to the entire genotypes-by-environments matrix of phenotypic data. But algorithms for QTL detection could potentially combine strength across environments by incorporating systematic trends, rather than analysing each environment in isolation.

High-throughput genotyping has been widely acclaimed as a tremendous asset for plant breeders. But phenotypic data also enter into QTL scans and related research in molecular breeding. Currently the nearly universal practice of breeders is to use the noise-rich naïve estimator for the phenotypic data, thereby compromising accuracy and efficiency. For accelerating improvements in yield and other traits of agronomic importance in the future, the winning combination will be high-throughput genotyping *and* high-efficiency phenotyping. Neither can substitute for the other. Best practices are needed for both genotyping and phenotyping, especially since dramatic cost reductions in genotyping have rendered phenotyping the most costly part of QTL research and breeding.

The simple intuition that prompted this empirical investigation is the plausible expectation that better phenotypic data can result in more accurate QTL scans. But phenotypic data enter into many kinds of biological, agricultural, and medical research besides QTL scans, including association analysis and genomic selection. Accordingly, the principles illustrated here regarding Ockham's hill should also be considered in these other contexts. Furthermore, although the present examples concern agricultural crops, the new strategies developed here for detecting and understanding QEI concern statistical principles of equal applicability across microbial and plant populations when studied in multiple environments, and may be adapted to animal and human genetic studies.

4.6. Supplementary material

Figure S4.1. Scale for preharvest sprouting from 0, on the left, to 10, on the right.

Chapter 5

5. A complex trait with unstable QTLs can follow from component traits with stable QTLs: an illustration by a simulation study in pepper

Abstract

Complex traits are traits whose phenotypic variation is driven by a set quantitative trait loci (QTLs) that are typically environment dependent. The environment dependence of complex traits can be observed at the phenotypic level as genotype-by-environment interaction (GEI) and at the genetic level as QTL-by-environment interaction (QEI). Genetic improvement of complex traits requires strategies for dealing with GEI and QEI. We illustrate a strategy for modeling of GEI and QEI in complex traits that departs from dissection of a target complex trait in a number of component traits, where each of the component traits is purely genotype dependent. An eco-physiological genotype-to-phenotype model converts the set of genotype specific component traits into the complex target trait by integrating the components with environmental inputs over the duration of the growing season. For component traits with a simple genetic basis, consisting of a few additive QTLs, an attractive scenario for marker assisted selection of the corresponding complex trait appears. First, identify the QTLs for the components. For new genotypes, then use molecular markers linked to the QTLs to predict the phenotypes for the components. Subsequently, use an appropriate genotype-to-phenotype model to integrate the components with environmental inputs to produce predictions for the complex target trait. In this chapter, we demonstrate the viability of our modeling approach for complex traits by a case study in sweet pepper (*Capsicum annuum* L.). We developed a seven component eco-physiological model for yield in pepper and simulated for a back cross population yield and yield components, where the yield components were given a simple QTL basis. We show how credible patterns of GEI and QEI for yield can be simulated from genotype specific yield components with a simple QTL basis. Our results can be instrumental in breeding strategies for the improvement of complex traits.

To be submitted as: Rodrigues, P.C., Heuvelink, E., Bink, M.C.A.M., Marcelis, L.F.M. and van Eeuwijk, F.A. A complex trait with unstable QTLs can follow from component traits with stable QTLs: an illustration by a simulation study in pepper.

5.1. Introduction

Genotype-by-environment interaction (GEI) is the phenomenon that the performance of genotypes is dependent on the environment. For example, a genotype that is superior under well watered conditions may yield poorly under dry conditions. A trait that shows strong GEI is hard to predict, especially when it concerns predictions for new genotypes and new environments. GEI is common in complex traits, traits whose phenotypic variation depends on many genes, or quantitative trait loci (QTLs), with relatively small effects, that are also environment dependent. A common example of a complex trait is yield. Understanding of GEI can lead to better predictions of complex traits and is a fundamental for improvement of such traits.

In a statistical genetic context, GEI in a complex trait can be tackled by regressing the GEI part of phenotypic responses on molecular marker variation to identify QTLs that show environment dependency, or QTL-by-environment interaction (QEI). The QEI can be further modeled in relation to environmental covariables, so that GEI can be predicted from markers linked to QTLs for the complex trait and environmental inputs. A well-known class of genotype-to-phenotype (G-P) models that can be subsumed under this approach are mixed linear and non-linear models. Examples of this approach can be found in Boer et al. (2007), Malosetti et al. (2004), Malosetti et al. (2010) and van Eeuwijk et al. (2005).

A physiologically inspired alternative approach to GEI is based on crop growth simulation models. Crop growth models represent a class of G-P models based on prior biological knowledge (Spitters 1990, van Ittersum et al. 2003) that has proved to be useful for understanding GEI and QEI (van Eeuwijk et al., 2010, van Eeuwijk et al., 2005, Cooper et al., 2009, Bertin et al., 2010, Letort et al., 2008, Chenu et al., 2009). A particularly strong point of crop growth models in comparison to more statistical G-P models is that they contain explicit representations of development over time and especially this feature may be useful in describing GEI (Chenu et al., 2009). In recent years, a wide spectrum of physiological models was offered for better interpretation of GEI and QEI, that aimed at traits of varying complexity like yield (Yin et al., 2000, Tardieu, 2003, Yin et al., 2004, Chenu et al., 2008), leaf elongation (Reymond et al., 2003, Reymond et al., 2004, Chenu et al., 2008), chemical concentration in seed grains (Ishii et al., 2010) and fruit quality (Quilot et al., 2005).

Most papers that aim at combining crop growth modelling approaches with quantitative genetic approaches give little attention to an integral analysis and understanding of the patterns of GEI and QEI that occur across environments. Typically, phenotypic and genotypic variation is analysed environment by environment and integration of the results is done in a narrative way without the use of a formal statistical framework. In the current chapter, a major objective is to investigate GEI and QEI for a complex trait in relation to its known genetic and physiological basis, as generated from a crop growth model in which key physiological parameters have been assigned an explicit QTL basis. An important question to be answered is whether QTLs for a complex trait will appear at genomic locations where QTLs for component traits were known to be present and whether the QTLs for the complex trait will show environment dependency (QEI), although the component traits were known to possess no environment dependency at

all. To answer this and related questions, we will use simulations based on a model system. We will be generating yield as a complex trait in the species sweet pepper (*Capsicum annuum* L.). The choice for sweet pepper follows from the fact that this chapter is part of the European project Smart tools for the Prediction and Improvement of Crop Yield (EU-SPICY, www.spicyweb.eu).

Yield is a complex trait that is notoriously difficult to improve, due to many contributing QTLs that exhibit QEI. We want to know whether an approach that aims at dissecting a complex trait with QEI in a number of components without QEI is potentially viable. If a complex trait with GEI and QEI can follow from component traits without GEI and QEI, we would be able to predict the complex trait from simple molecular marker profiles for the component traits together with environmental inputs. Of course, a first critical condition to be fulfilled is that we can define a sufficiently flexible G-P model that translates component traits and environmental inputs into a complex trait with realistic variation. A second condition requiring fulfillment is that QTLs explain a sufficiently large proportion of the variation in the component traits.

As G-P model, we developed a relatively simple crop growth model containing a small set of component traits. By generating the component traits from a QTL basis, we achieve an integration of crop growth models and statistical genetic models in the spirit of Yin et al. (2000, 2004). We report below on the results of a simulation study in sweet pepper focusing on the questions of (i) whether credible patterns of variation in GEI and QEI for yield could be generated using a simple crop growth model, with seven physiological parameters that did not contain any GEI; and (ii) whether the main effect QTLs, without QEI, used to generate the component traits, the physiological parameters, could be identified in a QTL analysis for yield and whether these yield QTLs showed QEI.

The structure of the chapter is as follows. We first describe and motivate the structure of our pepper crop growth model. Our crop growth model is a genotype specific extension of a more general species specific crop growth model. The complex trait, yield, is produced from a small set of genotype specific and environment independent physiological component traits. Values for the component traits were based on prior experiments and literature. Environmental inputs were obtained from actual environmental characterizations in earlier growing seasons. For simulation purposes, breeding populations (back crosses) were simulated in which the variation in the components traits was assigned a genetic basis in terms of one or more underlying QTLs and some residual genetic variation. The simulation framework is thus defined by i) the structure of the crop growth model and ii) its inputs, the genotype specific component traits, generated from underlying QTLs, and the environmental inputs. After the description of this framework, we briefly describe some statistical techniques that will be used to analyse the simulated data for the patterns in GEI and QEI. All of these statistical and statistical genetic ways of analysing the simulated data can be seen as special types of sensitivity analyses. Finally, we will address interpretation of QTL analyses for component traits and the resulting complex trait yield. Pleiotropic QTLs for component traits and complex trait, where the first don't show QEI and the second do, may allow the identification of beneficial marker profiles for the complex trait.

5.2. Materials and methods

5.2.1. Description of the Model: genotype-to-phenotype model

The eco-physiological G-P model (Figure 5.1) is based on the LINTUL crop modelling approach (Spitters, 1990, Spitters and Schapendonk, 1990, van Ittersum et al., 2003). Cumulative dry matter production ($TDM_{i,j}$ for genotype i in environment j; g m^{-2}) is the product of cumulative intercepted light and light use efficiency ($LUE_{i,j}$; g mol^{-1}):

$$TDM_{i,j} = \sum_{t=t_{0,j}}^{t_{f,j}} \left[(1 - exp(-K_i \times LAI_{i,j,t})) \times I_{j,t} \right] \times LUE_{i,j}, \quad (5.1)$$

where $t_{0,j}$ and $t_{f,j}$ represent the first and last day of the growing season in environment j, K_i is the light extinction coefficient for genotype i, $LAI_{i,j,t}$ represents the leaf area index (m^2 leaf area m^{-2} ground area) for genotype i, in environment j on day t, $I_{j,t}$ represents the photosynthetic active radiation (PAR; mol m^{-2} d^{-1}) on top of the crop in environment j and on day t.

Light use efficiency is assumed to increase with CO_2 and with temperature according to a saturating response:

$$LUE_{i,j} = LUE_i^{max} \times \{1 - exp(c_j \times [CO_2]_j)\} \times \{1 - exp[-Z_i(T_j - T_{LUE,j})]\} \quad (5.2)$$

where LUE_i^{max} is the light use efficiency of genotype i, when both CO_2 concentration and temperature are not limiting $LUE_{i,j}$, c_j, Z_i and $T_{LUE,j}$ are scaling constants, T_j represents the 24-h average temperature. CO_2 concentration and temperature are kept constant over the whole growing season in our simulations.

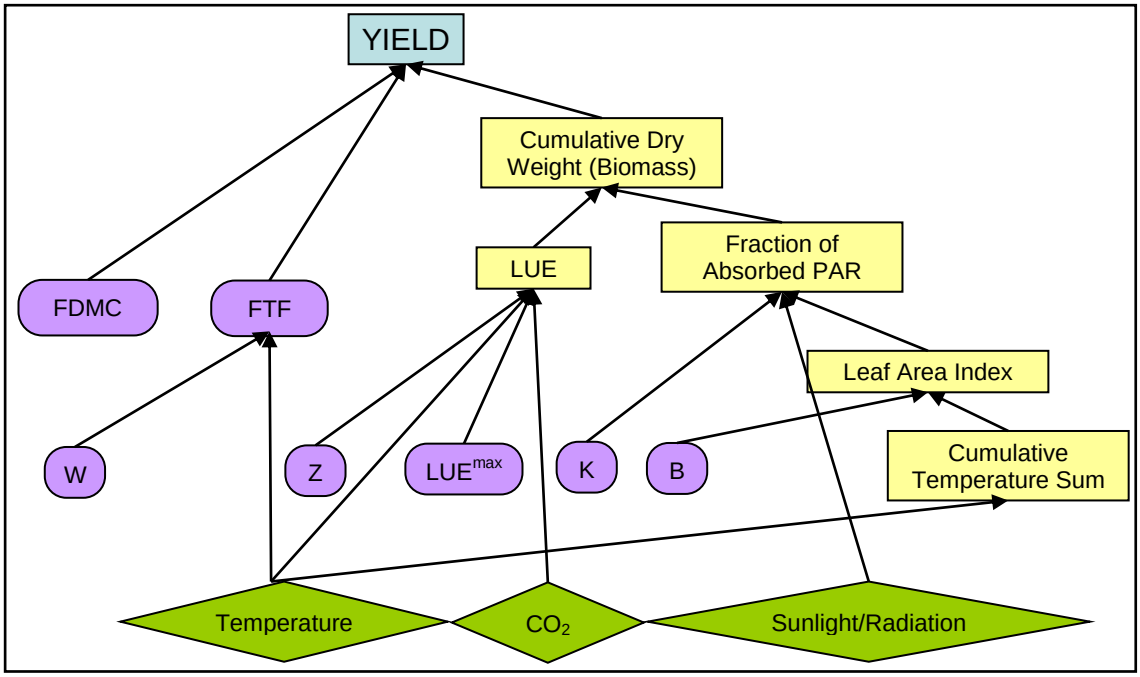

Figure 5.1. Schematic diagram of the crop growth model with seven physiological parameters. The diamonds represent input data, rectangles are states, ellipses are parameters and lines represent transfer of matter or information. The seven physiological parameters that are assumed to have genotype-specific values are: (1) maximum light use efficiency (***LUE^{max}***); (2) light extinction coefficient (***K***); (3) slope for the leaf area increase with temperature sum (***B***); (4) fraction of dry weight partitioned into the fruits (***FTF***, harvest index); (5) slope of the linear reduction in harvest index with temperature above 15ºC (***W***); (6) fruit dry matter content (*FDMC*); and (7) slope of the linear reduction in *LUE* for temperatures below 20ºC (***Z***).

The leaf area index ($LAI_{i,j,t}$) is the product of leaf area per shoot and shoot density, and is assumed to increase linearly with temperature sum (Marcelis et al., 2006). Considering a and B_i a genotype independent intercept and a genotype specific slope for the regression of leaf area per stem (m²) on temperature sum (ºC d), $(T_j - T_{base}) \times (\mathrm{t} - \mathrm{t}_0)$, the LAI for genotype i in the environment j at day t, $LAI_{i,j,t}$, can be calculated as follows:

$$LAI_{i,j,t} = \left[a + B_i(T_j - T_{base}) \times (t - t_0)\right] \times Sd, \tag{5.3}$$

where T_{base} is the base temperature, t represents the t-th day of the growing season ($t = t_0$ is the day of the first flowering), and Sd is the stem density.

The photosynthetically active radiation (PAR) incident on the crop on day t in environment j, $I_{j,t}$, is the product of (i) global radiation at day t in environment j, $RAD_{j,t}$, (ii) fraction of PAR in global radiation (F_{PAR}), and (iii) greenhouse transmissivity in environment j (Tr_j), i.e.

$$I_{j,t} = RAD_{j,t} \times F_{PAR} \times Tr_j. \tag{5.4}$$

Fresh fruit yield is calculated from cumulative dry matter production by multiplying the latter with a partitioning index ($FTF_{i,j}$) and dividing by fruit dry matter content ($FDMC_i$):

$$Yield_{i,j} = TDM_{i,j} \times FTF_{i,j} \times \frac{1}{FDMC_i}. \tag{5.5}$$

Partitioning index $FTF_{i,j}$ decreases linearly with temperature in a genotype specific manner

$$FTF_{i,j} = FTF_i \times \left(1 - W_i \times \left(T_j - T_{FTF}\right)\right), \tag{5.6}$$

in which W_i and T_{FTF} are scaling constants.

To the yield figures resulting from the application of the crop growth model (5.5), a normally distributed error was added for each environment individually, such that the coefficient of variation amounted to 10%.

5.2.2. Parameterization of the model

A greenhouse sweet pepper breeding population of genotypes was created by assigning values to the seven genotype specific, and environment independent, physiological component traits of the G-P model above. For each trait we assumed a Gaussian distribution with mean values based on *a priori* knowledge, as specified in Table 5.1. The component traits were assumed to be independent and the coefficient of variation was assumed to be 0.10 for each of them. Table 5.2 presents values for a series of constants in the G-P model.

5.2.3. Environments

Thirty six environments were defined, a 3×2×2×3 full factorial combination of four environmental factors: 1) three levels of daily radiation based on annual weather data for 2) two countries (1994, 2000 and 2008 for Spain, and 1998, 2003 and 2007 for The Netherlands; a year with low radiation, a year with high radiation and a year with an average radiation level), 3) two levels of CO_2 concentration (370 µmol

mol^{-1} – open environment, and 1000 μmol mol^{-1} – closed greenhouse with CO_2 enrichment), and 4) three levels of daily average temperature (15, 20 and 25ºC). In addition, the growing season in Spain was considered to start on September 10 and end on April 30 (232 days) and in The Netherlands from January 10 to November 30 (324 days). The greenhouse transmissivity (*Tr*) was considered to be 0.75 for The Netherlands and 0.60 for Spain because usually high tech glass greenhouses are used in The Netherlands and plastic greenhouses in Spain.

Table 5.1. The seven genotype specific, environment independent physiological parameters in the yield model, parameterized for greenhouse sweet pepper. For each parameter the mean value and the standard deviation (s.d.) are given. The last column presents the references for the chosen values. (index *i* refers to genotype *i*).

Parameter	Mean	s.d.	Reference
LUE_i^{max} 1)	0.87	0.174	(Nederhoff, 1994, Heuvelink, 1995)
Z_i 2)	0.6	0.05	(de Swart et al., 2006)
K_i	0.7	0.04	(Marcelis et al., 1998)
B_i	0.000378	3.78×10^{-5}	(Marcelis et al., 2006)
FTF_i	0.65	0.04	(Rijsdijk and Houter, 1993, Gelder et al., 2007)
W_i 3)	0.04	0.011	(Wubs et al., 2009, Wubs et al., 2010)
$FDMC_i$	0.0774	0.00508	(Wubs et al., 2009)

1) Mean value of LUE_i^{max} and c (Table 5.2) are chosen such that LUE at a CO_2 concentration of 370 μmol mol^{-1} is 0.65 g DM mol^{-1} PAR (Heuvelink, 1995) and the relative increase LUE when CO_2 concentration rises to 1000 μmol mol^{-1} agrees with Nederhoff (1994).

2)Mean value of Z_i and T_{LUE} (Table 5.2) are chosen such that LUE is not much different between 20 and 25°C, but is reduced at 15°C in agreement with De Swart et al. (2006).

3)Mean value of W_i and T_{FTF} (Table 5.2) are chosen such that the linear reduction in fraction partitioning to the fruits for temperatures about 15ºC agrees with Wubs et al. (2009, 2010).

Table 5.2. Parameterization of the constants in the model for sweet pepper. For each constant, the equation number, the chosen values and reference/section with further explanations are given.

Constant	Equation	Value (s)	Reference/Section
$t_{0,j}$	(5.1)	January 10 (NL); September 10 (SP)	Parameterization of the model
$t_{f,j}$	(5.1)	November 30 (NL); April 30 (SP)	Parameterization of the model
c_j	(5.2)	-0.004	(Nederhoff, 1994, Heuvelink, 1995)
$CO_{2,j}$	(5.2)	370, 1000 (μmol mol^{-1})	Parameterization of the model
$T_{LUE,j}$	(5.2)	13 (ºC)	(de Swart et al., 2006)
a	(5.3)	0.03372	(Marcelis et al., 2006)
T_{base}	(5.3)	10 (ºC)	(Marcelis et al., 2006)
Sd	(5.3)	7 (per m^2)	Common practice in The Netherlands
T_j	(5.3)	15, 20, 25 (ºC)	Parameterization of the model
$RAD_{j,t}$	(5.4)	Numerical variable	Historical data
F_{PAR}	(5.4)	0.5	(Goudriaan and Laar, 1994)
Tr_j	(5.4)	0.75 (NL); 0.60 (SP)	Parameterization of the model
T_{FTF}	(5.6)	15 (ºC)	(Wubs et al., 2009, Wubs et al., 2010)

5.2.4. Simulation of the population

We want to study the generation of GEI for a complex trait by simulating yields for a set of genotypes belonging to a segregating breeding population, where we chose a back cross for simplicity, using a

genotype specific crop growth model (LINTUL). In our crop growth model we combined genotype specific physiological parameters, without GEI, with environment specific inputs to produce yield with GEI. To investigate whether stable QTLs for physiological parameters, i.e., without GEI or QEI, could be translated into unstable yield QTLs, i.e., with GEI and QEI, we simulated for a series of runs of the crop growth model, a population of 500 back cross lines. This population size was expected to produce clear test profiles for the QTLs and is becoming realistic in current QTL studies.

Chromosome lengths and numbers of markers were based on the pepper population described by from Barchi et al. (2007). Marker positions were drawn from a continuous uniform distribution defined over the full length of the corresponding chromosomes, and ensuring markers to appear at both ends of each chromosome (Figure 5.2). Marker positions and alleles were generated by the function *sim.map* in package qtl of Software R (Broman and Sen, 2009).

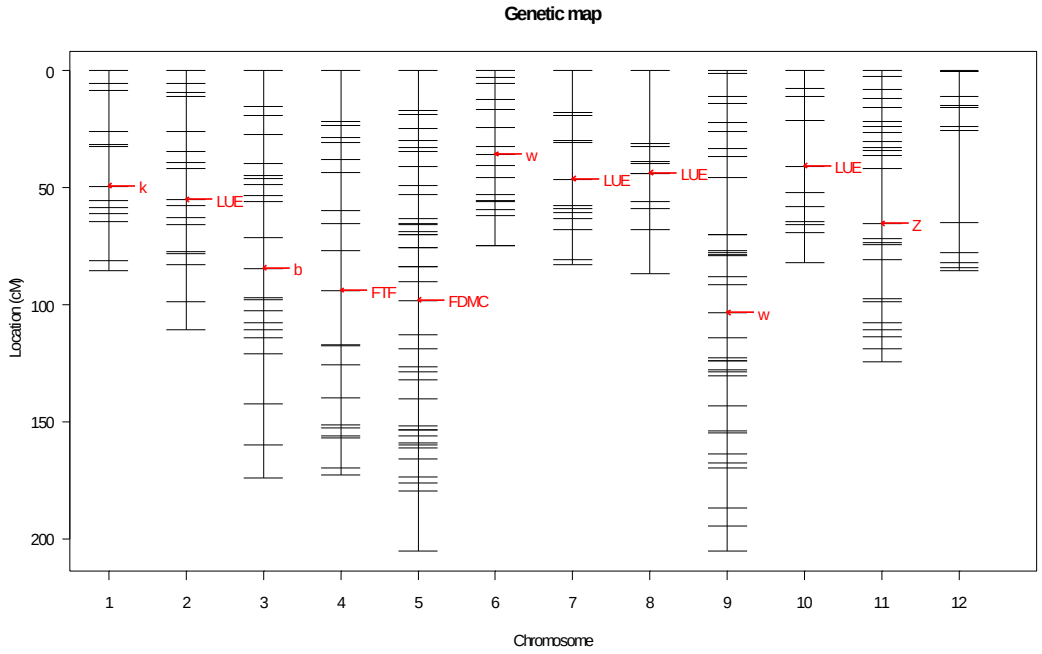

Figure 5.2. Genetic map for pepper, based on the lengths of chromosome and number of markers per chromosome in Barchi et al. (2007) and Barchi et al. (2009). The marker positions were taken as random. The arrows with the name of the 7 physiological parameters point the place where the QTL were placed.

QTLs underlying the seven physiological parameters were allocated as described in Table 5.3. QTL genotypes were converted into yield phenotypes by (5.5). Since the physiological parameters LUE^{max} and W were found to have stronger impact on the final phenotypic data (higher proportion of variance explained in sensitivity analysis – Table 5.4, and higher $-\log_{10}$(P-value) values in a preliminary QTL analysis), we decided to make these parameters dependent on more than one QTL.

Yield given by (5.5) depends on seven physiological parameters, each of which was considered to depend on a given number of QTLs (Figure 5.2 and Table 5.3). We can also express the physiological parameters of (5.5) in terms of QTL effects. For example,

$$LUE_i^{max} = g\_LUE_i^{max} = x_i\alpha + g^*\_LUE_i^{max}, \quad (5.7)$$

where we first assume the phenotypic differences for the physiological parameters to be equal to the genetic differences, or, the heritability for the physiological parameters is 1. Next, we partition the genetic differences in a QTL part and a genetic residual, with x_i representing the QTL genotype, a function of flanking marker genotype information, α the QTL allele substitution effect, assumed to be constant across all environments, and $x_i\alpha + g^*\_LUE_i^{max}$ the residual of the genetic effect for LUE^{max}. A development as in (5.7) can be inserted in (5.5) for each of the physiological parameters. Before inserting the component traits as defined in (5.7), the mean and variance of the component traits were scaled to comply with the specifications given in Table 5.1.

Table 5.3. Genetic architecture in the studied model. For each physiological parameter, the number of QTL responsible for its genetic variation, the location of the QTL and their heritability are presented. All the 11 QTL were placed next to the closest marker to the middle of the given chromosome.

Parameter	Number of QTL	Location of the QTL (Chromosome)	h^2
LUE^{max}	4	2, 7, 8, 10	0.12 each
W	2	6, 9	0.16 each
FTF	1	4	0.64
$FDMC$	1	5	0.64
Z	1	11	0.64
K	1	1	0.95
B	1	3	0.95

A population of 500 back cross lines was simulated for each of the 36 environmental conditions described above, resulting in a two-way table with 500 genotypes (rows) and 36 environments (columns). These simulations were used in an extensive study of GEI and QEI. As a final step in the simulation, for each environment the realized average yield was calculated and subsequently a normally distributed error was added to the yields such that the coefficient of variation became 10%. For investigating the robustness of the GEI and QEI patterns that were generated, also data were generated with CV's of 20% and 30%.

5.2.5. Sensitivity analyses

When dealing with a complex simulation model that depends on a set of parameters, a question will arise on the absolute and relative importance of the individual parameters. This question can be answered with a sensitivity analysis. The assessment of the relative importance of the individual component physiological parameters on the complex target trait yield was accomplished by applying various well known statistical methods for investigating two-way tables of genotype-by-environment means: factorial regression (van Eeuwijk et al., 1996), AMMI analysis (Gollob, 1968, Mandel, 1969, Gauch, 1988), and principal component analysis, or GGE analysis (Yan and Kang, 2002). A brief overview of all these techniques is presented in van Eeuwijk (1995).

5.2.6. Factorial regression

Factorial regression can best be understood as the imposition of contrasts on the levels of the row and column factor in a two-way table. We can use contrasts in the direction of the genotypes to partition the original variation between genotypes in a part due to a contrast and a residual. This is valid for both the genotype main effect and the GEI. Our intention is to use the yield components to define the contrast on the genotypes. In a sense, this use of contrasts is very similar to regression or covariance analysis. As the yield components were generated to be uncorrelated, the interpretation of the decomposition of genotype main effect and GEI is relatively straightforward. For the environments, we have a more detailed look at the variation due to the 36 environments by focusing on the parts of the environmental main effect and the GEI that can be attributed to the initial four generating factors: Country, Temperature, CO_2 and Radiation.

5.2.7. Bilinear models: AMMI and GGE

As a follow up on the above sensitivity analyses by factorial regression, where we used explicitly defined covariates or contrasts, we also studied the series of simulated yields for the 500 back cross lines in 36 environments with explorative linear-bilinear techniques (van Eeuwijk, 1995). These techniques combine additive and multiplicative terms. Well known representatives of this class of models are: 1) the model underlying principal components analysis (PCA) of the genotype-by-environment table, also called GGE biplot model (with GGE standing for genotypic main effects and GEI), see Yan and Kang (2003), and 2) the additive main effects and multiplicative interaction (AMMI) model, which is a combination of analysis of variance (ANOVA) for the genotypic and environmental main effects and PCA for the residuals from additivity (Gollob, 1968, Mandel, 1969, Gauch, 1988, Gauch, 1992). The PCA/GGE model is

$$y_{i,j} = \mu + E_j + \sum_{n=1}^{N} b_{i,n} z_{j,n} + \varepsilon_{i,j} , \quad (5.8a)$$

while the AMMI model can be written as,

$$y_{i,j} = \mu + G_i + E_j + \sum_{n=1}^{N} b_{i,n} z_{j,n} + \varepsilon_{i,j}, \quad (5.8b)$$

where $y_{i,j}$ is the yield of genotype i in environment j, μ is the grand mean, G_i are the genotype mean deviations (genotype means minus the grand mean), E_j are the environment mean deviations, $b_{i,n}$ and $z_{i,n}$ are the genotypic and environmental parameters (scores) for the n-th multiplicative interaction term (i.e. the genotype and environment principal component scores and loadings for PCA axis n), N is the number of principal component (IPC) axes retained, and $\varepsilon_{i,j}$ is a residual. In the PCA/GGE model, the genetic main effects and GEI are modelled simultaneously in terms of multiplicative terms, whereas in the AMMI model only the GEI is modelled multiplicatively.

5.2.8. QTL analysis

For the QTL analysis we used the mixed model QTL framework described by (Malosetti et al., 2004, Boer et al., 2007, Malosetti et al., 2010) as implemented in GenStat (Payne et al., 2011) A major point of

interest was whether could detect QEI for yield and see whether we could interpret this QEI in terms of underlying QTLs for the physiological parameters. The QTL model that we used uses explicit marker derived information to describe the GEI in terms of QTLs in their dependence on the environments (i.e. the QEI). The inclusion of this marker information, genetic predictors, allows testing whether the phenotypic trait (e.g. yield) is affected by the DNA at a particular genome position, and whether this effect depends on the environment. A mixed linear model definition following (Boer et al., 2007) is

$$\begin{aligned} y_{i,j} &= [\mu + E_j] + [G_i + (G.E)_{i,j}] \\ &= [\mu_j] + [\textstyle\sum_{p=1}^{P} x_{k,i}\alpha_{k,j} + \varepsilon_{i,j}] \end{aligned} \quad (5.9)$$

where μ_j is the intercept for each environment, $x_{k,i}$ is derived from marker genotype information for genotype i, $\alpha_{k,j}$ the QTL allele substitution effect for environment j, P is total number of QTL underlying $y_{i,j}$ (e.g. yield), and $\varepsilon_{i,j}$ follows a multivariate normal distribution with zero mean vector and a given variance-covariance (VCOV) matrix. The choice of the best VCOV structure was done following the procedure described in Malosetti et al. (2004) and Boer et al. (2007).

5.3. Results

5.3.1. Factorial regression analysis

Table 5.4 shows the results of various types of factorial regression on the simulated genotype-by-environment tables of means. For the genotype main effect in yield, we see that especially the variation in LUE was important, while W, FTF and $FDMC$ contributed to a lesser extent to consistent yield differences across the 36 environments. For the GEI, from the genotypic point of view, W seems the most important variable, followed by LUE.

For the environmental main effects, Country was the most important factor. Temperature and CO_2 were about half as important as Country, while Radiation was again about half of Temperature and CO_2, and about 4 times less than Country. Combinations of environmental factors were not found to add substantially to the average differences between environments. For the GEI, it is mainly Temperature that had influence, while Country also had influence, but three times less than Temperature. Other factors or factor combinations could be ignored.

5.3.2. GGE and AMMI analysis

Figure 5.3 shows the GGE biplot. The variation due to environments follows principally from temperature differences, in correspondence with the results of the factorial regressions. Figure 5.3 shows zones of cross over interactions between temperature regimes in the sectors II, III, IV and V. Sector VI shows genotypes that were above average in yield everywhere, sector I shows genotypes that were below average everywhere. Sector II shows genotypes that were below average in 20 degree environments, but above average in 25 degree environments, sector III shows genotypes that were below average in 15

degree environments, but above average 20 degree environments. In a similar way, the sectors IV and V can be interpreted in terms of cross over interactions.

The interpretation of GGE and AMMI biplots is very similar. For more details on the interpretation of GEI, see the description of the AMMI analysis below. Figure 5.3 shows that cross over interactions can be generated for the complex trait yield from a set of component traits without yield.

Table 5.4. Sensitivity analysis for the physiological parameters and environmental characterizations (3 temperature levels, 3 radiation levels, 2 countries and 2 CO_2 concentrations) based on the factorial regression model. The percentages for the main effects show the quotient between the type II sum of squares for each of the parameters [environments] and the type II sum of squares for the genotypic [environmental] main effects of the additive model. The percentages for the interaction show the quotient between: (i) the sum of type II sum of squares for the interaction between the physiological parameter [environmental variable] and each of the environmental variables [physiological parameters], and (ii) the type II sum of squares for the residuals of the additive model. The minimum (min), maximum (max) and average are obtained based on 10 random runs of the eco-physiological crop growth model.

	Main effects			Interaction		
	Min	Max	Average	Min	Max	Average
Physiological Parameters						
B	0.66%	0.90%	0.74%	0.27%	0.51%	0.39%
K	0.21%	0.46%	0.30%	0.09%	0.24%	0.17%
Z	0.29%	0.51%	0.37%	1.16%	1.51%	1.32%
LUE	**64.85%**	**74.81%**	**69.82%**	13.71%	14.29%	13.93%
W	11.69%	13.23%	12.68%	**24.59%**	**27.08%**	**25.88%**
FTF	6.09%	6.87%	6.43%	0.81%	1.58%	1.32%
FDMC	7.15%	8.41%	7.66%	1.14%	2.03%	1.59%
Environmental Variable						
Country	**41.04%**	**41.98%**	**41.49%**	8.78%	9.57%	9.16%
Temperature	19.50%	20.19%	19.85%	**28.49%**	**30.37%**	**29.41%**
CO2	16.36%	17.12%	16.71%	3.64%	4.32%	3.92%
Radiation	9.10%	9.63%	9.37%	1.90%	2.32%	2.11%

Note: All other combinations of genotypic parameters and environmental variables represent at most 1.1% of the main effects or interactions.

Table 5.5 gives a summary ANOVA table for the AMMI2 model. The ranges for the proportions of variance explained by genotypes, environments and GEI were [0.30; 0.36], [0.55; 0.64], and [0.12; 0.14], respectively. In the simulated phenotypic data, the GEI was responsible for about 29.0% (mean value for the 10 runs, with values between 28.1% and 29.7%) of the genotype related sum of squares (SS), i.e. genotype main effects plus GEI, GGE). The first two IPC were responsible for 16.0% of the GGE SS (Table 5.5).

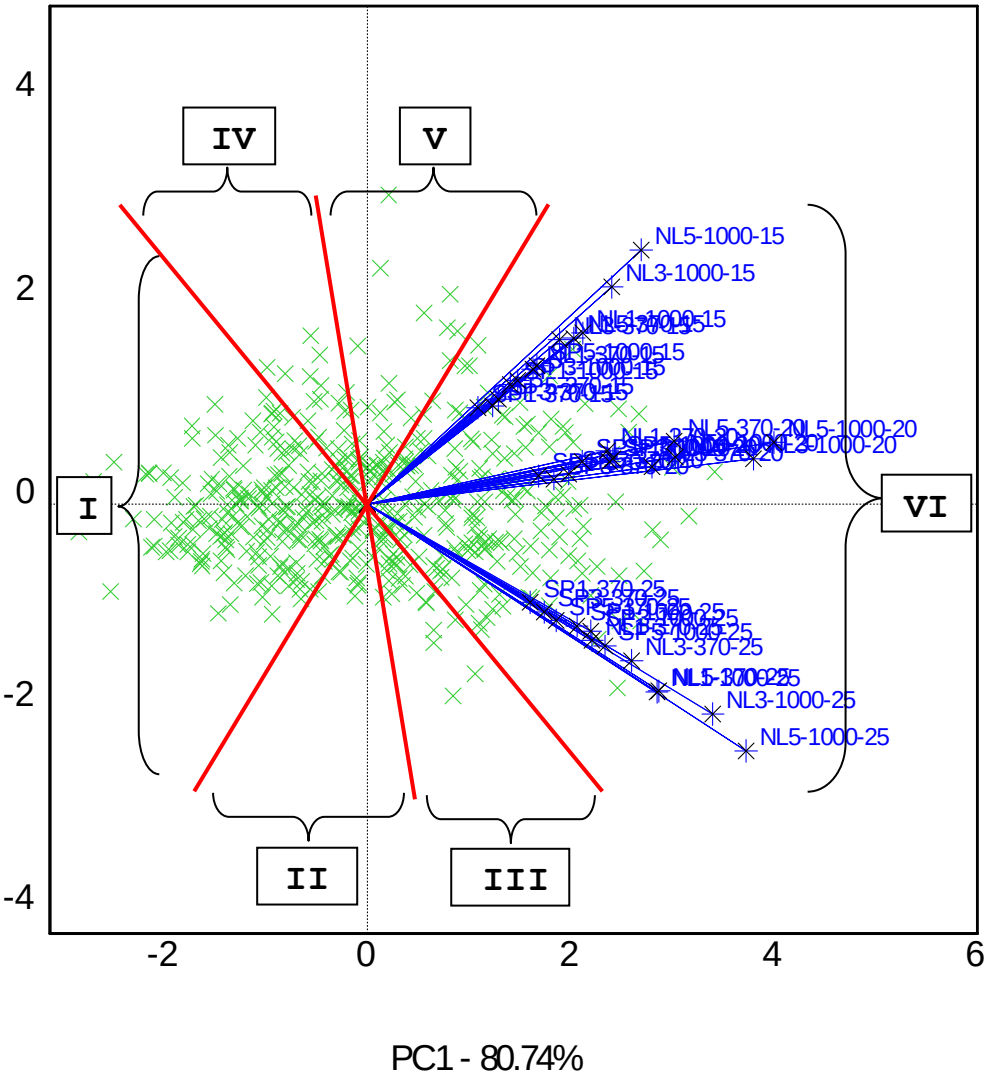

Figure 5.3. GGE biplot for one random realization of the two-way table with 500 genotypes and 36 environments. The abscissa shows the PC1 scores and the ordinate shows the PC2 scores. The 36 environments are marked by their code names (e.g. NL1-370-15 represents a Dutch environment with the minimum yearly average radiation in the considered historical period (NL1), CO2=370 µmol mol^{-1} and daily average temperature of 15ºC). The first and second axes explain a total of 87.72%.

Table 5.5. ANOVA for the AMMI model with 2 interaction principal components. In the column for the sum of squares (SS) the mean values of 10 independent runs of our model are reported, and between brackets the range (minimum and maximum). For the mean squares (MS) only the values associated with the mean SS are reported. The grand mean is 20.014 kg m^{-2}.

Source	df	SS		MS
Total	17999	1274940.8	(1133523; 1331845)	70.83
Genotypes	499	400543.5	(377161; 427486)	802.69
Environments	35	727589.8	(719865; 736996)	20788.28
GEI	17465	163075.7	(159201; 170435)	9.34
IPC1	533	52475.4	(49531; 57342)	98.45
IPC2	531	37925.4	(35552; 41196)	71.42
Residual	15872	72675.0	(71704; 74125)	4.58

Figure 5.4 gives a typical AMMI2 plot of the environmental scores. The first two IPC for this run explained 57.1% of the GEI SS and 16.1% of the GGE SS. The environments appear as three diagonal bands in the plot, with from left to right diagonals for 25, 20 and 15 degrees Celsius. Figure 5.4 thus endorses the results from the factorial regression analysis and the GGE biplot analysis (Table 5.4 and Figure 5.3), both emphasizing the dominant role of temperature. Besides temperature, also country plays an important role, as shown by the factorial regression as well, with Spanish environments being located in the upper right corner of the plot and Dutch environments in the left and lower parts. CO_2 pushes environments to the lower left of the plot within the diagonal groups defined by the temperatures, less than what the factor country does, but more than what radiation does.

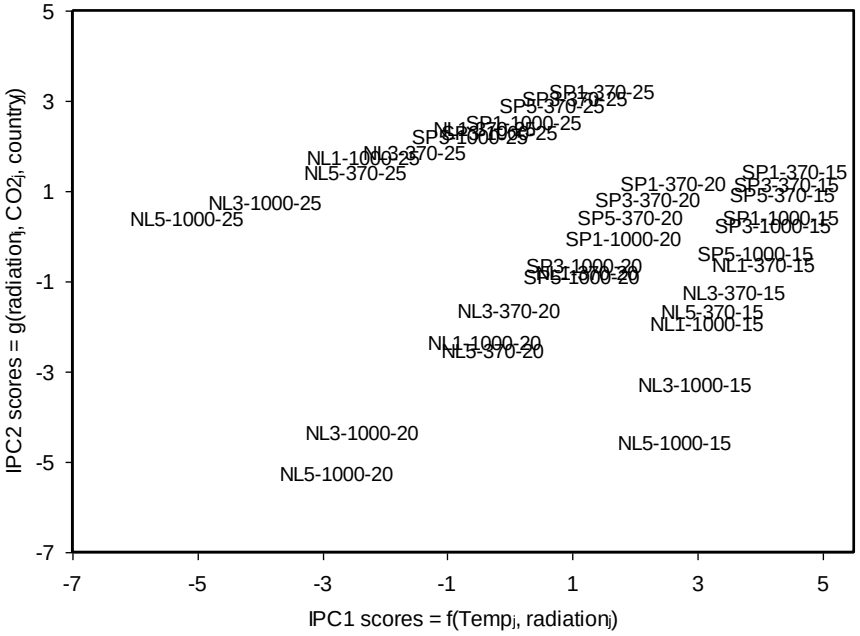

Figure 5.4. AMMI2 biplot for one random realization of the two-way table with 500 genotypes and 36 environments. The abscissa shows the IPC1 scores and the ordinate shows the IPC2 scores. The 36 environments are marked by their code names (e.g. NL1-370-15 represents a Dutch environment with the minimum yearly average radiation in the considered historical period (NL1), CO2=370 μmol mol^{-1} and daily average temperature of 15ºC). The first and second IPC explain 32.5% and 24.6% of the GEI, respectively, for a total of 57.1%.

5.3.3. QTL analyses

QTL analyses identify regions of the genome that contribute to variation in a quantitative trait (e.g. yield). We have chosen one run (one seed) of the model in (5.5) for illustration of a QTL analysis, other runs produced very comparable results.

A preliminary analysis of the VCOV structure was carried out in order to model the genetic variances and correlations across environments. Both, Akaike information criterion (AIC) and Schwarz information criterion (SIC, also known as Bayesian information criterion) pointed to the Factor Analytic with two multiplicative terms (FA2) as the best model for the genetic variances and correlations, following the procedure described by Malosetti et al. (2004) and Boer et al. (2007).

The genetic architecture used in this study comprised 11 QTL (Table 5.3 and Figure 5.2).

The results from the QTL analysis, using composite interval mapping (Zeng, 1994), and the Factor Analytic with two multiplicative terms as VCOV structure, are presented in Table 5.6. As shown in Table 5.6, Table S5.1 and Figure S5.1 (supplementary material, detailed effects across all the 36 environments), 10 out of the 11 QTL were found in many of the environments, often showing QEI, variations across environments for the QTL effect size.

We can observe QEI related to country for the QTL associated to LUE^{max} (Chr. 2, 7, 8 and 10), to FTF (Chr.4), and to $FDMC$ (Chr. 5), when changing from Spanish to Dutch locations. This finding is in agreement with the factorial regressions in Table 5.4 and the AMMI biplot (Figure 5.4), where the IPC2 is a function of the country.

For W (Chr. 6 and 9) and Z (Chr.11) we also clearly observed QEI related to daily average temperature. The QTL in chromosome 3 (B) had a consistent effect across environments. No QTL was found for K (chromosome 1).

Table 5.6 shows the QTL substitution effects for a few environmental contrasts. In the bottom section we show the 36 environments assembled into 6 groups based on combinations of Country and Temperature. We see larger differences between Spain and The Netherlands at 20ºC than at other temperatures (marked in bold). The differences for Country by Temperature combinations are larger than the average differences between the two countries (first two rows of Table 5.6).

When considering the subsets of environments categorized by temperature in Table 5.6, it is clear that the QTL effects of LUE, FTF, $FDMC$ present a curvilinear trend across temperature levels, while the remaining have a linear trend. These trends in the QTL effects can be tested using the standard errors presented in Table 5.6. The significant terms with more than 90% of confidence are marked with bold in Table 5.6, taking into account the linear/curvilinear trend of the parameters within each subset of environments.

We can conclude that the patterns that we observed in the QEI for the various QTLs can be understood in terms of the nature of the underlying yield component and the environmental factors determining yield in particular environments. The QTL analyses in Table 5.6 are thus in good agreement with the factorial regressions in Table 5.4 and the GGE and AMMI biplots in Figures 5.3 and 5.4.

5.4. Discussion

5.4.1. The importance of studying and understanding the GEI and QEI in simulation studies

Genotype-to-phenotype crop growth models have been widely used to study and understand the behaviour of plant development along the growing season. These studies focus mostly on the analysis of GEI and quantitative trait loci (QTL) and, sometimes, on the analysis of QEI. For example, Reymond et al. (2003) combined QTL analysis and a physiological model to analyse the influence of temperature and water deficit on leaf growth in maize. Chenu et al. (2009) simulated the impact of QTL controlling leaf

and silk elongation for maize under drought. Ishii et al. (2010) also presented a simple simulation study to analyse chemical concentration in seed grains.

However, all these studies leave a gap between extensive statistical analysis of GEI and QEI, and understanding the relation between the physiological parameters and the final phenotypic outcome, which is bridged by this work. To the best of our knowledge the most similar approach was presented by Letort et al. (2008), but these authors considered only one environment.

Simulation studies are powerful tools for complementing real breeding programs. Their use opens the possibility of controlling all the input parameters to better "model" the reality. There is also a much lower cost to re-run a simulation study than to re-do the field experiment. The current simulations show how to obtain insight in the factors determining a complex trait like yield by using additional genotypic and environmental information in the analysis of multiple environment data for the complex trait. QEI analysis for yield using additional information on yield components and environmental characterizations allows the partial unravelling of the genotype-to-phenotype function and the genetic architecture involved.

Table 5.6. QTL effects and (standard errors) for the 10 detections (Chr.2 to Chr.11) for several subsets of environments. The influence of the QTL increases with the absolute value of the coefficients. The signal of the coefficients represents the parent responsible for the QTL. The considered VCOV structured was the Factor Analytic with two multiplicative terms. The last column has the mean standard error (s.e.) for each environmental group. The bold values represent significant overall differences between the levels of each subset of environments, with a confidence level of at least 90%.

Parameter	LUE	B	FTF	FDMC	W	LUE	LUE	W	LUE	Z	s.e.
Chromosome	2	3	4	5	6	7	8	9	10	11	
SP	**4.46**	0.87	**3.85**	**-3.70**	-1.72	**3.49**	**3.91**	-1.58	**4.29**	1.68	0.86
NL	**6.84**	1.05	**6.41**	**-5.82**	-2.10	**5.70**	**5.90**	-1.78	**6.68**	2.68	0.71
CO2 = 370	5.10	0.80	4.65	-4.14	-1.40	4.12	4.44	-1.17	5.07	1.97	0.85
CO2 = 1000	6.20	1.12	5.62	-5.39	-2.42	5.07	5.36	-2.19	5.89	2.38	0.73
T = 15ºC	**4.56**	1.45	**4.62**	**-3.88**	**0.25**	4.21	**3.91**	**1.04**	**4.66**	**3.47**	0.62
T = 20ºC	**6.77**	1.11	**6.29**	**-5.86**	**-1.06**	5.67	**6.05**	**-0.86**	**6.69**	1.97	0.84
T = 25ºC	**5.62**	0.32	**4.49**	**-4.55**	**-4.92**	3.90	**4.75**	**-5.22**	**5.10**	**1.09**	0.89
SP, Temp = 15ºC	3.58	1.23	3.46	-3.01	0.20	3.18	3.10	**0.83**	3.74	2.78	0.68
SP, Temp = 20ºC	5.28	0.98	4.58	-4.39	**-1.37**	4.08	4.69	**-1.33**	4.98	1.44	0.91
SP, Temp = 25ºC	4.52	0.39	3.52	-3.71	**-3.98**	3.20	3.94	**-4.24**	4.15	0.81	0.97
NL, Temp = 15ºC	**5.55**	1.66	**5.77**	**-4.75**	0.31	**5.23**	**4.71**	**1.25**	**5.58**	4.16	0.55
NL, Temp = 20ºC	**8.26**	1.24	**8.00**	**-7.32**	**-0.75**	**7.26**	**7.42**	**-0.39**	**8.39**	2.51	0.76
NL, Temp = 25ºC	**6.71**	0.24	**5.47**	**-5.39**	**-5.87**	**4.61**	**5.56**	**-6.21**	**6.06**	1.37	0.80

5.4.2. How complex should a crop growth model be to generate GEI and QEI?

The integration of statistical-genetics and crop growth modelling for reliable and robust prediction of phenotypic traits, on the basis of genotypic-specific and stable physiological parameters and environmental characterizations, is the object of extensive research in plant sciences (Tardieu, 2003, Chenu et al., 2008, Malosetti et al., 2010). Very often these models are so complex and have so many parameters that it is almost impossible to apply them from a practical point of view because of the difficulty to obtain realistic estimates for the parameters.

In this study we intended to consider a parsimonious model with a small number of parameters. As a starting point for the present physiological model, five parameters were considered, in which the partitioning to the fruits (harvest index) was considered to be constant. This model appeared to be too simple and, although some interactions could be found between the genotypes and environments, no crossovers were detected.

With the inclusion of two parameters to force a linear reduction in harvest index for temperatures above 15ºC (Wubs et al., 2009, Wubs et al., 2010) and a genotypic-specific exponential reduction in LUE for temperatures below 25ºC which starts to have more impact below 20ºC (de Swart et al., 2006), the model became more realistic while still being simple. Despite its simplicity, the model simulated of GEI and QEI, including crossovers (Figure 5.3), for yield.

Still, it needs to be admitted that GEI and QEI with crossovers could only be obtained by "penalization" of yield components in relation to environmental factors (e.g. lower partitioning to the fruits for higher temperatures; higher fruit abortion at higher temperatures; drought influence; pests; etc.). So, the model was able to simulate significant GEI and QEI, including crossovers, while the physiological parameters were environment independent, but the estimates for the physiological parameters may still require prior experiments including some relevant environmental contrasts (Tardieu, 2003).

5.4.3. The genetic architectures and the transmission of information from the original physiological parameters to the final phenotypic two-way table

One of the main achievements of this simulation study was the detection of QTL for yield, while simulating only QTL for physiological parameters. After defining a set of genotype specific model parameters, which can serve as features for a QTL-analysis, independent QTL (and independent of the environments) were assigned to each of the 7 model parameters. The QTL analysis of the two-way table with yields revealed QTL directly linked to 5 out of the original 7 parameters in all the simulation runs of the model for different simulation seeds (only the QTL for *K* and *B* were not consistent because of their lower importance). This is visible in Tables 5.6 and S5.1 and in the lower panel of Figure S5.1. It is also possible to observe the lack of importance of CO_2 levels in the final yield.

Several genetic architectures, concerning the number of QTL per physiological parameter and their respective effects, were considered. The number of QTL was chosen from 7 to 11 while the percentage of variance explained (i.e. the heritability) by them varied from 4% to 95%, to cover a wide range of combinations.

The number and importance of QTL for each parameter was chosen based on the relative importance of the parameter for the model (Table 5.4): (i) LUE^{max} (4 QTL with heritability of 12%) and W (2 QTL with heritability of 16%) were chosen to depend on more than one QTL; (ii) *K* and *B* were considered to be responsible for an high proportion of variance (i.e. 95% of the variance in the parameter was due to the QTL); and (iii) the remaining parameters were of average importance and the proportions of variance assigned to the respective QTL were chosen accordingly (64% of the variance in the parameter due to the QTL).

In a more extreme scenario, where only 4 to 6% of the heritability in LUE^{max} was explained by each QTL and/or the number of genotypes was reduced, the QTL in LUE^{max} were still detected. Similarly, if the heritability of the remaining parameters was reduced to half (of the one considered in this study), only the QTL for W was detected consistently in all the environments with 25ºC. This evidences the importance of LUE^{max} and W in the outcome of this crop growth model.

The most important QTL are those that are relatively stable across environments, i.e. QTL that present QTL main effects without environment-specific deviations. The genetic information in LUE^{max} is clearly the most important in this sense, being followed by $FDMC$ and FTF (Tables 5.6 and S5.1 and lower panel of Figure S5.1). The QTL assigned to W have higher $-\log_{10}$(P-value) values for the QTL effects at higher temperatures, while the QTL for Z have lower $-\log_{10}$(P-value) values for the QTL effects at higher temperatures (Figure S5.1 and Table 5.6), which is in accordance with the factorial regression (Table 5.4). It is expected to find higher yields for daily average temperatures around 20ºC.

The QTL associated to the most important parameters, for the trait in study, can be detected in the exact same place where they were allocated during the simulation. These detections were made using only the final phenotypic data and the genetic map with the marker information. This result underlines the penetrance of genetic information on component traits in a physiological model, to the final yields across a wide set of environmental conditions.

5.5. Supplementary material

Table S5.1. QTL effect for the 10 detections (Chr.2 to Chr.11) for each of the 36 environments. The influence of the QTL increases with the absolute value of the coefficients. The signal of the coefficients represents the parent responsible for the QTL. The considered VCOV structured was the Factor Analytic with two multiplicative terms. The last column has the mean standard error (s.e.) for each environment.

Parameter	LUE	B	FTF	FDMC	W	LUE	LUE	W	LUE	Z	s.e.
Chromosome	2	3	4	5	6	7	8	9	10	11	
NL1-1000-15	5.01	1.76	5.52	-4.39	0.59	5.52	5.08	0.97	5.48	4.34	0.71
NL1-1000-20	6.56	1.58	6.55	-6.29	-2.25	6.96	6.81	-2.04	7.37	1.85	0.96
NL1-1000-25	6.60	1.33	4.80	-4.85	-6.05	3.71	5.47	-5.77	5.51	1.19	0.97
NL1-370-15	4.83	0.54	4.77	-3.65	0.15	3.61	3.41	1.05	4.37	2.99	0.54
NL1-370-20	7.25	0.25	7.58	-5.99	0.87	6.67	5.93	1.02	7.03	2.11	0.91
NL1-370-25	4.86	-0.04	4.06	-4.41	-5.02	3.42	4.41	-5.00	4.46	0.93	0.72
NL3-1000-15	6.88	1.54	6.11	-6.09	0.62	5.57	4.76	1.35	5.87	4.74	0.82
NL3-1000-20	9.80	1.36	7.96	-7.83	-2.95	8.00	8.22	-1.62	8.93	2.28	1.17
NL3-1000-25	8.46	-0.25	6.39	-6.31	-6.26	5.66	5.65	-7.33	6.23	2.27	1.11
NL3-370-15	4.90	1.72	5.63	-3.97	0.35	4.54	4.88	1.14	5.27	3.87	0.62
NL3-370-20	7.51	1.49	7.91	-7.67	1.28	7.29	7.05	0.36	8.87	2.53	1.05
NL3-370-25	6.08	0.62	4.91	-4.40	-4.49	4.36	5.28	-5.33	5.88	1.93	0.85
NL5-1000-15	6.14	2.01	7.35	-6.61	0.52	6.86	5.85	1.89	7.24	5.03	0.92
NL5-1000-20	9.39	1.86	9.60	-9.04	-2.98	7.25	8.37	-2.82	8.93	2.29	1.25
NL5-1000-25	7.76	0.11	7.73	-7.14	-7.38	6.30	7.44	-8.15	8.57	1.15	1.19
NL5-370-15	5.55	2.41	5.24	-3.77	-0.40	5.29	4.30	1.13	5.24	3.98	0.68
NL5-370-20	9.06	0.90	8.42	-7.12	1.53	7.36	8.15	2.79	9.23	3.99	1.17
NL5-370-25	6.49	-0.30	4.91	-5.26	-6.00	4.19	5.09	-5.69	5.70	0.75	0.97
SP1-1000-15	3.79	1.32	3.55	-3.26	0.27	3.31	3.11	0.74	3.76	2.89	0.48
SP1-1000-20	5.41	0.73	4.94	-5.04	-1.52	3.88	5.02	-0.73	4.36	1.73	0.67
SP1-1000-25	4.60	0.88	3.73	-4.24	-4.07	3.42	4.33	-4.54	3.66	0.87	0.67
SP1-370-15	2.99	0.84	2.54	-2.42	0.22	2.77	2.68	0.99	3.05	2.13	0.37
SP1-370-20	4.04	0.40	3.85	-3.56	-1.23	3.78	3.73	-1.33	4.15	1.37	0.53
SP1-370-25	3.69	0.34	2.86	-2.99	-2.95	2.69	3.55	-3.42	3.42	0.73	0.52
SP3-1000-15	4.09	1.20	3.69	-3.28	0.16	3.22	3.06	1.27	4.20	3.08	0.50
SP3-1000-20	5.57	1.02	5.36	-4.95	-1.48	4.80	5.98	-1.62	5.90	1.87	0.74
SP3-1000-25	5.22	0.24	4.05	-4.23	-4.49	4.00	3.99	-5.09	4.40	1.39	0.71
SP3-370-15	2.95	1.13	3.09	-2.62	0.26	3.15	2.90	0.66	3.24	2.17	0.42
SP3-370-20	4.67	1.15	4.06	-3.25	-0.76	3.41	4.11	-1.34	4.79	0.63	0.60
SP3-370-25	3.91	0.56	3.29	-3.32	-4.03	2.37	3.21	-3.70	3.91	0.40	0.56
SP5-1000-15	4.55	1.86	4.57	-3.75	-0.12	3.70	3.70	0.63	4.44	3.80	0.55
SP5-1000-20	6.26	1.31	5.16	-5.80	-1.54	4.95	5.03	-1.81	6.16	1.31	0.76
SP5-1000-25	5.47	0.28	4.01	-3.90	-4.59	4.05	4.69	-4.84	5.08	0.80	0.76
SP5-370-15	3.08	1.02	3.32	-2.73	0.41	2.92	3.15	0.67	3.78	2.62	0.44
SP5-370-20	5.72	1.27	4.11	-3.75	-1.69	3.68	4.25	-1.15	4.54	1.74	0.62
SP5-370-25	4.25	0.06	3.19	-3.60	-3.75	2.64	3.90	-3.83	4.41	0.66	0.60

Figure S5.1. Genome scan for the yield data. The top panel presents the scan with the $-\log_{10}$(P-value) values for the QTL effects, including main effects and environment-specific effects. The red horizontal line is the 5% genomewide significance threshold. The bottom panel depicts the environment specific QTL effects with the environment labels on the left hand side. The green in the first row summarizes the top panel. The blue colour represents the increasing effect of one parent in yield and the red the decreasing of that same parent. Darker colours mean stronger effect while light colours weaker effects. The considered VCOV structured was the Factor Analytic with two multiplicative terms.

Chapter 6

6. Weighted AMMI to study genotype-by-environment interaction and QTL-by-environment interaction

Abstract

Genotype-by-environment interactions (GEI) and quantitative trait locus (QTL) -by-environment interactions (QEI) are common phenomena in multiple-environment trials and represent a major challenge for breeders that want to select better adapted genotypes. The additive main effects and multiplicative interaction (AMMI) model is a widely used tool in the analysis of multiple-environment trials, but in its standard form it doesn't take in to account the heterogeneity of error variance across environments that is typical for many multiple-environment data with strong GEI.

In this chapter we introduce a generalization of AMMI model that accounts for heterogeneity of error variance across environments, the weighted AMMI, or WAMMI. WAMMI is useful for studying GEI as well as QEI. For QEI, we perform an initial analysis by WAMMI, and take the predicted values from this analysis as starting point for QTL analyses per environment. We look at the performance of this strategy in relation to QTL scans on the actual data and AMMI predicted values. We also look at a full mixed model approach to QTL mapping for multiple-environments. We used two data sets for making comparisons: (i) data from a simulated pepper (*Capsicum annuum*) back cross population using a crop growth model to relate genotypes to phenotypes; and (ii) a doubled-haploid barley (*Hordeum vulgare* L.) population. Our results demonstrate that the QTL scans of the WAMMI predicted values outperform the QTL scans for the actual data and for the AMMI predicted values, being very similar to the QTL mixed model approach, with respect to the number of QTLs detected. WAMMI for GEI and QEI has wide applicability.

To be submitted as: Rodrigues, P.C., Malosetti, M., Gauch, J.G. and van Eeuwijk, F.A. Weighted AMMI to study genotype-by-environment interaction and QTL-by-environment interaction.

6.1. Introduction

A differential response of genotypes across environments (often, location by year combinations) is frequent in multi-environment trials (METs) and is known as genotype-by-environment interaction (GEI). Data from METs are often summarized in two-way tables with genotypes in the rows and environments in the columns. GEI occurs in various forms, with the most extreme consisting of crossovers, when there is a change of ranking of genotypes across environments, e.g., a genotype that is superior under well watered conditions may yield poorly under dry conditions. The study and understanding of GEI is a major challenge in the improvement of complex traits like yield across environmental gradients.

The additive main effects and multiplicative interaction (AMMI) model (Gauch, 1992) is one of the most widely used statistical methods to understand and structure interactions between genotypes and environments. In essence the AMMI model applies the singular value decomposition (SVD) to the residuals of the analysis of variance (ANOVA). However, if there is a strong GEI in the data, we also expect the trials to have heterogeneous error variances, and this is not taken into account by the standard AMMI model. Therefore, we propose a generalization of the AMMI model that is able to take into account heterogeneity of error variance by using a weighted low-rank SVD, the weighted AMMI (WAMMI) model. Although this generalization occurs in a fixed model, WAMMI offers a reasonable approximation to mixed model methodology for GEI, which is considered to be more appropriate in case of heterogeneous error variances.

A natural follow up to the analysis of GEI, is the study of the genetic factors underlying GEI: QTL (quantitative trait locus) and environment interaction, QEI. Gauch et al. (2011) proposed the AQ analysis where the AMMI model is used to obtain predicted values for genotype-by-environment combinations, which are then used in QTL mapping. We believe that the weighting by the (reciprocal) of error variances in the AMMI model will not only improve the analysis of GEI, but equally so that of subsequent QTL analysis. Thus, we present a generalization of the AQ analysis that is able to account for heterogeneity in both genetic variances, captured by the interaction principal components in AMMI, and error variances, by weighting; we replace the AMMI model by the WAMMI model. The weighted version of the AQ analysis, WAQ, can be conducted in three stages: (i) compute the weights for each environment based on the error variances; (ii) fit the WAMMI model to the GEI data table; and (iii) perform the QTL scans using the predictions from the WAMMI model as response variable. In the spirit of AMMI, with our WAMMI approach we expect to separate signal, GEI and QEI patterns, from noise.

WAQ is compared with the QTL analyses on the actual data, with the AQ analysis (Gauch et al., 2011), and with a QTL mixed models approach (Boer et al., 2007, Malosetti et al., 2004). Two data sets were used. The first one deals with yield simulated from a backcross pepper (*Capsicum annuum*) population using a crop growth (physiological) genotype-to-phenotype model (Rodrigues et al., 2012a). The motivation for using a crop growth model to transform genotypic information to phenotypic information was that we wanted a biologically realistic data set, while still wanting to know the underlying genetic architecture. The second data set concerned 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, Hayes et al., 1996).

WAMMI is applicable to a wide range of fields to which also AMMI has been applied, including more than 200 articles referenced by the ISI web of knowledge within the last ten years. In addition applications in plant breeding, crop sciences and genetics, AMMI was applied to microarray experiments (Crossa et al., 2005), rDNA studies (Adams et al., 2002), plant and microbial populations' growth across several environmental conditions (Culman et al., 2008, Culman et al., 2009), and animal sciences (Barhdadi and Dube, 2010).

6.2. Materials and methods

6.2.1. Plant materials

The primary data set of this study is a two-way table with $I = 200$ genotypes and $J = 12$ environments (Table 6.1) of the complex trait yield. These data were simulated by assuming that the final yield set equals the signal plus the noise. The signal for genotype i in environment j was simulated from a eco-physiological genotype-to-phenotype crop growth model (CGM) for pepper (Rodrigues et al., 2012a) and is a function of physiological parameters and environmental characterizations. The model can be written as:

$$Yield_{i,j} = Signal + Noise$$
$$= \frac{FTF_i \times [1 - W_j \times (T_j - T_{FTF})]}{FDMC_i} \times LUE_{i,j} \times \sum_{t=t_0}^{t_f} \left[1 - exp\left(-K_i \times LAI_{i,j,t}\right)\right] \times I_{j,t} + \varepsilon_{i,j} \quad (6.1)$$

where t_0 and t_f represent the beginning and the end of the growing season, in days, $LAI_{i,j,t}$ the leaf area index for genotype i, environment j and day t, and $I_{j,t}$ is the photosynthetic active radiation incident on the crop for environment j on day t, and $\varepsilon_{i,j}$ is the error (or noise) for the $Yield_{i,j}$.

The model (6.1) is a function of seven physiological parameters: K (light extinction coefficient); LUE (maximum light use efficiency); B (slope for the leaf area increase with temperature sum, used to define LAI); FTF (fraction of dry weight partitioned to the fruits); $FDMC$ (fruit dry matter content); W (slope of the linear reduction in harvest index with temperature above 15ºC); and Z (slope of the linear reduction in LUE for temperatures below 20ºC, used to define LUE); and three environmental variables: temperature, radiation and country (Table 6.1). More details can be found in Rodrigues et al. (2012a).

The main motivation for using a nonlinear physiological genotype-to-phenotype model instead of a statistical model is to ensure that the simulated data is close to a biologically credible model, where we have full information on the biological background.

Each of the seven physiological parameters (component traits) was simulated as a sum of a number of QTLs (Table 6.2) plus a residual effect. This was done for 200 simulated pepper genotypes, characterized by 237 markers covering all the 12 chromosomes (Barchi et al., 2007, Barchi et al., 2009). In this simulation several QTLs were placed along the 12 chromosomes of the pepper genome. The exact positions of these QTLs and their heritability are described in Table 6.2. The simulations were made using

the package qtl (Broman and Sen, 2009) of the statistical software R. More details on the model and physiological parameters can be found in Rodrigues et al. (2012a).

The noise for yield $\varepsilon_{i,j}$ in equation (6.1) was simulated from a Gaussian distribution with zero mean and variance $\sigma^2_{\varepsilon_j}, j = 1, \ldots, J$, depending on the environment (Table 6.1) and the chosen heritability for yield (h^2, Table 6.2), i.e.

$$\sigma^2_{\varepsilon_j} = \frac{1 - h^2}{h^2}\sigma^2_{g_j},$$

where $\sigma^2_{\varepsilon_j}$ and $\sigma^2_{g_j}$ are the error and genetic variance (from the eco-physiological genotype-to-phenotype model) for the environment j, $j = 1, \ldots, J$ (Table 6.1). The final yield data is the result of the sum of the signal and the noise as in equation (6.1). This simulation was repeated 100 times resulting in 100 two-way tables with 200 genotypes and 12 environments.

Table 6.1. The 12 environments used in the simulated yield data for pepper. Description of the environments considered in the genotype-to-phenotype crop growth model. The first column represents the code for the environments which is used in the text and figures. The countries were chosen to represent different environmental and practical conditions (Rodrigues et al., 2012a). Radiation has two levels (years) based on historical data. Temperature contains three levels of daily average temperature. The heritability for the environments was set to be $h^2 = 0.5$. The mean genetic and error variances, for the 100 simulated data sets, are reported in the last two columns.

Environment	Country	Radiation	Temperature	Genetic variance	Error variance
NL1-15	Netherlands	Lower	15ºC	21.27	21.32
NL1-20	Netherlands	Lower	20ºC	39.29	39.60
NL1-25	Netherlands	Lower	25ºC	39.22	39.19
NL5-15	Netherlands	Higher	15ºC	35.08	34.82
NL5-20	Netherlands	Higher	20ºC	65.14	64.81
NL5-25	Netherlands	Higher	25ºC	65.25	64.77
SP1-15	Spain	Lower	15ºC	9.91	9.79
SP1-20	Spain	Lower	20ºC	19.06	19.27
SP1-25	Spain	Lower	25ºC	19.66	19.96
SP5-15	Spain	Higher	15ºC	13.16	13.24
SP5-20	Spain	Higher	20ºC	25.43	25.75
SP5-25	Spain	Higher	25ºC	26.29	26.46

Table 6.2. Genetic architecture of the simulated yield data for pepper (signal). The first columns give the name of the parameter related to the QTL, the code for the closest marker, the chromosome, the position and its heritability when included in the physiological genotype-to-phenotype model. The last column gives a summary on which environments the QTLs are expected to be detected, being the FTF and $FDMC$ much weaker and harder to detect than the LUE.

Parameter	Marker	Chromosome	Position (cM)	Heritability	Importance[1]
K	D1M6	1	38.0	0.95	None
LUE	D2M13	2	55.0	0.16	All
B	D3M10	3	87.3	0.95	None
FTF	D4M9	4	83.1	0.80	All
$FDMC$	D5M25	5	103.3	0.80	All
W	D6M12	6	36.7	0.20	20ºC; 25ºC
LUE	D7M5	7	42.5	0.16	All
LUE	D8M7	8	38.8	0.16	All
W	D9M15	9	100.4	0.20	20ºC; 25ºC
LUE	D10M5	10	43.1	0.16	All
Z	D11M11	11	62.4	0.80	15ºC

[1]Based on a sensitivity analysis with heritability of 1 for all environments (Table 6.1).

The second data set in our study is a subset of the grain yield data from the Steptoe x Morex (SxM) cross, produced by the North American barley genome mapping project (Hayes et al., 1993, Hayes et al., 1996). The data contain 150 doubled haploid genotypes evaluated in 16 environments during 1991 and 1992, in USA and Canada. The genotypes were characterized by 116 markers covering all seven chromosomes. For inclusion in our data set, environments needed to have either a complete replication (block) or a complete replication block and an additional partial replication. The 13 chosen environments are presented in Table 6.3. The trials conducted in 1991 had a full replicate/block and a second one containing only 50 genotypes. For trials in 1992 two complete replications were available.

Table 6.3. The 13 environments used in the SxM analysis. The first column gives the code for the environment (location year combination) which is used in the text and figures. The information about full replication of partially replication, genetic and error variances, and heritability are presented in the next columns.

Environment	Full replication	Genetic variance	Error variance	Heritability
ID91	No	0.94	0.74	0.56
ID92	Yes	0.55	0.42	0.57
MAN92	Yes	0.38	0.20	0.66
MIN92	Yes	0.35	0.37	0.49
MTd91	No	0.39	0.11	0.78
MTd92	Yes	0.43	0.31	0.58
MTi91	No	0.36	0.23	0.61
MTi92	Yes	0.43	0.31	0.58
NY92	Yes	0.20	0.67	0.23
ONT92	Yes	0.21	0.31	0.40
OR91	No	0.26	1.66	0.14
WA91	No	0.72	0.71	0.50
WA92	Yes	0.20	0.25	0.44

6.2.2. AMMI analysis

The additive main effects and multiplicative interaction (AMMI) model (Gauch, 1988, Gauch, 1992) combines together the features of analysis of variance (ANOVA) and singular value decomposition (SVD), the SVD is applied to the residuals from the additive ANOVA, i.e. to the GEI. In the ANOVA part, the additive main effects are estimated, whereas the SVD models the interaction via N axes, or N interaction principal components, IPCs, $N \leq \min(I-1, J-1)$, with I the number of genotypes (rows) and J the number of environments (columns). The model is usually written as (Gauch, 1992)

$$y_{i,j} = \mu + \alpha_i + \beta_j + \sum_{n=1}^{N} \lambda_n \gamma_{n,i} \delta_{n,j} + \varepsilon_{i,j}, \quad (6.2)$$

where $y_{i,j}$ is the yield of genotype i in environment j, μ the grand mean, α_i the genotype deviations from μ, β_j the environment deviations from μ, λ_n is the singular value for the IPC axis n, $\gamma_{n,i}$ and $\delta_{n,j}$ the genotype and environment IPC scores (i.e. the left and right singular vectors) for axis n, and $\varepsilon_{i,j}$ the residual containing both multiplicative terms not included in the model (6.2) as well as an experimental error. A matrix formulation of equation (6.2) can be given by

$$\boldsymbol{y} = \mathbf{1}_I \, \mathbf{1}_J^T \mu + \boldsymbol{\alpha} \, \mathbf{1}_I^T + \mathbf{1}_J^T \, \boldsymbol{\beta} + \boldsymbol{U D V}^T + \boldsymbol{\varepsilon}, \quad (6.3)$$

where $\boldsymbol{y}$ is the $(I \times J)$ two-way data table, $\mathbf{1}_I \, \mathbf{1}_J^T \mu$ is a $(I \times J)$ matrix with the grand mean μ in all positions, $\boldsymbol{\alpha} \, \mathbf{1}_I^T$ is a $(I \times J)$ matrix with the genotype deviations from the grand mean (equal rows), $\mathbf{1}_J^T \, \boldsymbol{\beta}$ is a $(I \times J)$ matrix with the environment deviations from the grand mean (equal columns), $\boldsymbol{U}$ is a $(I \times N)$ matrix whose columns contain the left singular vectors of the multiplicative part of the data, i.e. $\tilde{\boldsymbol{y}} = \boldsymbol{Y} - \boldsymbol{\mu} - \boldsymbol{\alpha} - \boldsymbol{\beta}$, $\boldsymbol{D}$ a $(N \times N)$ diagonal matrix containing the singular values of $\tilde{\boldsymbol{y}}$ in the diagonal, $\boldsymbol{V}$ is a $(J \times N)$ matrix whose columns contain the right singular vectors of $\tilde{\boldsymbol{y}}$, and is the $(I \times J)$ matrix with the residuals. With this procedure we are aiming at a low rank (N) approximation to the matrix $\tilde{\boldsymbol{y}}$, i.e. the interaction.

The number of interaction terms in the model, N, has to be chosen wisely as it will affect all the subsequent results (Gauch et al., 2008, Yang et al., 2009). In this chapter we use a cross-validation based method proposed by Krzanowski (1987). By considering the model for the GEI data table $x_{i,j} = \sum_{n=1}^{N} \lambda_n \gamma_{n,i} \delta_{n,j} + \varepsilon_{i,j}$, we are able to compute the average squared discrepancy between the actual and predicted values:

$$PRESS(n) = \frac{1}{IJ} \sum_{i=1}^{I} \sum_{j=1}^{J} \left(\hat{x}_{i,j}^{(n)} - x_{i,j} \right)^2, \tag{6.4}$$

and, consequently,

$$W_n = \frac{PRESS(n-1) - PRESS(n)}{I+J-2n} \div \frac{PRESS(n)}{D_r}, \tag{6.5}$$

where D_r can be obtained by sequential subtraction from $(I-1)(j-1)$ of $I + J - 2N$. The W_n represent the increase in predictive information supplied by the n-th component, divided by the average predictive information in each of the remaining components (Krzanowski, 1987). Krzanowski (1987) suggested that the optimal number of components is the highest number of n such that W_n is greater than 0.6.

6.2.3. Weighted AMMI analysis

When the two-way data table $\boldsymbol{y}$ has missing cells and/or the error variance is not constant across environments, the cells of the table should have different weights for their squared residuals in the estimation procedure for the model parameters. To account for heterogeneity of error variances across environments, our proposal is to replace the standard low-rank SVD in equation (6.3) by a weighted low-rank SVD (Gabriel and Zamir, 1979). The approach we use here is based on an expectation-maximization (EM) procedure and, while the sum of squares of the difference between two consecutive iterations, $\boldsymbol{X}^{(t+1)}$ and $\boldsymbol{X}^{(t)}$, is greater than some small value, e.g. 10^{-9}, we run

$$\boldsymbol{X}^{(t+1)} = SVD\left(\boldsymbol{W} \odot \boldsymbol{y} + (\mathbf{1} - \boldsymbol{W}) \odot \boldsymbol{X}^{(t)} \right) \tag{6.6}$$

where $\boldsymbol{W}$ is a $(I \times J)$ matrix with weights, $W_{i,j}$, $0 \leq W_{i,j} \leq 1$, $\mathbf{1}$ is a $(I \times J)$ matrix with ones in all positions, $\odot$ the Hadamard (or entrywise) product of matrices, and t is the iteration number (Srebro and Jaakkola, 2003). $\boldsymbol{X}$ should be initialized to $\boldsymbol{X}^{(0)} = \boldsymbol{y}$ or to $\boldsymbol{X}^{(0)} = \mathbf{0}$. The outputs of this procedure are the matrices $\boldsymbol{U}_N$, $\boldsymbol{D}_N$ and $\boldsymbol{V}_N$ such that $\tilde{\boldsymbol{y}} \approx \boldsymbol{U}_N \boldsymbol{D}_N \boldsymbol{V}_N{}'$, being N the rank of approximation. The R code for this algorithm, with detailed explanation, can be found in the File S1 of the supplementary material.

Applying the weighted low-rank SVD (6.6) to the matrix $\tilde{\boldsymbol{y}}$ and replacing in equation (6.3) will result in the weighted AMMI (WAMMI) model. This generalization of the AMMI model is now able to account

for differences in error variances across environments and/or missing cells, and can be applied to all data sets where the AMMI model has been used. Of course, we need to be able to estimate the error variance for an environment, so we need at least partial replication per environment. It should be remarked that in this chapter we first estimate the main effects without using weights, then produce the residuals from additivity and finally approach these residuals by a weighted SVD. We could also have used the weights already in the estimation of the main effects, but this approach may be less robust. We feel that this topic merits further study.

With partial replication, cell means based on more replication will have smaller variances than those with less replication. The scheme of weights should reflect the number of replications per cell. The a (I×J) matrix with weights, $W_{i,j},\ 0 \leq W_{i,j} \leq 1$, can be calculated from the Hadamard (or entrywise) product of two matrices: (i) a matrix in which the entries are column wise constant, being the inverse of the error variance; and (ii) a matrix with the proportion of replications per cell, i.e.

$$W = \begin{bmatrix} \frac{1/\sigma_1^2}{m} & \frac{1/\sigma_2^2}{m} & \cdots & \frac{1/\sigma_J^2}{m} \\ \frac{1/\sigma_1^2}{m} & \frac{1/\sigma_2^2}{m} & \cdots & \frac{1/\sigma_J^2}{m} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1/\sigma_1^2}{m} & \frac{1/\sigma_2^2}{m} & \cdots & \frac{1/\sigma_J^2}{m} \end{bmatrix} \odot \begin{bmatrix} \frac{Nrep_{1,1}}{Nrep} & \frac{Nrep_{1,2}}{Nrep} & \cdots & \frac{Nrep_{1,J}}{Nrep} \\ \frac{Nrep_{2,1}}{Nrep} & \frac{Nrep_{2,2}}{Nrep} & \cdots & \frac{Nrep_{2,J}}{Nrep} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{Nrep_{I,1}}{Nrep} & \frac{Nrep_{I,2}}{Nrep} & \cdots & \frac{Nrep_{150,J}}{Nrep} \end{bmatrix}, \quad (6.7)$$

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

6.2.4. Weighted AQ analysis

Gauch et al. (2011) suggested a new approach for detecting and understand QEI, the AQ analysis, where the QTL scans are made based on AMMI predictions (instead of direct QTL scans on the actual data). In this chapter we make use of the above proposed weighted version of the AMMI model, the WAMMI, to generalize the AQ analysis to account for both heterogeneous genetic (SVD) and error variances (weighs) across environments. We use a fixed effects model as an alternative to the QTL mixed models approach (Boer et al., 2007, Malosetti et al., 2004) that can be fitted with standard statistical software for linear models. The weighted AQ analysis can be conducted in three stages: (i) compute the weights for each environment based on the error variances, i.e. the weights are given by the inverse of the error variance in each environment and are (usually) constant for all genotypes in an environment; (ii) fit the WAMMI model to the GEI data table and obtain the predicted values for each combination of genotype and environment; and (iii) perform the QTL scans using the WAMMI predicted values as response variable for each environment separately. This approach can potentially improve the power for QTL detection as it uses improved genotypic predictions as response variable that showed to be better than the means from the ANOVA model. The environments can then be ordered by AMMI and WAMMI parameters that summarize GEI and QEI information to reveal consistent patterns and

systematic trends that often can be explained in terms of environmental conditions (Gauch et al., 2011, Gauch, 1992).

6.2.5. Linear mixed model

As a kind of bench mark for QTL analysis, we analysed the simulated pepper and Steptoe x Morex barley data also by a QTL analysis based on mixed models, as described by Boer et al. (2007), and implemented in Genstat 14 (Payne et al., 2011). The input for the QTL analysis consists of the genotype-by-environment means and corresponding weights, defined in the way also used in WAMMI and WAQ. The QTL analysis fits fixed environment specific QTLs, i.e., actually the sum of QTL main effect and QEI, to the genotypic main effects and GEI, say GGEI, jointly. The model contains a multivariate normal distribution for the GGEI effects allowing heterogeneity of genetic variances and correlations.

6.3. Results for the simulated pepper data

6.3.1. Preliminary analysis

Table 6.2 gives the simulation conditions and, therefore, the "true" genetic architecture for the pepper population under study. Figure 6.1 (blue line) depicts the single trait single environment QTL scans for 6 environments, of the complex trait yield simulated from the physiological genotype-to-phenotype model with seven physiological parameters (Rodrigues et al., 2012a). The six environments were chosen to represent the three levels of temperature with the lowest and highest error variances. Comparing the "true" genetic architecture in Table 6.2 and the QTLs detected in Figure 6.1 and Figure S6.1 (QTL scans for all 12 environments) for the actual data, only those associated with the parameters LUE (22 out of the expected 48 = 12 environments times 4 chromosomes, Table 6.2) and W (three out of the expected 16 = 8 environments with temperatures of 20°C or 25°C times 2 chromosomes, Table 6.2) were found, which represents a poor outcome of this single trait single environment analysis.

6.3.2. AMMI analysis

Table 6.4 gives the ANOVA for the model AMMI5 based on one randomly chosen realization of a genotype-by-environment two-way data table. Similar results are obtained for other two-way data tables simulated from the model in use (Rodrigues et al., 2012a). The genotypes, environments and GEI account for 31.5, 34.4 and 34.1% of the treatment sum of squares (SS). Two interaction principal components were chosen for the AMMI model as in Rodrigues et al. (2012a). This choice was confirmed by the cross-validation proposed by Krzanowski (1987): the W_n values from equation (6.5) for the first five components are 10.371; 1.385; 0.579; 0.475; 0.306, and show the "best" model to have two principal components because only two W_n values are above the cut-off of 0.6.

The AMMI2 biplot is depicted on the left hand side of Figure 6.2. In this figure, the environments with higher genetic variance (NL5-20 and NL5-25, Table 6.1) are farthest away from the origin, showing

an extreme behaviour when compared with the remaining environments. However these environments with higher genetic variance also have higher error variance (Table 6.2), which should be down-weighted to produce a more trustable result. This can be achieved by giving smaller weights to the environments with higher error variance.

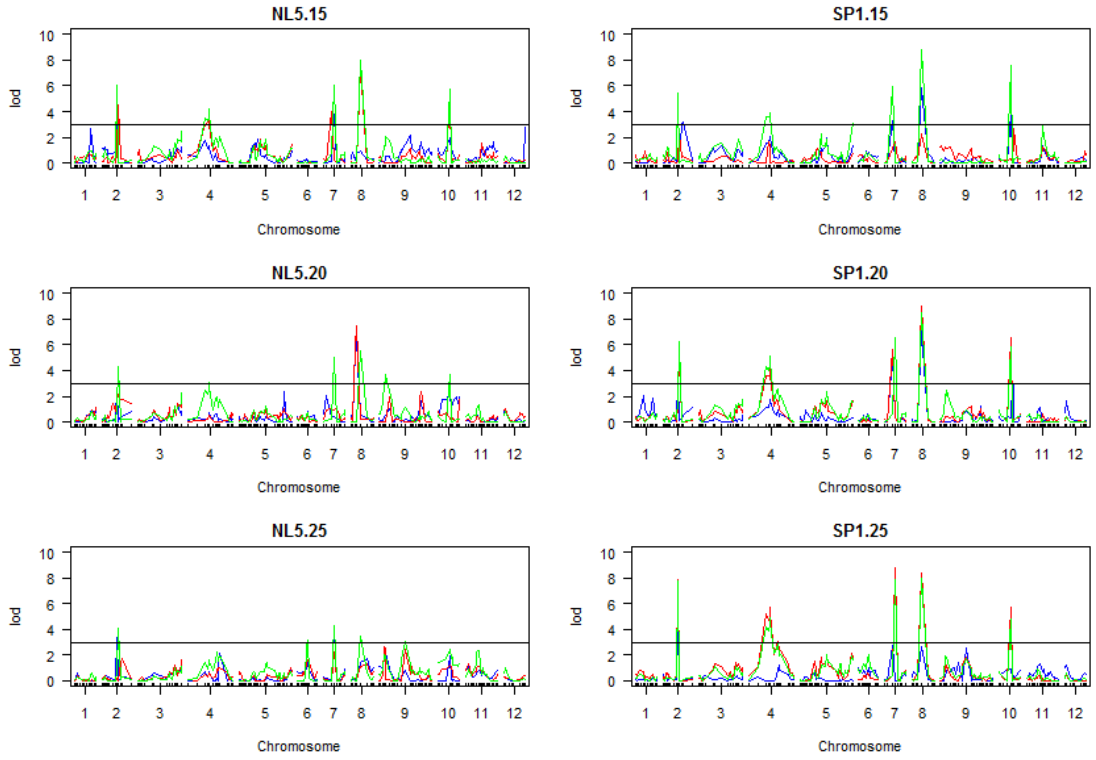

Figure 6.1. QTL scans for 6 environments of the yield data for pepper simulated from the genotype-to-phenotype model with seven physiological parameters (Rodrigues et al. 2012a). Each row represents a different level of temperature. The plots on the left correspond to the highest error variance in this simulated data table and the ones on the right to the lowest. The blue line represents the scans for the actual data, the red for the AMMI2 predicted values, and the green for the WAMMI2 predicted values. All the scans are based on composite interval mapping. The horizontal lines correspond to the thresholds for a LOD score of 3. 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 6.1.

Table 6.4. ANOVA of the AMMI5 model for the simulated yield data for pepper. Results based on one randomly chosen realization of the genotype-to-phenotype crop growth model. The columns of the table show the source of variation, the degrees of freedom (df), the sums of squares (SS) and the mean squares (MS).

Source	df	SS	MS
Total	2399	256089	106.7
Genotypes	199	80774	405.9
Environments	11	88054	8004.9
GEI	2189	87261	39.9
IPC1	209	18122	86.7
IPC2	207	14740	71.2
IPC3	205	11470	56.0
IPC4	203	10074	49.6
IPC5	201	7916	39.4
IPC6—IPC11	1164	24938	21.4

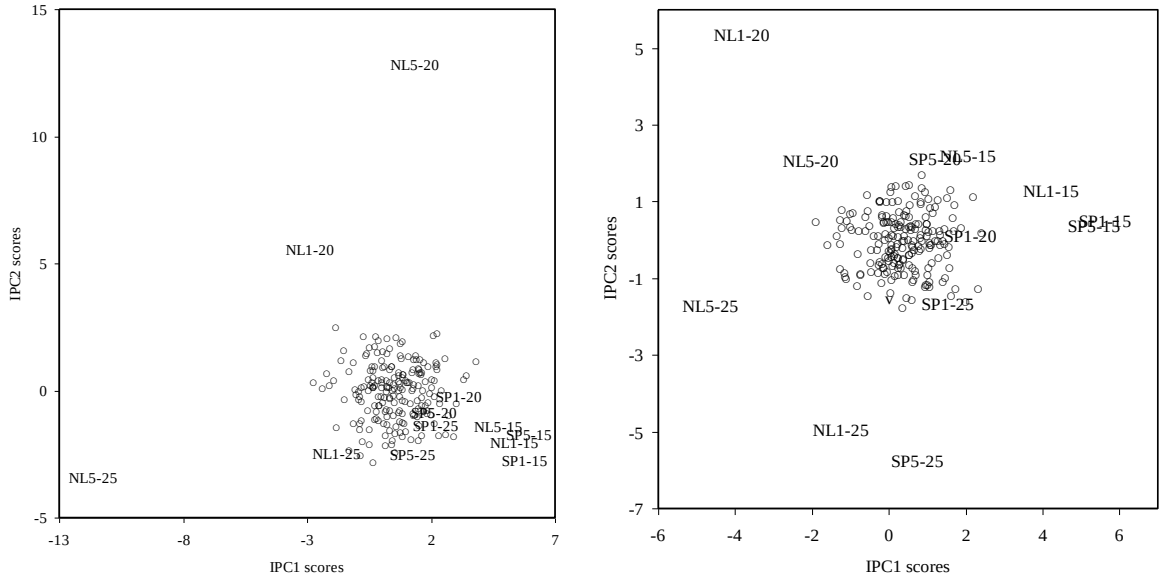

Figure 6.2. AMMI2 (left) and WAMMI2 (right) biplots for one randomly chosen realization. The abscissa represents the first multiplicative term and the ordinate the second. The open dots represent the 200 genotypes and the codes for the 12 environments are defined in Table 6.1.

6.3.3. Weighted AMMI analysis

To avoid considering environments with high error variance as outliers (Gauch et al., 2011) or letting them influence (too much) the results, the weighted AMMI analysis described above was used where the contribution (i.e. the weight) of a given environment to the model fit is the inverse of its error variance (Table 6.1). The WAMMI biplot is given in Figure 6.2 (right). In this plot the environments SP5-20 and SP5-25 ceased to show extreme behaviour. There is also a visible pattern in the environments: (i) the right hand side presents more Spanish environments whereas the left hand side has more Dutch environments; and (ii) the right top corner shows environments with temperatures of 15ºC, the left bottom corner shows environments with temperatures of 25ºC, and in between are placed the environments with temperatures of 20ºC.

6.3.4. AQ analysis and weighted AQ analysis

The AQ analysis is the AMMI analysis followed by QTL scans on the AMMI predicted values (Gauch et al., 2011). The weighted AQ (WAQ) analysis is a generalization of the AQ analysis, where the AMMI analysis is replaced by the weighted AMMI analysis proposed before. This WAMMI and WAQ analyses are particularly useful to analyse data sets whose environments show high heterogeneity in their error variances.

Figure 6.1 shows the AQ (red line) and WAQ (green line) analyses for models with two IPCs. There is a clear improvement from the QTL scans of the actual data to the AQ and WAQ analysis in both the number of detected QTLs and higher LOD scores. As in the biplots of Figure 6.2, the improvement from

the unweighted to the weighted method is visible in Figure 6.1 for AQ and WAQ analysis. As an example, all peaks of environment SP1-15 (lowest error variance, Table 6.1) are below the LOD 3 threshold when using AQ analysis but five (true) QTLs are detected when using the WAQ analysis. This happens because, being SP1-15 the most accurate environment (lowest error variance), its weight is expected to be underestimated by the AMMI2 model but corrected with the WAMMI2.

6.3.5. The 100 simulated data sets and comparison between methods

A more detailed comparison for all the 100 simulated data sets is presented in Figures 6.3 and 6.4. As expected, the worst performance (in terms of detected QTLs) is obtained by the QTL scans of the actual data. Better are the QTL scans on the AMMI2 predicted values (AQ analysis), which, however, do not detect QTLs for some environments in some runs. The WAQ analysis and QTL mixed model framework are the best options in the presence of heterogeneity of error variance across environments. Although the mixed model detects slightly more QTLs, the fixed effects WAQ analysis shows less variance for the number of QTLs (Figures 6.3 and 6.4).

The analysis and interpretation was clearly improved by using the error variances in each environment, which leads us to conclude that the WAMMI biplot is also an improved version (closer to the reality) of the AMMI biplot (Figure 6.2).

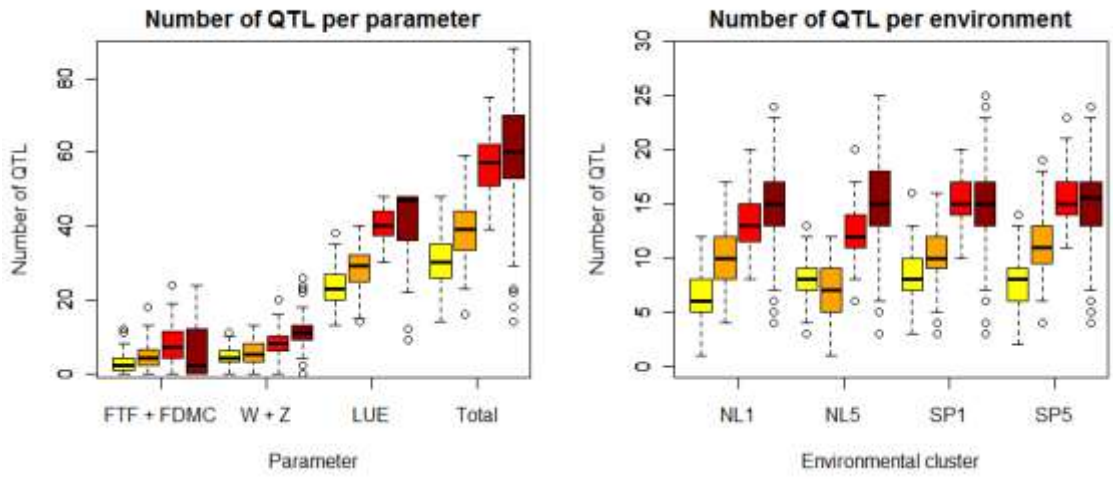

Figure 6.3. Summary of the number of detected QTLs for the actual data (yellow), AMMI2 predicted values (orange), WAMMI2 predicted values (red) and linear mixed model (dark red). The graph on the left hand side shows the box plots for the number of QTLs per model parameter (Table 6.2), and on the right hand side the number of QTLs per environmental cluster (Table 6.1). These values are for QTLs detected when considering an interval of 20 cM centred on the right QTL position. These plots are for a heritability of 0.5 in all environments.

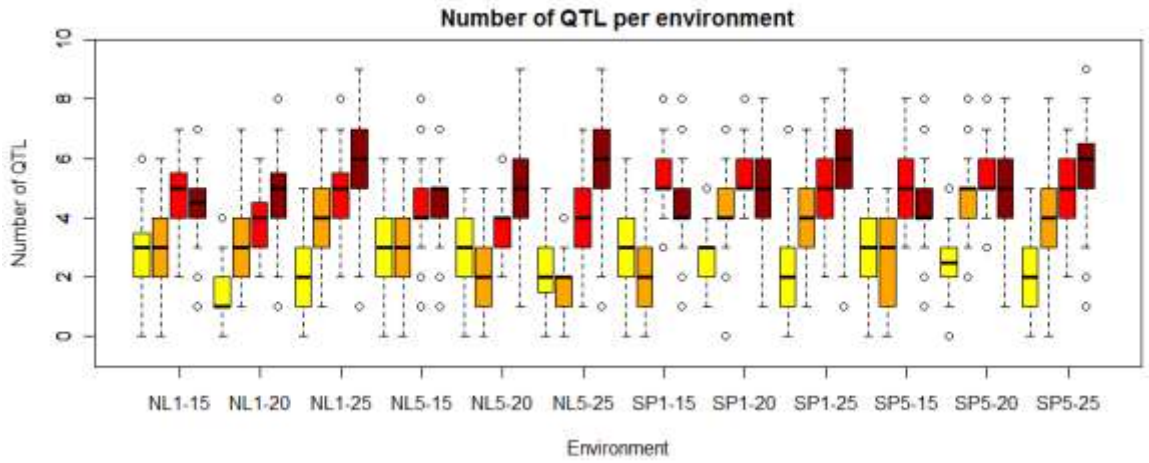

Figure 6.4. Number of QTLs detected per environment for an expected maximum of 7 (environments with temperature of 15ºC or 20ºC, Table 6.2) or 8 (environments with temperature of 25ºC, Table 6.2). The box plots are presented for the actual data (yellow), AMMI2 predicted values (orange), WAMMI2 predicted values (red) and linear mixed model (dark red). These plots are for a heritability of 0.5 in all environments.

6.4. Results for the barley experiment

6.4.1. Preliminary analysis

Two previous studies have applied the AMMI model to the SxM yield data to improve and better understand QTL detections (Romagosa et al., 1996, Gauch et al., 2011). Here we used the genotype-by-environment means for 13 environments (Table 6.3), where the experiment was partially replicated, instead of the means for the original 16 environments. Table S6.1 gives a short summary of findings in the literature about detected QTLs on the SxM yield data. Figure 6.5 (blue line) depicts the QTL scans for the actual data of the 13 environments.

6.4.2. AMMI analysis

Table 6.5 gives the ANOVA for the AMMI5 model. The genotypes, environments and GEI account for 9.2, 67.4 and 23.4% of the treatments sum of squares (SS). The amount of noise in the GEI can be estimated by the product of the interaction degrees of freedom (df) with the error mean square (MS), namely 768.8, which by difference from the total of 2157 implies a GEI signal of 1388.2, or 64.4% (Gauch, 1992, Voltas et al., 2002). IPC1 captures a SS of 566, IPC2 412 and IPC3 287, which includes the most of the signal and little noise because the first principal components tend to capture more signal and less noise (Gauch, 1992).

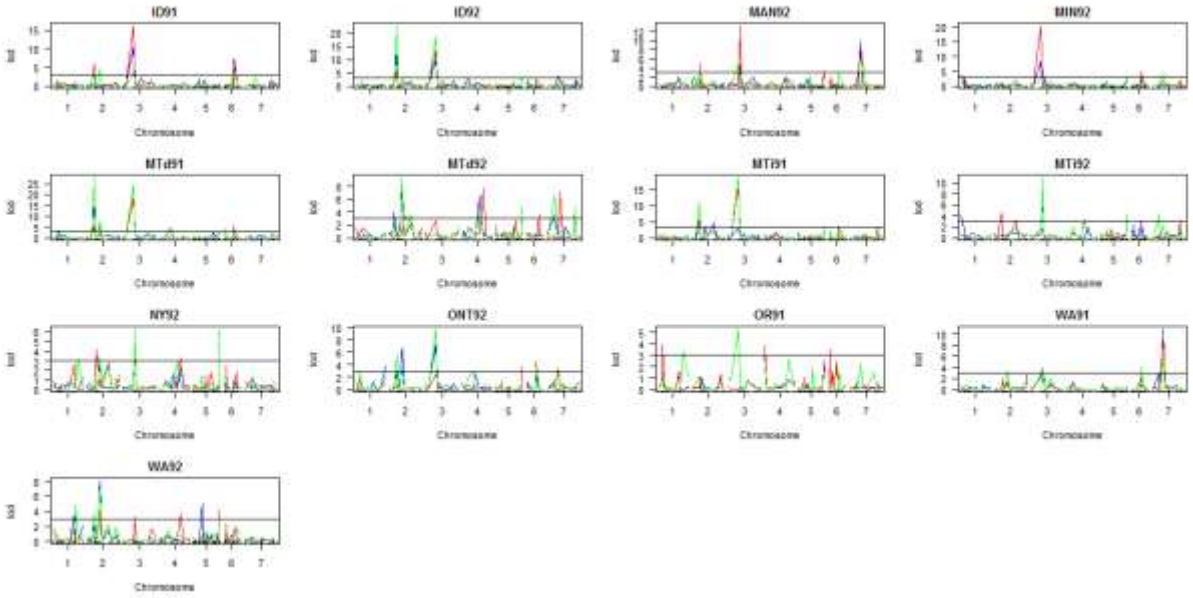

Figure 6.5. QTL scans for the 13 environments for the means of the SxM yield data (blue line), AMMI3 predicted values (red line), and WAMMI3 predicted values (green line). The results are for composite interval mapping and the threshold was set to a LOD score of 3.

Table 6.5. ANOVA of the AMMI5 model for the SxM yield data.

Source	df	SS	MS
Total	3399	9829	2.89
Treatments	1949	9202	4.72
Genotypes	149	844	5.66
Environments	12	6201	516.77
GEI	1788	2157	1.21
IPC1	160	566	3.54
IPC2	158	412	2.61
IPC3	156	287	1.84
IPC4	154	227	1.47
IPC5	152	137	0.90
IPC6—IPC11	1008	528	0.52
Intra Block Error	1450	626	0.43

Within the two studies where the AMMI model was applied to the SxM yield data, Romagosa et al. (1996) found QTLs in the first four IPCs. Subsequently Gauch et al. (2011) considered the AMMI3 based on the Ockham's valley for the root mean squared prediction error following from a jackknife procedure. For this particular data set the cross-validation procedure described before and introduced by Krzanowski (1987) was considered. When computing the W_n values for the first five components of the SxM yield we obtain: 11.019; 0.141; 0.825; 0.675; 0.395. This results in the same limitations as in the example presented by Krzanowski (1987), i.e. the W_n values are not monotonic. Therefore adopting the suggestion provided by Krzanowski (1987) where the W_n values are ordered (11.019; 0.825; 0.675; 0.395; 0.395; 0.141;…) and the number of components should correspond to the number of W_n values greater than 0.6 (Krzanowski, 1987), we consider three principal components. I.e. we used an AMMI model with three interaction principal components.

The first two axes of the AMMI3 model are depicted in Figure 6.6 (left). As before, the environments with higher error variance tend to be placed away from the origin. A similar pattern was found by Gauch et al. (2011) where the environment OR91 was considered as an outlier.

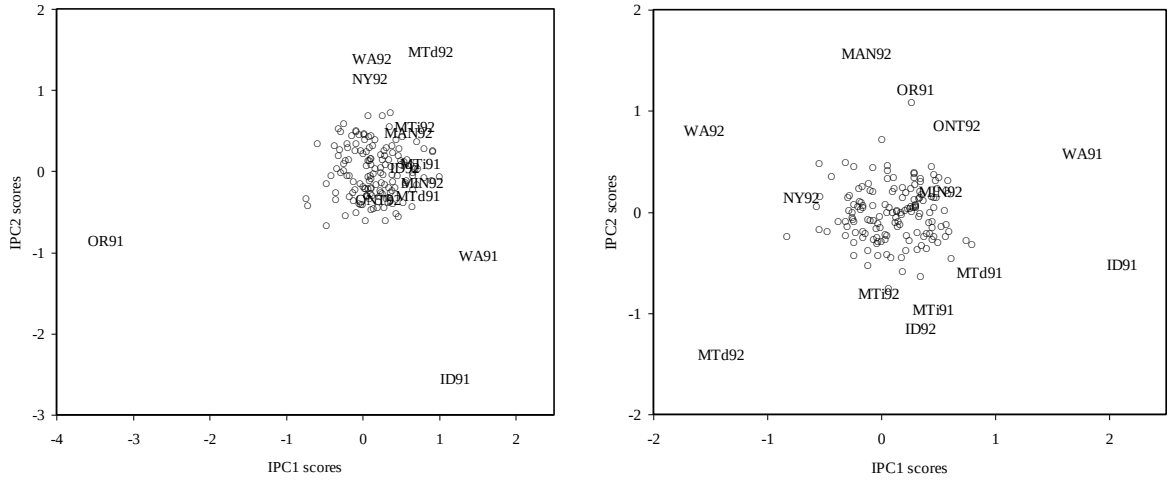

Figure 6.6. Biplots for the first two axes of AMMI3 (left) and WAMMI3 (right) models, for the SxM yield data. The abscissa represents the first multiplicative term and the ordinate the second. The open dots represent the 150 genotypes and the codes for the 13 environments are defined in Table 6.3.

6.4.3. Weighted AMMI analysis

Since the data in use is partially replicated, the cell means bring more information when there are two observations for the genotype environment combination. Therefore we have used the matrix of weights as defined by equation (6.7), with $m = max_j(1/\sigma^2_{\varepsilon_j})$, $\sigma^2_{\varepsilon_j}, j = 1, \ldots, 13$, is the error variance for environment j, $Nrep_{i,j}, i = 1, \ldots 150, j = 1, \ldots, 13$, is 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 6.6 (right) shows the first two axes of the WAMMI3 model, weighted by $\boldsymbol{W}$ as in (6.7), and represents 75.7% of the total variance explained by the WAMMI3 model. As in the first example (simulated data), the environments with higher influence in the AMMI analysis have a more homogeneous distribution when using the WAMMI3 model (right hand side of Figure 6.6).

6.4.4. AQ analysis and weighted AQ analysis

Figure 6.5 shows the QTL scans for the AMMI3 (red line) and WAMMI3 (green line) predicted values. The LOD scores show an increase when the QTL scans are made for the AMMI3 predicted values instead of the actual data. The same pattern is observed for most of the environments. When the AQ analysis is replaced by the WAQ analysis the LOD scores become higher for the three environments with lowest LOD scores in the actual data and AMMI3 predicted values: OR91, NY92 and WA92. The two QTLs on chromosome 2 (Malosetti et al., 2004) are now visible in Figure 6.5 (red and green lines) for the most of the environments (more clear for WAQ analysis).

6.4.5. Weighted AQ analysis and comparison with QTL mixed linear models

Figure 6.7 presents a general comparison between the four approaches used here: direct QTL scans of the actual data; AQ analysis; WAQ analysis; and QTL mixed model framework. The exact positions can be found in Table S6.1. Most of the QTLs, detected with the WAQ analysis and the QTL mixed model framework , were either found in previous analyses or are very close to those QTLs (Romagosa et al., 1996, Romagosa et al., 1999, Hayes et al., 1993, Lacaze et al., 2009, Larson et al., 1996, Malosetti et al., 2004, Zhu et al., 1999, Gauch et al., 2011). Figure S6.6 shows the genome scan for the SxM yield data using the QTL mixed model framework.

On chromosome 1, between 116.7 and 170.1 cM, the linear mixed model identifies a QTL in two environments and WAQ analysis in four environments. For chromosome 2 two QTLs are identified by linear mixed models (in six and eight environments, respectively) and WAQ analysis (in five and five environments, respectively). The QTL on chromosome 3 is identified in 11 of the 13 environments by the linear mixed model and WAQ analysis. No QTL is detected by the linear mixed model on chromosomes 4 and 5. However, the WAQ analysis identifies a QTL on chromosome 4 at the same place as Lacaze et al. (2009) did. A QTL on chromosome 5 is detected for five environments by the WAQ analysis. Between 53.1 cM and 72.5 cM of chromosome 6 there is a QTL detection for linear mixed model and WAQ analysis, in five and four environments, respectively. A QTL is detected between 45.6 cM and 78.2 cM of chromosome 7 for linear mixed model and WAQ analysis, in seven and four environments, respectively.

When comparing the methods under study we can conclude again that the QTL scans of the actual data have the worst performance in terms of QTL detection (Figures 6.5 and 6.6). The AQ analysis finds more QTLs but the WAQ analysis is more similar with the results from the QTL mixed model analysis (Figure 6.6 and Table S6.1), which we believe are more credible.

6.5. Discussion

6.5.1. Weighted AMMI analysis

The WAMMI model proposed here is a generalization of the standard AMMI analysis (Gauch, 1992) that is able to account for heterogeneity of error variances across environments in a multiple-environment trial. This extension also allows the generalisation of the AQ analysis where the QTL scans are based on the AMMI predicted values (Gauch et al., 2011), which makes the results become similar to the often used QTL mixed model framework (Boer et al., 2007, Malosetti et al., 2004, Malosetti et al., 2008).

In this chapter we used an algorithm based on the EM procedure proposed by Srebro and Jaakkola (2003) to conduct the weighted low-rank approximation. However, many alternatives can be found in the literature: maximum likelihood principal component analysis (Wentzell et al., 1997); a steepest descent algorithm and a Newton-like algorithm (Manton et al., 2003); and the use of a weighted rank correlation coefficient instead of the usual Pearson's (da Costa et al., 2011), among others. We chose the EM approach because of its easy implementation and good behaviour for our type of data.

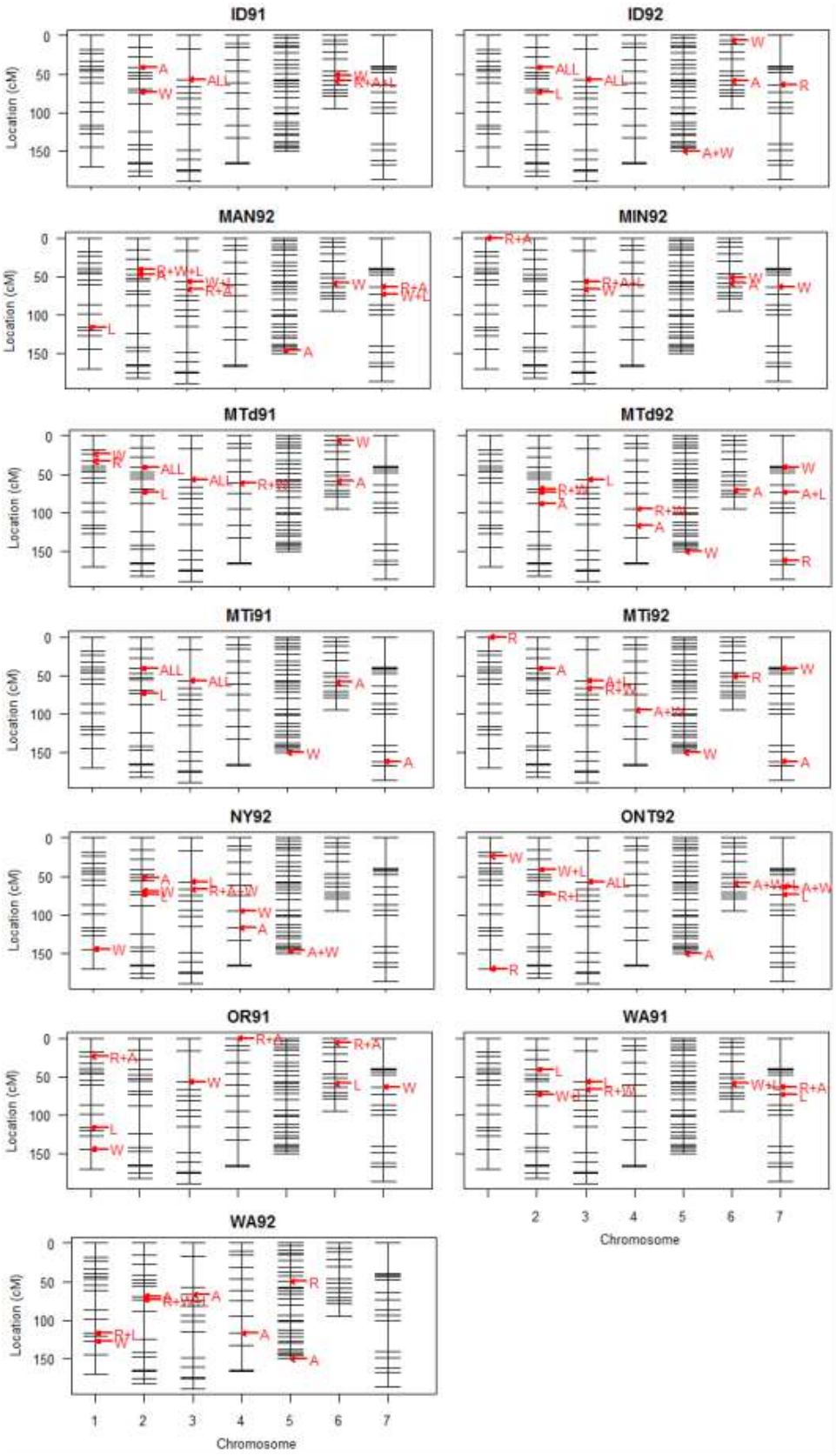

Figure 6.7. Genetic map with the information of the place where a QTL was detected for each of the four approaches: QTL scans of the actual data (R); QTL scans of the AMMI3 predicted values (A); QTL scans of the WAMMI3 predicted values (W); and linear mixed model framework (L), for the SxM yield data in all 13 environments. "ALL" means that the QTL was detected with all the four approaches.

6.5.2. AMMI model selection

Much research has been done about the choice of the "optimal" number of principal components in a PCA in general, and the number of multiplicative terms in the AMMI model in particular. Besides the cross-validation proposed by Krzanowski (1987) and used in this chapter to decide on the number of multiplicative terms in the AMMI model, there are many options widely in AMMI literature. These include the signal to noise ratio (Gauch, 1992), the Ockham's valley (MacKay, 1992, Gauch, 2006), a cross-validation (Gauch, 1992, Gauch, 1988, Piepho, 1994) which can be performed with , e.g. the software MATMODEL version 3.0 (Gauch, 2007), and significance tests. Gollob (1968) was the first proposing *F* tests to help choosing the number of multiplicative terms. Other *F* tests were also suggested: F_{GH2} (Cornelius et al., 1992, Cornelius, 1993), and F_R (Piepho, 1995). The cross-validation is usually seen as a conservative method because it refers to the modelling of a subset of the original data, which is expected to be less accurate than to model all data after the model choice (Cornelius, 1993, Annicchiarico, 1997a). Although it is commonly used in AMMI literature, the cross-validation procedure proposed by Gauch (1992) is based on within experiment variation, which may not be the most suitable form of variation for a cross-validation procedure for a multiple environment data set. Gollob's *F* test is too liberal (Piepho, 1997), F_{GH2} is less conservative than F_R (Annicchiarico, 1997a) and both tend to retain a higher number of multiplicative terms than the cross-validation. However, these *F* tests are used for determining the number of non-null multiplicative terms, which is different from finding the optimal number of terms for a prediction purposes (Cornelius, 1993, Piepho, 1997). The number of multiplicative terms for a predictive model should then be lower than the found significant by a significance *F* test (Piepho, 1997). Other alternatives widely used in PCA but not in AMMI modelling are parallel analysis (Horn, 1965); minimum average partial (MAP) de Velicer (Velicer, 1976), and very simple structure (VSS) (Revelle and Rocklin, 1979). These methods can only be applied to the two-way table of means, and should be applied to the multiplicative part of the data, i.e. after removing the genetic and environmental main effects.

The variety of possible methods is wide and so is the outcome. From an exhaustive analysis of related literature usually two or three axes are used to model the data because one component is (usually) not enough to capture the entire signal present in the data, and more than three components are already capturing a big amount of noise and are more difficult to visualize graphically. Moreover, in multi-environment trials, usually, there is no further information beyond two or three principal components.

6.5.3. The influence of the heritability in the results

In field crops, the range of heritabilities is wide, and may vary from about 0.3 for yield in cereals in open environments (Clarke and Townleysmith, 1986, Saeed et al., 2007), to more than 0.7 for tomato (Reif et al., 2009) or pepper (do Rego et al., 2011, Sood et al., 2009) in greenhouse experiments. When the heritability of the environments under study decreases the WAQ analysis tend to out-perform the QTL mixed model framework (h^2=0.3, Figures S6.2 and S6.3); whereas for higher heritability of the

environments the QTL mixed model framework out-performs the WAQ analysis (h^2=0.8, Figures S6.4 and S6.5) but detects some QTLs which are likely to be false positives (e.g. a few detections on chromosome 12 and the very unlikely scenario of 10 QTLs found in several environments, Figure S6.5, Table 6.2). Both WAQ analysis and QTL mixed model framework out-perform AQ analysis, and all of them out-perform the QTL scans of the actual data. For all these comparison we should bear in mind that the thresholds for the WAQ analysis are fully comparable with the thresholds for AQ analysis and with the thresholds for the QTL scans on the actual data. However, because of the different methodologies and different software, the thresholds for WAQ analysis are not fully comparable with the QTL mixed model framework, but an approximation for illustration purposes. It should be remarked that the mixed model QTL mapping was used with default multiple testing corrections as set in Genstat 14 (Payne et al., 2011). Some playing around with those settings might have produced results closer to WAQ.

6.5.4. Alternatives to the QTL mixed model methodology

Boer et al. (2007) suggested the possibility of having different methodologies performing as well as the QTL mixed model approach. They named Bayesian based methods and penalized regression as possibilities for similar analyses. The AQ analysis, i.e. use the AMMI expected values in the QTL scans across environments, first proposed by Gauch et al. (2011), is also an alternative to the QTL mixed model framework. However, the AQ analysis is not general enough to account for different error variances across environments which this chapter generalizes by introducing the weighted SVD in the AMMI analysis.

The results presented in this chapter are very encouraging because of several factors: (i) the WAQ analysis can be performed with the package qtl (Broman and Sen, 2009) of the open source R software (Team, 2009); (ii) the computation time to obtain the QTL scans and its summary is much shorter than the QTL mixed model framework in GenStat (Payne et al., 2011); and (iii) the results are very similar with the QTL mixed model output (Figures 6.3, 6.4, 6.7 and Table S6.1). It is also remarkable how the inclusion of the information about the error variances improves that much the results when the heritability of the trait/environment decreases (Figures 6.3, S6.2 and S6.4), comparing with the AQ analysis and the QTL scans of the actual data, and makes them very similar to the QTL mixed model methodology (Figure 6.3, 6.4, 6.7 and Table S6.1). So, the WAQ analysis is easy to apply with open source software and faster to run when compared with the QTL mixed linear model framework. Moreover, the WAMMI model and WAQ analysis are fully applicable to a wide range of fields such as plant breeding, crop sciences, genetics, microarray experiments (Crossa et al., 2005), rDNA studies (Adams et al., 2002); plant and microbial populations' growth across several environmental conditions (Culman et al., 2008, Culman et al., 2009) and animal sciences (Barhdadi and Dube, 2010).

6.6. Supplementary material

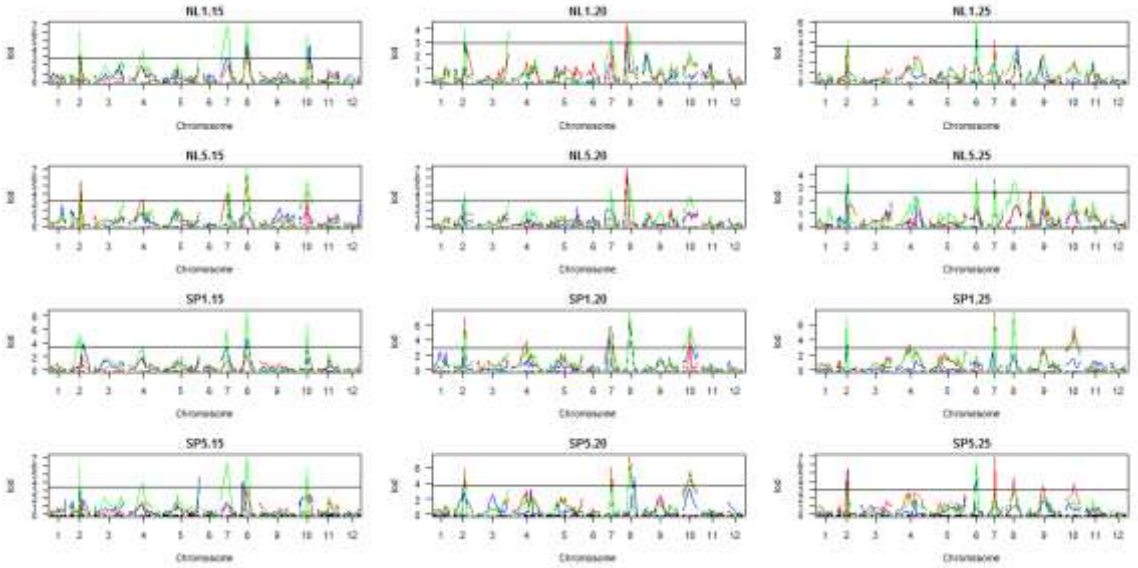

Figure S6.1. QTL scans for the 12 environments of the yield data for pepper simulated from the physiological genotype-to-phenotype model with seven physiological parameters (Rodrigues et al., 2012a). Each column represents a different level of temperature. The first row corresponds to the highest error variance in this realization and the second to the lowest. The black line represents the scans for the actual data, the blue for the AMMI2 predicted values, and the red for the WAMMI2 predicted values. All the scans are based on the composite interval mapping. The horizontal lines correspond to the thresholds for a LOD score of 3. These scans are based on one randomly chosen realization out of the 100 simulations.

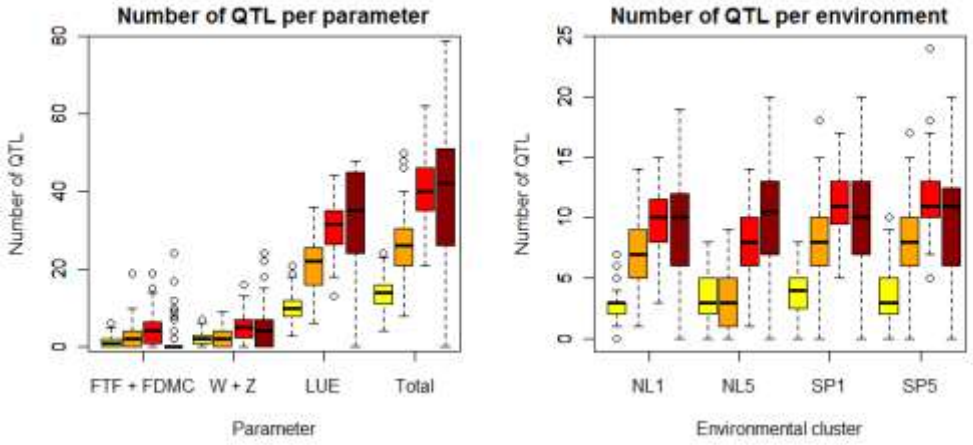

Figure S6.2. Summary of the number of detected QTLs for the actual data (yellow), AMMI2 predicted values (orange), WAMMI2 predicted values (red) and linear mixed model (dark red). The graph on the left hand side shows the box plots for the number of QTLs per model parameter (Table 6.2), and on the right hand side the number of QTLs per environmental cluster (Table 6.1). These values are for QTLs detected when considering an interval of 20 cM centred in the right position. These plots are for a heritability of 0.3 in all environments.

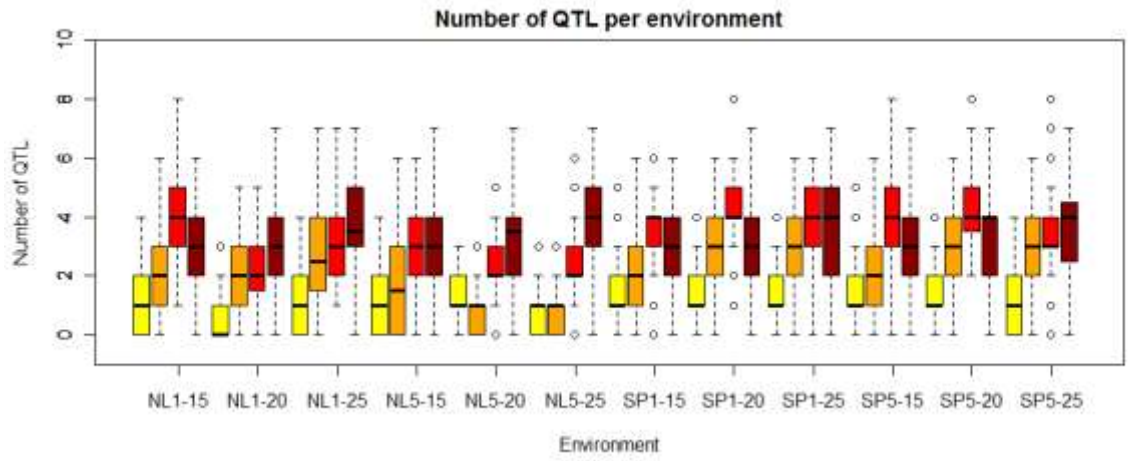

Figure S6.3. The top panel shows the number of QTLs detected per environment for a expected maximum of 7 (environments with temperature of 15ºC or 20ºC, Table 6.2) or 8 (environments with temperature of 25ºC, Table 6.2). The bottom panel shows the LOD scores per environment. The box plots are presented for the actual data (yellow), AMMI2 predicted values (orange), WAMMI2 predicted values (red) and linear mixed model (dark red). These plots are for an heritability of 0.3 in all environments.

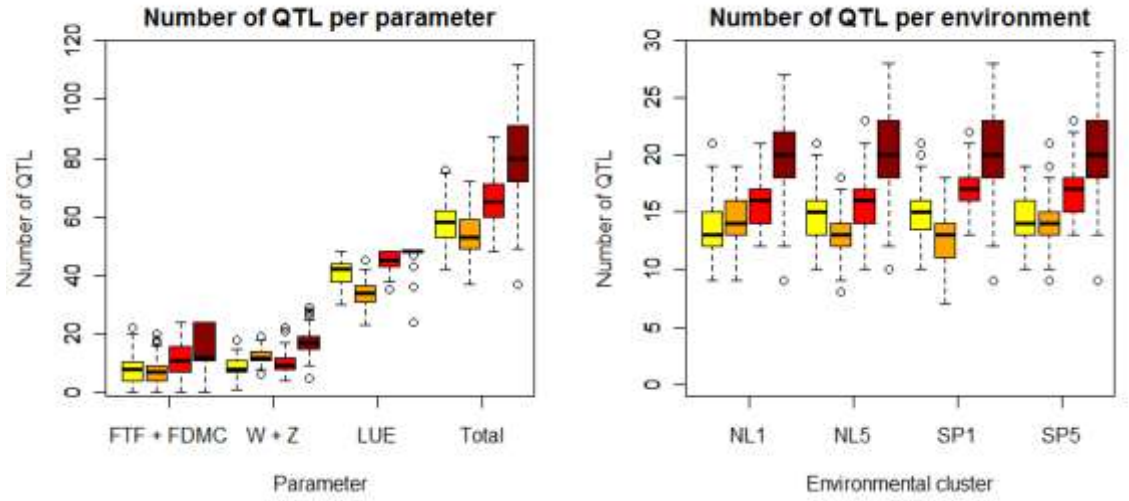

Figure S6.4. Summary of the number of detected QTLs for the actual data (yellow), AMMI2 predicted values (orange), WAMMI2 predicted values (red) and linear mixed model (dark red). The graph on the left hand side shows the box plots for the number of QTLs per model parameter (Table 6.2), and on the right hand side the number of QTLs per environmental cluster (Table 6.1). These values are for QTLs detected when considering an interval of 20 cM centred in the right position. These plots are for a heritability of 0.8 in all environments.

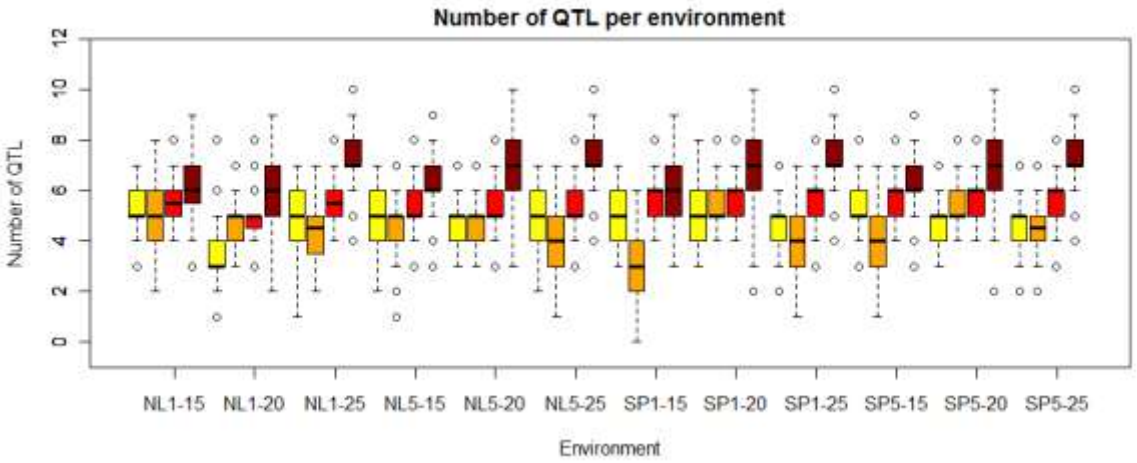

Figure S6.5. The top panel shows the number of QTLs detected per environment for a expected maximum of 7 (environments with temperature of 15ºC or 20ºC, Table 6.2) or 8 (environments with temperature of 25ºC, Table 6.2). The bottom panel shows the LOD scores per environment. The box plots are presented for the actual data (yellow), AMMI2 predicted values (orange), WAMMI2 predicted values (red) and linear mixed model (dark red). These plots are for an heritability of 0.8 in all environments.

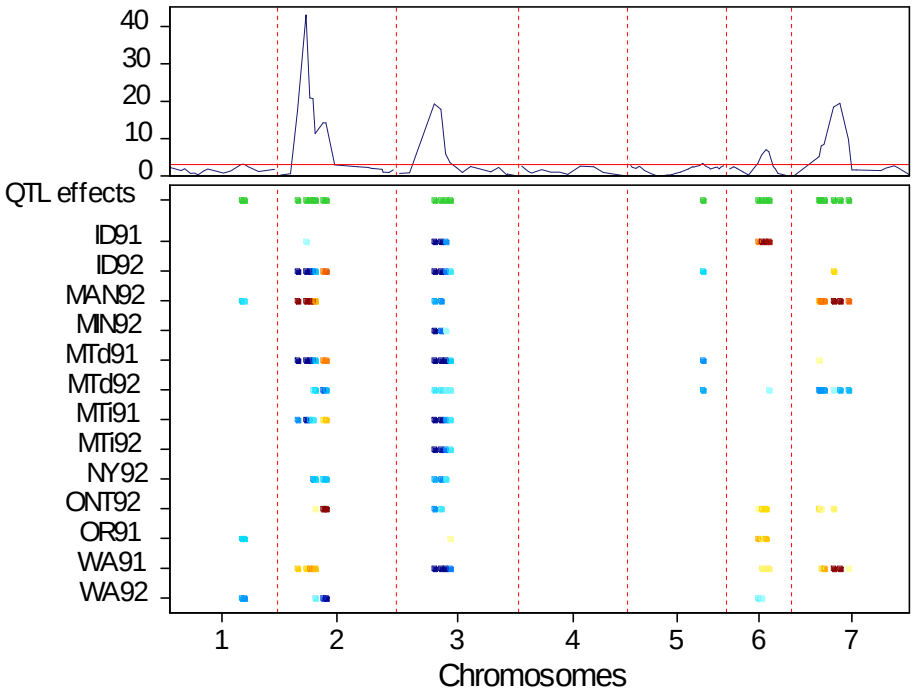

Figure S6.6. Genome scan for the means of the SxM yield data. The $-\log_{10}(p)$-values for the QTL main effects plus QEI are shown. The red horizontal line is the 5% genomewide significance threshold. The green horizontal line in the bottom section summarizes the top panel. The environment specific QTL effects are shown. Blue (red) indicates that parent Steptoe (Morex) has significantly higher yield contribution. The considered variance-covariance (VCOV) structure was the factor analytic with two multiplicative terms.

Table S6.1. Chromosome (Chr) and respective positions (Pos) where a QTL was detected for each of the four approaches: QTL scans of the actual data (R); QTL scans of the AMMI3 predicted values (A); QTL scans of the WAMMI3 predicted values (W); and linear mixed model framework (L), for the SxM yield data in all 13 environments. "ALL" means that the QTL was detected with all the four approaches. The last column gives the reference where the same detection was observed. "None" indicates that no reference was found where a similar QTL was detected.

Chr	Pos	ID91	ID92	MAN92	MIN92	MTd91	MTd92	MTi91	MTi92	NY92	ONT92	OR91	WA91	WA92	Reference
1	[0; 33.5]				R+A	R+W			R		W	R+A			(Hayes et al., 1993)
1	[116.7; 170.1]			L						W	W	W+L		R+W+L	None
2	[41.2; 52.6]	A	ALL	ALL		ALL		ALL	A	A	W+L		L		(Gauch et al., 2011) (Hayes et al., 1993) (Lacaze et al., 2009) (Malosetti et al., 2004) (Romagosa et al., 1996) (Romagosa et al., 1999) (Zhu et al., 1999)
2	[68.8; 88.2]	W	L			L	R+W+L	L		W+L	R+L		W+L	ALL	(Gauch et al., 2011) (Malosetti et al., 2004)
3	[73; 83.6]	ALL	ALL	ALL	ALL	ALL	L	ALL	ALL	ALL	ALL	W	R+W+L	A	(Gauch et al., 2011) (Hayes et al., 1993) (Lacaze et al., 2009) (Larson et al., 1996) (Romagosa et al., 1999)
4	1.4											R+A			None
4	63.2					R+W									(Lacaze et al., 2009)
4	96.5						R+W		A+W	W					None
4	118.3						A			A				A	None
5	49.6													R	None
5	[146.3; 150.8]		A+W	A			W	W	W	A+W	A			A	None
6	8.1		W			W						R+A			None
6	[53.1; 72.5]	ALL	A	W	A+W	A	A	A	R		A+W+L	L	W+L		(Gauch et al., 2011) (Romagosa et al., 1996) (Romagosa et al., 1999)
7	[45.6; 78.2]		R	ALL	W		A+W+L		W		A+W+L	W	ALL		(Gauch et al., 2011) (Romagosa et al., 1996)

File S 1. The R code for the weighted low-rank SVD (Srebro and Jaakkola, 2003) (adapted from the Marlin's MatLab code in http://www.cs.toronto.edu/~marlin/code/wsvd.m).

```
# Imputs
# Y – (IxJ) data matrix
# W – (IxJ) weight matrix with 0 <= Wij <= 1; I = 1,…,I; j = 1,…,J
# N – rank of approximation
#
# Outputs
# U,D,V such that Y~UDV'

Y<- read.csv("data.csv", header=T)      # Read the original data set (IxJ)
X<- matrix(0,ngen,nenv)                 # Matrix (IxJ) with zero in all positions to initialize the algorithm
aux<- matrix(1,ngen,nenv)
Xold=Inf*aux
Err=Inf                                 # Initial distance between consecutive iterations – X(i) and X(i+1)
eps<- 1e-10                             # Maximum admissible distance between consecutive iterations

while(Err>eps){                         # Repeats the code until the distance between X(i) and X(i+1) is below eps=1e-10
   Xold=X                               # Update Xold to X(i)
   wsvd<- svd(W*Y + (1-W)*X)            # Weighted SVD
   U<- wsvd$u                           # Left singular vectors
   D<- diag(wsvd$d)                     # Singular values
   V<- wsvd$v                           # Right singular vectors
   D[(N+1):length(wsvd$d),(N+1):length(wsvd$d)]<- 0      # Discard singular values above N
   X<- U %*% D %*% t(V)                                  # Update X (i.e. compute X(i+1))
   Err=sum(sum((X-Xold)^2))                              # Update the distance between consecutive iterations (Err)
}
```

Chapter 7

7. General Discussion

7.1. Summary

In this thesis, we have described and applied the most standard techniques to analyze and to structure genotype-by-environment interaction (GEI), as well as quantitative trait loci (QTL) –by-environment interaction (QEI). Despite the wide range of available references and techniques to explore and better understand GEI and QEI (Malosetti et al., 2010), not all of them are available to all breeders and researchers. In some cases, the statistical methods are too complex to be computationally implemented and applied by non-statisticians, in other cases, although these complex techniques are already well implemented in statistical packages, the software is commercial and too expensive for developing countries budgets, where the statistical improvements are slow to arrive.

One of the goals of this thesis was to provide a strategy to simulate and to model GEI and QEI in complex traits, with the example of yield, based on a number of physiological parameters purely genotype dependent. This was done by using an eco-physiological genotype-to-phenotype model with seven parameters defined with a simple QTL basis.

One other goal of this thesis is to propose strategies and methodologies to improve the detection and understanding of QTLs, especially those exhibiting QEI in the context of multi-environment trials, using open source software (e.g. QTL Cartographer, MATMODEL and R/qtl). The first of the strategies proposed in this thesis is a two-stage approach where the QTL scans are based on the AMMI predicted values (AQ analysis). This allows gaining accuracy in the phenotypic data, because each "new environment" (i.e. new environmental predicted values) gains "strength" from the other environments. This improvement happens for parsimonious models, where only the signal is taken in consideration and the noise components discarded. The single trait single environment QTL scans obtained with the AMMI predicted values can then be ordered by AMMI scores in order to analyse patterns with ecological or biological interpretation. The second strategy is a three-stage approach that uses a weighted version of the AMMI model (WAMMI, proposed in this thesis) to obtain the WAMMI predicted values to be used in the QTL scans (WAQ analysis). The WAMMI model generalizes the AMMI model for the cases where the error variance across environments is heterogeneous, by giving higher weights to environments with lower error variances (and consequently more accurate).

The AQ and WAQ analyses were compared with the QTL scans of the actual data and with the QTL mixed model framework. It is remarkable how the inclusion of the information about the error variances improves that much the results when the heritability of the trait/environment decreases (Figures 6.3, S6.2

and S6.4), comparing with the AQ analysis and the QTL scans of the actual data, and makes them very similar to the QTL mixed model methodology (Figure 6.3, 6.4, 6.7 and Table S6.1).

7.2. The usefulness of simulation models

The results of eco-physiological genotype-to-phenotype model described and analysed in Chapter 5 were compared with greenhouse experiments. The simulation model considered 36 environments defined by:

- 2 locations: Almeria, Spain, and Wageningen, The Netherlands;
- 3 years of historical daily global radiation data for each location;
- 2 CO_2 levels: 370 or 1000 μmol mol^{-1};
- 3 levels of temperature constant along the growing season: 15, 20 or 25ºC.

For each of these 36 environments, 500 genotypes were simulated based on drawings from a multivariate normal distribution for the seven physiological parameters described in Chapter 5, i.e.

$$Phen = [LUE^{max}, K, B, FTF, W, FDMC, Z] \sim MVN(\boldsymbol{\mu}, \boldsymbol{V}),$$

with

$$\boldsymbol{\mu} = [0.87,\ 0.6,\ 0.7,\ 0.000378,\ 0.65,\ 0.04,\ 0.0774];$$

$$\boldsymbol{V} = diag\,(0.1742,\ 0.052,\ 0.042,\ (3.78 \times 10^{-5})^2,\ 0.042,\ 0.0112,\ 0.005082).$$

The pepper experiments comprise 149 genotypes from recombined inbred lines of pepper carried out in 2 locations: Almeria, Spain, and Wageningen, The Netherlands; and during 2 growing seasons: January–June 2010 (SP1, NL1), July–December 2010 (SP2, NL2). The population of recombinant inbred lines was obtained from a cross between Yolo Wonder and CM334 (Figure 7.1).

Figure 7.1. Parents (Yolo Wonder and CM334) and F1 of the recombined inbred lines of pepper population (left) and glasshouse experiments (right).

The yield obtained with the greenhouse experiments in the two Spanish environments (SP1 and SP2) was compared with the output of the genotype-to-phenotype crop growth model. The model was calibrated (Chater 5) by using the QTL information observed in several quantitative traits such as leaf area, dry weight, number of internodes, proportion of total biomass due to fruit, etc. (Alimi et al., 2012).

A preliminary comparison is presented in Figure 7.2 which shows similar behaviour in the observed and simulated yield. Here we are more interested to see whether the top-yielding genotypes are the same in the observed and simulated data and not in the magnitude of the yield itself. We can conclude that 67.1% of the top-yielding 10% genotypes are the same for the simulated and observed data, in both environments SP1 and SP2.

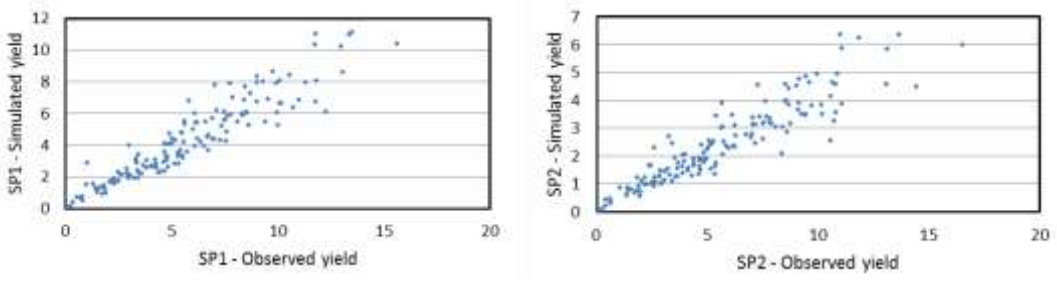

Figure 7.2. Observed (abscissa) and simulated (ordinate) yield for the pepper population in SP1 (left) and SP2 (right).

When comparing the QTL analyses of the observed and simulated data we can find some similarities. The QTL scans for the simulated and observed yield for the two environments (SP1 and SP2) under study are presented in Figure 7.3. When comparing the top panel with the bottom panel in SP1, it is clear the detection of the same QTL in chromosome 9. The QTL on the observed data in chromosome 4 is also a peak in the simulations and the opposite for the QTL detected for the simulated data in chromosome 2b. When the comparison is made for SP2, the similarities are greater:

1. the same QTLs are detected in both simulated and observed data in chromosomes 2b and 9;
2. the QTLs detected in chromosomes 4 and 11a for the observed data have high peaks in the simulated data;
3. an high peak in chromosome 7a is observed for both simulated and observed data.

This preliminary analysis show that the patterns of the observed data are being detected by the simulation model when the QTL information is included, which will allow better predictions of the behaviour of a given genotype along the growing season. However, further analyses should be done considering all the genetic information available (i.e. all the detected QTLs) to try to get (even) better results.

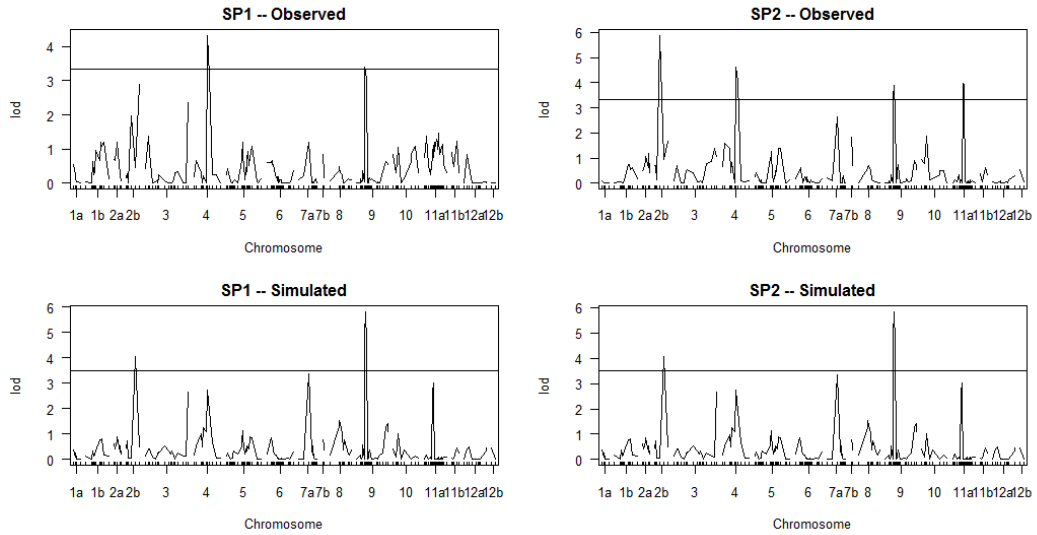

Figure 7.3. QTL scans for the observed (top panel) and simulated (bottom panel) yield in SP1 (left) and SP2 (right).

7.3. Final remarks

The aim of this project was to compare and to develop new methodologies that can potentially improve the detection and the understanding of GEI and QEI in the context of multi-environment trials: the AQ analysis and the WAQ analysis. These new methodologies were proposed in Chapters 4 and 6 and were compared with each other and with the QTL mixed model framework (Boer et al., 2007, Malosetti et al., 2004) by using real data and simulated data using a genotype-to-phenotype crop growth model (Chapter 5).

The improvement in AQ analysis by considering the information about error variances in the WAMMI model (i.e. the WAQ analysis) is remarkable, being the results very similar with the QTL mixed model methodology.

The WAQ analysis is easy to apply with open source software and faster to run when compared with the QTL mixed linear model framework. Moreover, the WAMMI model and WAQ analysis are fully applicable to a wide range of fields such as plant breeding, crop sciences, genetics, microarray experiments (Crossa et al., 2005), rDNA studies (Adams et al., 2002); plant and microbial populations' growth across several environmental conditions (Culman et al., 2008, Culman et al., 2009) and animal sciences (Barhdadi and Dube, 2010).

References

AASTVEIT, A. H. & MEJZA, S. 1992. A selected bibliography on statistical methods for the analysis of genotype x environment interaction. *Biuletyn Oceny Odmian,* 24-25, 83-97.

ADAMS, G. C., WU, N. T. & EISENBERG, B. E. 2002. Virulence and double-stranded RNA in Sphaeropsis sapinea. *Forest Pathology,* 32, 309-329.

ALARCÓN, S. A., PEÑA, M. G., DIAS, C. T. S. & KRZANOWSKI, W. J. 2010. An alternative methodology for imputing missing data in trials with genotype-by-environment interaction. *Biometrical Letters,* 47, 1-14.

ALIMI, N. A., BINK, M. C. A. M., DIELEMAN, A., VOORRIPS, R. E., NICOLAÏ, M., PALLOIX, A. & VAN EEUWIJK, F. A. 2012. Mapping of Quantitative Trait Loci for crop growth traits of pepper in multiple environments (to be submited).

ANDERSON, J. A., SORRELLS, M. E. & TANKSLEY, S. D. 1993. Rflp Analysis of Genomic Regions Associated with Resistance to Preharvest Sprouting in Wheat. *Crop Science,* 33, 453-459.

ANNICCHIARICO, P. 1997a. Additive main effects and multiplicative interaction (AMMI) analysis of genotype-location interaction in variety trials repeated over years. *Theoretical and Applied Genetics,* 94, 1072-1077.

ANNICCHIARICO, P. 1997b. Joint regression vs AMMI analysis of genotype-environment interactions for cereals in Italy. *Euphytica,* 94, 53-62.

ANNICCHIARICO, P. 2002. Genotype x Environment Interactions - Challenges and Opportunities for Plant Breeding and Cultivar Recommendations. *FAO Plant Production and Protection Papers* [Online].

ANNICCHIARICO, P. 2009. Coping with and exploiting genotype-by-environment interactions. *In:* CECCARELLI, S., E.P., G. & WELTZIEN, E. (eds.) *Plant breeding and farmer participation.* Rome: FAO.

ANNICCHIARICO, P., BELLAH, F. & CHIARI, T. 2005. Defining subregions and estimating benefits for a specific-adaptation strategy by breeding programs: A case study. *Crop Science,* 45, 1741-1749.

ANNICCHIARICO, P., BELLAH, F. & CHIARI, T. 2006. Repeatable genotype X location interaction and its exploitation by conventional and GIS-based cultivar recommendation for durum wheat in Algeria. *European Journal of Agronomy,* 24, 70-81.

ANNICCHIARICO, P., ROYO, C., BELLAH, F. & MORAGUES, M. 2009. Relationships among adaptation patterns, morphophysiological traits and molecular markers in durum wheat. *Plant Breeding,* 128, 164-171.

BARCHI, L., BONNET, J., BOUDET, C., SIGNORET, P., NAGY, I., LANTERI, S., PALLOIX, A. & 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, 51-60.

BARCHI, L., LEFEBVRE, V., SAGE-PALLOIX, A. M., LANTERI, S. & PALLOIX, A. 2009. QTL analysis of plant development and fruit traits in pepper and performance of selective phenotyping. *Theoretical and Applied Genetics,* 118, 1157-1171.

BARHDADI, A. & DUBE, M. P. 2010. Testing for Gene-Gene Interaction with AMMI Models. *Statistical Applications in Genetics and Molecular Biology,* 9.

BARIL, C. P., DENIS, J. B., WUSTMAN, R. & VANEEUWIJK, F. A. 1995. Analyzing Genotype by Environment Interaction in Dutch Potato Variety Trials Using Factorial Regression. *Euphytica,* 82, 149-155.

BENJAMINI, Y. & HOCHBERG, Y. 1995. Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B-Methodological,* 57, 289-300.

BERGAMO, G. C., DIAS, C. T. D. S. & KRZANOWSKI, W. J. 2008. Distribution-free multiple imputation in an interaction matrix though singular value decomposition. *Scientia Agricola,* 65, 422-427.

BERTIN, N., MARTRE, P., GENARD, M., QUILOT, B. & SALON, C. 2010. Under what circumstances can process-based simulation models link genotype to phenotype for complex traits? Case-study of fruit and grain quality traits. *Journal of Experimental Botany,* 61, 955-967.

BOER, M. P., WRIGHT, D., FENG, L. Z., PODLICH, D. W., LUO, L., COOPER, M. & VAN EEUWIJK, F. A. 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.

BRADU, D. & GABRIEL, K. R. 1978. Biplot as a Diagnostic Tool for Models of 2-Way Tables. *Technometrics,* 20, 47-68.

BRANCOURT-HULMEL, M., BIARNÈS-DUMOULIN, V. & DENIS, J. B. 1997. Points de repère dans l'analyse de la stabilité et de l'interaction génotype-milieu en amélioration des plants. *Agronomie,* 17, 219-246.

BROMAN, K. W. & SEN, S. 2009. *A Guide to QTL Mapping with R/qtl,* New York, Springer-Verlag.

CALINSKI, T., CZAJKA, S., DENIS, J. B. & KACZMAREK, Z. 1992. EM and ALS algorithms applied to estimation of missing data in series of variety trials. *Biuletyn Oceny Odmian,* 24-25, 9-31.

CHENU, K., CHAPMAN, S. C., HAMMER, G. L., MCLEAN, G., SALAH, H. B. H. & TARDIEU, F. 2008. Short-term responses of leaf growth rate to water deficit scale up to whole-plant and crop levels: an integrated modelling approach in maize. *Plant Cell and Environment,* 31, 378-391.

CHENU, K., CHAPMAN, S. C., TARDIEU, F., MCLEAN, G., WELCKER, C. & HAMMER, G. L. 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.

CHURCHILL, G. A. & DOERGE, R. W. 1994. Empirical Threshold Values for Quantitative Trait Mapping. *Genetics,* 138, 963-971.

CLARKE, J. M. & TOWNLEYSMITH, T. F. 1986. Heritability and Relationship to Yield of Excised-Leaf Water-Retention in Durum-Wheat. *Crop Science,* 26, 289-292.

COOPER, M., VAN EEUWIJK, F. A., HAMMER, G. L., PODLICH, D. W. & MESSINA, C. 2009. Modeling QTL for complex traits: detection and context for plant breeding. *Current Opinion in Plant Biology,* 12, 231-240.

CORNELIUS, P. L. 1993. Statistical Tests and Retention of Terms in the Additive Main Effects and Multiplicative Interaction-Model for Cultivar Trials. *Crop Science,* 33, 1186-1193.

CORNELIUS, P. L., SEYEDSADR, M. & CROSSA, J. 1992. Using the Shifted Multiplicative Model to Search for Separability in Crop Cultivar Trials. *Theoretical and Applied Genetics,* 84, 161-172.

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. & ARENAS, D. 2005. Using linear-bilinear models for studying gene expression x teatment interaction in microarray experiments. *Journal of Agricultural Biological and Environmental Statistics,* 10, 337-353.

CROSSA, J., CORNELIUS, P. L. & YAN, W. K. 2002. Biplots of linear-bilinear models for studying crossover genotype x environment interaction. *Crop Science,* 42, 619-633.

CROSSA, J., FOX, P. N., PFEIFFER, W. H., RAJARAM, S. & GAUCH, H. G. 1991. AMMI Adjustment for Statistical-Analysis of an International Wheat Yield Trial. *Theoretical and Applied Genetics,* 81, 27-37.

CROSSA, J., GAUCH, H. G. & ZOBEL, R. W. 1990. Additive Main Effects and Multiplicative Interaction Analysis of 2 International Maize Cultivar Trials. *Crop Science,* 30, 493-500.

CULMAN, S. W., BUKOWSKI, R., GAUCH, H. G., CADILLO-QUIROZ, H. & BUCKLEY, D. H. 2009. T-REX: software for the processing and analysis of T-RFLP data. *BMC Bioinformatics,* 10.

CULMAN, S. W., GAUCH, H. G., BLACKWOOD, C. B. & THIES, J. E. 2008. Analysis of T-RFLP data using analysis of variance and ordination methods: A comparative study. *Journal of Microbiological Methods,* 75, 55-63.

DA COSTA, J. F. P., ALONSO, H. & ROQUE, L. 2011. A Weighted Principal Component Analysis and Its Application to Gene Expression Data. *Ieee-Acm Transactions on Computational Biology and Bioinformatics,* 8, 246-252.

DE SWART, E. A. M., MARCELIS, L. F. M. & VOORRIPS, R. E. 2006. Variation in relative growth rate and growth traits in wild and cultivated Capsicum accessions grown under different temperatures. *Journal of Horticultural Science & Biotechnology,* 81, 1029-1037.

DENIS, J. B. 1988. Two-way analysis using covariables. *Statistics,* 19, 123-132.

DENIS, J. B. & BARIL, C. P. 1992. Sophisticated models with numerous missing values: The multiplicative interaction model as an example. *Biuletyn Oceny Odmian,* 24-25, 33-45.

DIGBY, P. G. N. 1979. Modified Joint Regression-Analysis for Incomplete Variety X Environment Data. *Journal of Agricultural Science,* 93, 81-86.

DO REGO, E. R., DO REGO, M. M., CRUZ, C. D., FINGER, F. L. & CASALI, V. W. D. 2011. Phenotypic diversity, correlation and importance of variables for fruit quality and yield traits in Brazilian peppers (Capsicum baccatum). *Genetic Resources and Crop Evolution,* 58, 909-918.

EBDON, J. S. & GAUCH, H. G. 2011. Direct Validation of AMMI Predictions in Turfgrass Trials. *Crop Science,* 51, 862-869.

EBERHART, S. A. & RUSSELL, W. A. 1966. Stability Parameters for Comparing Varieties. *Crop Science,* 6, 36-40.

EFRON, B. & GONG, G. 1983. A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation. *American Statistician,* 37, 36-48.

EMEBIRI, L. C. & MOODY, D. B. 2006. Heritable basis for some genotype-environment stability statistics: Inferences from QTL analysis of heading date in two-rowed barley. *Field Crops Research,* 96, 243-251.

FINLAY, K. W. & WILKINSON, G. N. 1963. Analysis of Adaptation in a Plant-Breeding Programme. *Australian Journal of Agricultural Research,* 14, 742-754.

FISHER, R. A. & MACKENZIE, W. A. 1923. Studies in crop variation. II. The manurial response of different potato varieties. *The Journal of Agricultural Science,* 13, 311-320.

FREEMAN, G. H. 1973. Statistical-Methods for Analysis of Genotype-Environment Interactions. *Heredity,* 31, 339-354.

GABRIEL, K. R. 1971. Biplot Graphic Display of Matrices with Application to Principal Component Analysis. *Biometrika,* 58, 453-467.

GABRIEL, K. R. & ZAMIR, S. 1979. Lower Rank Approximation of Matrices by Least-Squares with Any Choice of Weights. *Technometrics,* 21, 489-498.

GALWEY, N. 2006. *Introduction to mixed modelling : beyond regression and analysis of variance,* Chichester, England ; Hoboken, NJ, Wiley.

GAUCH, H. G. 1988. Model Selection and Validation for Yield Trials with Interaction. *Biometrics,* 44, 705-715.

GAUCH, H. G. 1992. *Statistical analysis of regional yield trials: AMMI analysis of factorial designs,* Amsterdam, Elsevier.

GAUCH, H. G. 2006. Winning the accuracy game - Three statistical strategies - replicating, blocking and modeling - can help scientists improve accuracy and accelerate progress. *American Scientist,* 94, 133-141.

GAUCH, H. G. 2002. *Scientific method in practice,* Cambridge, Cambridge University Press.

GAUCH, H. G. 2007. MATMODEL version 3.0: Open source software for AMMI and related analyses.

GAUCH, H. G. & FURNAS, R. E. 1991. Statistical-Analysis of Yield Trials with Matmodel. *Agronomy Journal,* 83, 916-920.

GAUCH, H. G., PIEPHO, H. P. & 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. & SORRELLS, M. 2011. Two New Strategies for Detecting and Understanding QTL x Environment Interactions. *Crop Science,* 51, 96-113.

GAUCH, H. G. & ZOBEL, R. W. 1990. Imputing Missing Yield Trial Data. *Theoretical and Applied Genetics,* 79, 753-761.

GAUCH, H. G. & ZOBEL, R. W. 1997. Identifying mega-environments and targeting genotypes. *Crop Science,* 37, 311-326.

GELDER, A. D., RAAPHORST, M., HOON, M. D. & BREUGEM, F. 2007. *Paprikateelt in de gesloten kas : resultaten bij Themato in 2006,* Naaldwijk, Wageningen UR.

GOLLOB, H. F. 1968. A Statistical Model Which Combines Features of Factor Analysis and Analysis of Variance Techniques. *Psychometrika,* 33, 73-115.

GOUDRIAAN, J. & LAAR, H. H. V. 1994. *Modelling potential crop growth processes : textbook with exercises,* Dordrecht etc., Kluwer.

GUSMÃO, L. 1985. An Adequate Design for Regression-Analysis of Yield Trials. *Theoretical and Applied Genetics,* 71, 314-319.

HAYES, P. M., CHEN, F. Q., KLEINHOFS, A., KILIAN, A. & MATHER, D. E. 1996. Barley genome mapping and its applications. *In:* JAUHAR, P. P. (ed.) *Method of Genome Analysis in Plants.* Boca Raton, Florida: CRC press.

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. & 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.

HEFFNER, E. L., SORRELLS, M. E. & JANNINK, J. L. 2009. Genomic Selection for Crop Improvement. *Crop Science,* 49, 1-12.

HEUVELINK, E. 1995. Growth, Development and Yield of a Tomato Crop - Periodic Destructive Measurements in a Greenhouse. *Scientia Horticulturae,* 61, 77-99.

HILL, M. O., BUNCE, R. G. H. & SHAW, M. W. 1975. Indicator Species Analysis, a Divisive Polythetic Method of Classification, and Its Application to a Survey of Native Pinewoods in Scotland. *Journal of Ecology,* 63, 597-613.

HORN, J. 1965. A rationale and test for the number of factors in factor analysis. *Psychometrika,* 30, 179-185.

ISHII, T., HAYASHI, T. & YONEZAWA, K. 2010. Categorization of Quantitative Trait Loci by Their Functional Roles: QTL Analysis for Chemical Concentration in Seed Grains. *Crop Science,* 50, 784-793.

JIANG, C. J. & ZENG, Z. B. 1995. Multiple-Trait Analysis of Genetic-Mapping for Quantitative Trait Loci. *Genetics,* 140, 1111-1127.

KANG, M. S. & GAUCH, H. G. 1996. *Genotype -by- Environment Interaction,* Boca Raton, CRC Press.

KOROL, A. B., RONIN, Y. I. & NEVO, E. 1998. Approximate analysis of QTL-environment interaction with no limits on the number of environments. *Genetics,* 148, 2015-2028.

KRZANOWSKI, W. J. 1987. Cross-Validation in Principal Component Analysis. *Biometrics,* 43, 575-584.

LACAZE, X., HAYES, P. M. & KOROL, A. 2009. Genetics of phenotypic plasticity: QTL analysis in barley, Hordeum vulgare. *Heredity,* 102, 163-173.

LARSON, S. R., KADYRZHANOVA, D., MCDONALD, C., SORRELLS, M. & BLAKE, T. K. 1996. Evaluation of barley chromosome-3 yield QTLs in a backcross of F2 population using STS-PCR. *Theoretical and Applied Genetics,* 93, 618-625.

LETORT, V., MAHE, P., COURNEDE, P. H., DE REFFYE, P. & COURTOIS, B. 2008. Quantitative genetics and functional-structural plant growth models: Simulation of quantitative trait loci detection for model parameters and application to potential yield optimization. *Annals of Botany,* 101, 1243-1254.

MACKAY, D. J. C. 1992. Bayesian Interpolation. *Neural Computation,* 4, 415-447.

MALOSETTI, M., RIBAUT, J. M. & VAN EEUWIJK, F. A. 2010. The analysis of multi-environment data: modeling genotype by environment and QTL by environment interaction. *In:* MONNEVEUX, P. & RIBAUT, J. M. (eds.) *Drought phenotyping in crops: from theory to practice.*

MALOSETTI, M., RIBAUT, J. M., VARGAS, M., CROSSA, J. & VAN EEUWIJK, F. A. 2008. A multi-trait multi-environment QTL mixed model with an application to drought and nitrogen stress trials in maize (Zea mays L.). *Euphytica,* 161, 241-257.

MALOSETTI, M., VOLTAS, J., ROMAGOSA, I., ULLRICH, S. E. & VAN EEUWIJK, F. A. 2004. Mixed models including environmental covariables for studying QTL by environment interaction. *Euphytica,* 137, 139-145.

MANDEL, J. 1969. Partitioning of Interaction in Analysis of Variance. *Journal of Research of the National Bureau of Standards Section B-Mathematical Sciences,* B 73, 309-318.

MANDEL, J. 1971. New Analysis of Variance Model for Non-Additive Data. *Technometrics,* 13, 1-18.

MANTON, J. H., MAHONY, R. & HUA, Y. B. 2003. The geometry of weighted low-rank approximations. *Ieee Transactions on Signal Processing,* 51, 500-514.

MARCELIS, L. F. M., ELINGS, A., DIELEMAN, J. A., BRAJEUL, E., BAKKER, M. J. & HEUVELINK, E. 2006. Modelling dry matter production and partitioning in sweet pepper. *Acta Horticulturae* 718, 121-128.

MARCELIS, L. F. M., HEUVELINK, E. & GOUDRIAAN, J. 1998. Modelling biomass production and yield of horticultural crops: a review. *Scientia Horticulturae,* 74, 83-111.

MEXIA, J. T., AMARO, A. P., GUSMAO, L. & BAETA, J. 1997. Upper contour of a Joint Regression Analysis. *Journal of Genetics and Breeding,* 51, 253-255.

MEXIA, J. T., PEREIRA, D. G. & BAETA, J. 1999. L2 environmental indexes. *Biometrical Letters,* 36, 137-143.

MOOERS, C. A. 1921. The agronomic placement of varieties. *Journal of the American Society of Agronomy,* 13, 337-352.

MUNKVOLD, J. D., TANAKA, J., BENSCHER, D. & SORRELLS, M. 2009. Mapping quantitative trait loci for preharvest sprouting resistance in white wheat. *Theoretical and Applied Genetics,* 119, 1223-1235.

NEDERHOFF, E. M. 1994. *Effects of CO2 concentrations on photosynthesis, transpiration, and production of greenhouse fruit vegetable crops*. PhD, Wageningen University.

PADEREWSKI, J., GAUCH, H. G., MADRY, W., DRZAZGA, T. & RODRIGUES, P. C. 2011. Yield Response of Winter Wheat to Spatial Conditions Using AMMI and Cluster Analysis. *Crop Science,* 51, 969-980.

PATTERSON, H. D. & THOMPSON, R. 1971. Recovery of Inter-Block Information When Block Sizes Are Unequal. *Biometrika,* 58, 545-553.

PAYNE, R. W., MURRAY, D. A., HARDING, S. A., BAIRD, D. B. & SOUTAR, D. M. 2011. *An Introduction to GenStat for Windows (14th Edition)*, VSN International, Hemel Hempstead, UK.

PEIGHAMBARI, S. A., SAMADI, B. Y., NABIPOUR, A., CHARMET, G. & SARRAFI, A. 2005. QTL analysis for agronomic traits in a barley doubled haploids population grown in Iran. *Plant Science,* 169, 1008-1013.

PEREIRA, D. G. & MEXIA, J. T. 2008. Selection proposal of cultivars of spring barley in the years from 2001 to 2004, using Joint Regression Analysis. *Plant Breeding,* 127, 452-458.

PEREIRA, D. G. & MEXIA, J. T. 2010. Comparing double minimization and zigzag algorithms in Joint Regression Analysis: the complete case. *Journal of Statistical Computation and Simulation,* 80, 133-141.

PEREIRA, D. G., MEXIA, J. T. & RODRIGUES, P. C. 2007. Robustness of Joint Regression Analysis. *Biometrical Letters,* 44, 105-128.

PIEPHO, H. P. 1994. Best Linear Unbiased Prediction (Blup) for Regional Yield Trials - a Comparison to Additive Main Effects and Multiplicative Interaction (Ammi) Analysis. *Theoretical and Applied Genetics,* 89, 647-654.

PIEPHO, H. P. 1995. Robustness of statistical tests for multiplicative terms in the additive main effects and multiplicative interaction model for cultivar trials. *Theoretical and Applied Genetics,* 90, 438-443.

PIEPHO, H. P. 1997. Analyzing genotype-environment data by mixed models with multiplicative terms. *Biometrics,* 53, 761-766.

PIEPHO, H. P. 2000. A mixed-model approach to mapping quantitative trait loci in barley on the basis of multiple environment data. *Genetics,* 156, 2043-2050.

QUILOT, B., GENARD, M., LESCOURRET, F. & KERVELLA, J. 2005. Simulating genotypic variation of fruit quality in an advanced peach x Prunus davidiana cross. *Journal of Experimental Botany,* 56, 3071-3081.

R DEVELOPMENT CORE TEAM. 2009. R: A Language and Environment for Statistical Computing. Vienna, Austria.

REIF, J. C., KUSTERER, B., PIEPHO, H. P., MEYER, R. C., ALTMANN, T., SCHON, C. C. & MELCHINGER, A. E. 2009. Unraveling Epistasis With Triple Testcross Progenies of Near-Isogenic Lines. *Genetics,* 181, 247-257.

REVELLE, W. & ROCKLIN, T. 1979. Very Simple Structure: an Alternative Procedure for Estimating the Optimal Number of Interpretable Factors. *Multivariate Behavioral Research,* 14, 403-414.

REYMOND, M., MULLER, B., LEONARDI, A., CHARCOSSET, A. & TARDIEU, F. 2003. Combining quantitative trait loci analysis and an ecophysiological model to analyze the genetic variability of the responses of maize leaf growth to temperature and water deficit. *Plant Physiology,* 131, 664-675.

REYMOND, M., MULLER, B. & TARDIEU, F. 2004. Dealing with the genotypexenvironment interaction via a modelling approach: a comparison of QTLs of maize leaf length or width with QTLs of model parameters. *Journal of Experimental Botany,* 55, 2461-2472.

RIJSDIJK, A. A. & HOUTER, G. 1993. Validation of a model for energy consumption, CO2 consumption and crop production (epc-model). *Acta Agriculturae,* 328, 125-131.

RODRIGUES, P. C., HEUVELINK, E., BINK, M. C. A. M., MARCELIS, L. F. M. & VAN EEUWIJK, F. A. 2012a. A complex trait with unstable QTLs can follow from component traits with stable QTLs: an illustration by a simulation study in pepper. *(to be submited).*

RODRIGUES, P. C., MALOSETTI, M., GAUCH, H. G. & VAN EEUWIJK, F. A. 2012b. Weighted AMMI to study genotype-by-environment interaction and QTL-by-environment interaction. *(to be submited).*

RODRIGUES, P. C., PEREIRA, D. G. & 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., HAN, F., ULLRICH, S. E., HAYES, P. M. & WESENBERG, D. M. 1999. Verification of yield QTL through realized molecular marker-assisted selection responses in a barley cross. *Molecular Breeding,* 5, 143-152.

ROMAGOSA, I., ULLRICH, S. E., HAN, F. & 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. & THOMAS, W. T. B. 2009. Statistical analyses of genotype by environment data. *In:* CARENA, M. J. (ed.) *Cereals*. New York: Springer.

SAEED, A., HAYAT, K., KHAN, A. A., IQBAL, S. & ABAS, G. 2007. Assessment of Genetic Variability and Heritability in Lycopersicon esculentum Mill. *International Journal of Agriculture and Biology,* 9, 375-377.

SCHEFFÉ, H. 1959. *The analysis of variance,* New York,, Wiley.

SEARLE, S. R. 1971. *Linear models,* New York,, Wiley.

SEARLE, S. R., CASELLA, G. & MCCULLOCH, C. E. 1992. *Variance components,* New York, Wiley.

SEBER, G. A. F. & LEE, A. J. 2003. *Linear regression analysis,* Hoboken, N.J., Wiley-Interscience.

SETIMELA, P. S., VIVEK, B., BANZIGER, M., CROSSA, J. & MAIDENI, F. 2007. Evaluation of early to medium maturing open pollinated maize varieties in SADC region using GGE biplot based on the SREG model. *Field Crops Research,* 103, 161-169.

SHUKLA, G. K. 1972. Some Statistical Aspects of Partitioning Genotype Environmental Components of Variability. *Heredity,* 29, 237-&.

SOOD, S., SOOD, R., SAGAR, V. & SHARMA, K. C. 2009. Genetic Variation and Association Analysis for Fruit Yield, Agronomic and Quality Characters in Bell Pepper. *International Journal of Vegetable Science,* 15, 272-284.

SPITTERS, C. J. T. 1990. Crop growth models: their usefulness and limitations. *Acta Horticulture,* 267, 349-368.

SPITTERS, C. J. T. & SCHAPENDONK, A. H. C. M. 1990. Evaluation of Breeding Strategies for Drought Tolerance in Potato by Means of Crop Growth Simulation. *Plant and Soil,* 123, 193-203.

SREBRO, N. & JAAKKOLA, T. Year. Weighted Low-Rank Approximations. *In:* FAWCETT, T. & MISHRA, N., eds. Twentieth International Conference on Machine Learning (ICML-2003), 2003 Washington DC. The AAAI Press, Menlo Park, California, 600-607.

TARDIEU, F. 2003. Virtual plants: modelling as a tool for the genomics of tolerance to water deficit. *Trends in Plant Science,* 8, 9-14.

VAN EEUWIJK, F. A. 1995. Linear and Bilinear Models for the Analysis of Multi-Environment Trials .1. An Inventory of Models. *Euphytica,* 84, 1-7.

VAN EEUWIJK, F. A., BINK, M. C. A. M., CHENU, K. & 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., DENIS, J. B. & KANG, M. S. 1996. Incorporating additional information on genotypes and environments in models for two-way genotype by environment tables. *In:* KANG, M. S. & GAUCH, H. G. (eds.) *Genotype by Environment Interaction: New Perspectives*. Boca Raton: CRC Press.

VAN EEUWIJK, F. A., MALOSETTI, M., YIN, X. Y., STRUIK, P. C. & STAM, P. 2005. Statistical models for genotype by environment data: from conventional ANOVA models to eco-physiological QTL models. *Australian Journal of Agricultural Research,* 56, 883-894.

VAN ITTERSUM, M. K., LEFFELAAR, P. A., VAN KEULEN, H., KROPFF, M. J., BASTIAANS, L. & GOUDRIAAN, J. 2003. On approaches and applications of the Wageningen crop models. *European Journal of Agronomy,* 18, 201-234.

VELICER, W. 1976. Determining the number of components from the matrix of partial correlations. *Psychometrika,* 41, 321-327.

VERBEKE, G. & MOLENBERGHS, G. 2009. *Linear mixed models for longitudinal data,* New York, Springer.

VOLTAS, J., VAN EEUWIJK, F., IGARTUA, E., GARCÍA DEL MORAL, L. F., MOLINA-CANO, J. L. & ROMAGOSA, I. 2002. Genotype by environment interaction and adaptation in barley breeding: Basic concepts and methods of analysis. *In:* SLAFER, G. A., MOLINA-CANO, J. L., SAVIN, R., ARAUS, J. L. & ROMAGOSA, I. (eds.) *Barley science: Recent advances from molecular biology to agronomy of yield and quality*. New York: Haworth Press.

WANG, S., BASTEN, C. J. & ZENG, Z.-B. 2007. Windows QTL Cartographer 2.5. *Department of Statistics, North Carolina State University*. Raleigh, NC.

WENTZELL, P. D., ANDREWS, D. T., HAMILTON, D. C., FABER, K. & KOWALSKI, B. R. 1997. Maximum likelihood principal component analysis. *Journal of Chemometrics,* 11, 339-366.

WILLIAMS, E. J. 1952. The Interpretation of Interactions in Factorial Experiments. *Biometrika,* 39, 65-81.

WUBS, A. M., HEUVELINK, E. & MARCELIS, L. F. M. 2009. Abortion of reproductive organs in sweet pepper (Capsicum annuum L.): a review. *Journal of Horticultural Science & Biotechnology,* 84, 467-475.

WUBS, A. M., HEUVELINK, E., MARCELIS, L. F. M. & HEMERIK, L. 2010. Survival analysis as a tool to quantify effects of factors on abortion rates of reproductive organs. *(Submitted).*

YAN, W. & KANG, M. S. 2002. *GGE biplot analysis: A graphical tool for breeders, geneticists, and agronomists,* Boca Raton, Florida, CRC Press.

YAN, W. & KANG, M. S. 2003. GGE biplot analysis : a graphical tool for breeders, geneticists, and agronomists, Boca Raton, Fla., CRC Press.

YANG, R. C., CROSSA, J., CORNELIUS, P. L. & BURGUENO, J. 2009. Biplot Analysis of Genotype x Environment Interaction: Proceed with Caution. Crop Science, 49, 1564-1576.

YATES, F. & COCHRAN, W. G. 1938. The analysis of groups of experiments. The Journal of Agricultural Science, 28, 556-580.

YIN, X. Y., CHASALOW, S. D., DOURLEIJN, C. J., STAM, P. & KROPFF, M. J. 2000. Coupling estimated effects of QTLs for physiological traits to a crop growth model: predicting yield variation among recombinant inbred lines in barley. Heredity, 85, 539-549.

YIN, X. Y., STRUIK, P. C. & KROPFF, M. J. 2004. Role of crop physiology in predicting gene-to-phenotype relationships. Trends in Plant Science, 9, 426-432.

YIN, X. Y., STRUIK, P. C., VAN EEUWIJK, F. A., STAM, P. & TANG, J. J. 2005. QTL analysis and QTL-based prediction of flowering phenology in recombinant inbred lines of barley. Journal of Experimental Botany, 56, 967-976.

ZENG, Z. B. 1994. Precision Mapping of Quantitative Trait Loci. Genetics, 136, 1457-1468.

ZHANG, M., MONTOOTH, K. L., WELLS, M. T., CLARK, A. G. & ZHANG, D. B. 2005. Mapping multiple quantitative trait loci by Bayesian classification. Genetics, 169, 2305-2318.

ZHENG, B. S., LE GOUIS, J., DANIEL, D. & BRANCOURT-HULMEL, M. 2009. Optimal numbers of environments to assess slopes of joint regression for grain yield, grain protein yield and grain protein concentration under nitrogen constraint in winter wheat. Field Crops Research, 113, 187-196.

ZHU, H., BRICENO, G., DOVEL, R., HAYES, P. M., LIU, B. H., LIU, C. T. & ULLRICH, S. E. 1999. Molecular breeding for grain yield in barley: an evaluation of QTL effects in a spring barley cross. Theoretical and Applied Genetics, 98, 772-779.

ZOBEL, R. W., WRIGHT, M. J. & GAUCH, H. G. 1988. Statistical-Analysis of a Yield Trial. Agronomy Journal, 80, 388-393.