Dealing with Sparsity in Genotype×Environment Analyses

A thesis presented in partial
fulfilment of the requirements
for the degree of
Doctor of Philosophy
in Statistics

at Massey University,
Palmerston North, New Zealand.

A. Jonathan R. Godfrey
2004

Massey University
COLLEGE OF SCIENCES

INSTITUTE OF INFORMATION SCIENCES & TECHNOLOGY
Private Bag 11 222
Palmerston North
New Zealand
T 64 6 356 9099
F 64 6 350 5750
www.massey.ac.nz
www-ist.massey.ac.nz

Integrated research and teaching in the fields of
- Statistics
- Computer Science
- Electronics & InfoComm Engineering

SUPERVISOR'S DECLARATION

This is to certify that the research carried out for the Doctoral thesis entitled "Dealing with Sparsity in Genotype-by-Environment Analyses" was done by Anthony Jonathan Royce Godfrey in the Institute of Information Sciences and Technology, Massey University, Palmerston North, New Zealand. The thesis material has not been used in part or in whole for any other qualification, and I confirm that the candidate has pursued the course of study in accordance with the requirements of the Massey University regulations.

Supervisor's Name: Associate Professor Chin-Diew Lai

Signature: C. D. Lai

Date: 2-8-04

Massey University
COLLEGE OF SCIENCES

INSTITUTE OF INFORMATION SCIENCES & TECHNOLOGY
Private Bag 11 222
Palmerston North
New Zealand
T 64 6 356 9099
F 64 6 350 5750
www.massey.ac.nz
www-ist.massey.ac.nz

Integrated research and teaching in the fields of
- Statistics
- Computer Science
- Electronics & InfoComm Engineering

CERTIFICATE OF REGULATORY COMPLIANCE

This is to certify that the research carried out in the Doctoral Thesis entitled **"Dealing with Sparsity in Genotype-by-Environment Analyses"** in the Institute of Information Sciences and Technology, Massey University, New Zealand:

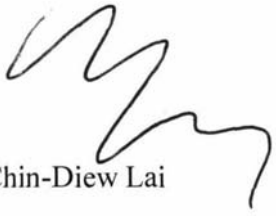
(a) is the original work of the candidate, except as indicated by appropriate attribution in the text and/or in the acknowledgements;

(b) that the text, excluding appendices/annexes, does not exceed 100,000 words;

(c) all the ethical requirements applicable to this study have been complied with as required by Massey University, other organisations and/or committees which had particular association with this study, and relevant legislation.

Please insert Ethical Authorisation code(s) here: **N.A.**

Candidate's Name: Anthony Jonathan Royce Godfrey

Signature:  **Date:**

Supervisor's Name: Chin-Diew Lai

Signature: C. D. Lai **Date:** 2/8/04

Massey University
COLLEGE OF SCIENCES

INSTITUTE OF INFORMATION SCIENCES & TECHNOLOGY
Private Bag 11 222
Palmerston North
New Zealand
T 64 6 356 9099
F 64 6 350 5750
www.massey.ac.nz
www-ist.massey.ac.nz

Integrated research and teaching in the fields of
- Statistics
- Computer Science
- Electronics & InfoComm Engineering

CANDIDATE'S DECLARATION

This is to certify that the research carried out for my Doctoral thesis entitled "*Dealing with Sparsity in Genotype-by-Environment Analyses*" in the Institute of Information Sciences and Technology, Massey University, Palmerston North, New Zealand, is my own work and that the thesis material has not been used in part or in whole for any other qualification.

Candidate's Name Anthony Jonathan Godfrey

Signature

Date 02/08/2004

Abstract

Researchers are frequently faced with the problem of analyzing incomplete and often unbalanced genotype-by-environment (G×E) matrices which arise as a trials programme progresses over seasons. The principal data for this investigation, arising from a ten year programme of onion trials, has less than 2,300 of the 49,200 combinations from the 400 genotypes and 123 environments. This 'sparsity' renders standard G×E methodology inapplicable. Analysis of this data to identify onion varieties that suit the shorter, hotter days of tropical and subtropical locations therefore presented a unique challenge.

Removal of some data to form a complete G×E matrix wastes information and is consequently undesirable. An incomplete G×E matrix can be analyzed using the additive main effects and multiplicative interaction (AMMI) model in conjunction with the EM algorithm but proved unsatisfactory in this instance.

Cluster analysis has been commonly used in G×E analyses, but current methods are inadequate when the data matrix is incomplete. If clustering is to be applied to incomplete data sets, one of two routes needs to be taken: either the clustering procedure must be modified to handle the missing data, or the missing entries must be imputed so that standard cluster analysis can be performed.

A new clustering method capable of handling incomplete data has been developed. 'Two-stage clustering', as it has been named, relies on a partitioning of squared Euclidean distance into two independent components, the G×E interaction and the genotype main effect. These components are used in the first and second stages of clustering respectively.

Two-stage clustering forms the basis for imputing missing values in a G×E matrix, so that a more complete data array is available for other G×E analyses. 'Two-stage imputation' estimates unobserved G×E yields using inter-genotype similarities to adjust observed yield data in the environment in which the yield is missing. This new imputation method is transferrable to any two-way data situation where all observations are measured on the same scale and the two factors are expected to have significant interaction. This simple, but effective, imputation method is shown to improve on an existing method that confounds the G×E interaction and the genotype main effect. Future development of two-stage imputation will use a parameterization of two-stage clustering in a multiple imputation process.

Varieties recommended for use in a certain environment would normally be chosen using results from similar environments. Differing cluster analysis approaches were applied, but led to inconsistent environment clusterings. A graphical summary tool, created to ease the difficulty in identifying the differences between pairs of clusterings, proved especially useful when the number of clusters and clustered observations were high. 'Cluster influence

diagrams' were also used to investigate the effects the new imputation method had on the qualitative structure of the data.

A consequence of the principal data's sparsity was that imputed values were found to be dependent on the existence of observable inter-genotype relationships, rather than the strength of these observable relationships. As a result of this investigation, practical recommendations are provided for limiting the detrimental effects of sparsity. Applying these recommendations will enhance the future ability of two-stage imputation to identify those onion varieties that suit tropical and subtropical locations.

Acknowledgements

I wish to acknowledge the efforts of many people that have culminated in the submission of this work. First, and foremost, let me thank my supervisors. Professor Graham Wood gave me the opportunity to start a PhD. I will remain eternally grateful for this, the endless supply of advice, kind words, and the occasional kick in the pants that all students need to achieve the best they can. Thanks to Dr Mike Nichols and Dr Ganes Ganesalingam also for their contributions. Other Massey staff have also played a part; Greg Arnold for S-PLUS debugging and advice, Mark Bebbington and Geoff Jones for LaTeXcommands and debugging, and everyone that came to my seminars and provided feedback over the years.

This work would not have been considered if the real world problem had not existed. Dr Lesley Currah has contributed on numerous occasions with the knowledge surrounding the trials programme. Her patience must at times have worn thin, but it is hoped that she will relish the opportunity that now exists to send our findings to the many collaborators around the world. E-mail has served us well, and the time difference has allowed us each to send well considered questions and responses. Thank you Lesley for everything, including the opportunities for publications, your assistance with proof reading, and generally acting as another supervisor.

There are in fact very many people that need to be remembered for their contributions. I hope each of them receives some joy at the news of the submission of my thesis.

Special mention needs to be paid to those individuals that have assisted in the reading of many papers, document preparation, and all things visually dependent; Zoë Wood, Vince Pegg (for weekends and weeknights), Allister Campbell, Will Samuel, Judy Cann, Graeme Robinson, Michele Bisset, and others.

How can we forget those that ensured that I reached the start of the PhD, and then sustained me through the years. My parents and family have provided the breaks away from Palmerston North that revitalize the spirit; Mum for finishing her PhD first, so that she had time for proof reading. My many flatmates from 1998 to 2002 who put up with my occasional rantings and experimental cooking. The All Blacks and the Black Caps for their various interruptions. Radio Sport for commentaries and the slow paced progress of work over the summer months. The staff at the MUSA shop are yet to vanquish my appetite for apple and cinnamon muffins, let alone the raspberry twists.

Inanimate items serve frequently without thanks. Praise be to the Concise Oxford English Dictionary and various sources of LaTeXnotes available on the Internet. I have now learned more about the English language and how to present it than I expected at the outset. I thought my command of the language and sentence construction was better than the constant proofing showed me. Sorry to the purists, but I boldly went, and here is the fruit of my labour.

Contents

List of Figures

List of Tables

Part I

Background to the Problem

Chapter 1

Setting the scene

1.1 Introduction

Daunting is the word that best describes the prospect of working with an extremely sparse data set of yields from an international programme of onion trials. More than 95% of the combinations for 400 genotypes and 123 trials, which form the principal data for this investigation, were never tested. Programme organizers still wish to know, however, which of the onion varieties should be grown in tropical and subtropical locations. This volume presents the most comprehensive attempt to date to analyze this large and complex set of data.

Collaborators in over fifty countries ran a total of 123 trials with limited time and physical resources. When the results of their efforts were combined into a genotype-by-environment (G×E) table less than 2,300 of the possible 49,200 combinations had been observed. This lack of data has been termed 'sparsity' throughout the investigation.

Making sense of the phenomena that led to the success of certain onion varieties in a wide range of tropical environments was not possible using existing methodology. The challenge of bringing order to the data was met by extending cluster analysis techniques to allow them to be used with incomplete G×E data. An imputation technique was also developed and applied to provide a complete G×E matrix for analysis. Ultimately the aim of this investigation is to be able to answer the principal research question:

> "Given a certain (possibly new) environment, which onion varieties are most likely to succeed in terms of their edible yield?"

This chapter provides an initial background to the investigation by first introducing some key definitions in Section 1.2. Section 1.3 then addresses some issues in G×E research which informed this investigation. This is followed in Section 1.4 by an outline of the path taken in seeking a solution to the principal research question, and includes the key findings of each chapter. The chapter concludes in Section 1.5 with a discussion of the significance of the theoretical and practical contributions arising from this investigation.

1.2 Some key definitions

This section presents some definitions of terms that are used throughout this volume. They are to be found in common use in the genotype-by-environment (G×E) literature, but may not be well known to a less agronomically minded reader.

The two principal terms of interest are 'genotype' and 'environment', but other terms also need to be mentioned here.

Genotype	Although the term has a formal definition to geneticists, the term 'genotype' refers to an identifiably specific subspecies. This variety, line, or cultivar is usually identified by a commercial name or code.
Environment	This term refers to a set of growing conditions under which genotypes are examined. These conditions are dependent on both the physical location and the timing of the trial, and include climatic, geographic and edaphic factors.
Edaphic	According to the Concise Oxford Dictionary, this term includes all factors that are produced or influenced by the soil.

When genotypes are being tested in a set of environments it is common for their differing genetic traits to cause them to have different responses to the environmental factors. Kang (1998) defined 'phenotype' as "physical appearance or discernible traits of an individual, which may be observable at a physical, morphological, anatomical, or biochemical level." The only phenotypic quality of interest in this investigation was genotype yield, and as such the term phenotype has not been required in this volume.

Observed differences in yield are reflected as either absolute or comparative advantages. An absolute advantage arises when one genotype outperforms another in all trials where they were both tested. A comparative advantage is gained when the margin of advantage to one genotype changes from environment to environment. This phenomenon is referred to as genotype-by-environment (G×E) interaction throughout this volume and the G×E literature.

1.3 Issues relevant to G×E research

In the substantial body of G×E literature, it was clear that G×E research has many differing aims which depend mainly on the focus of the particular project being discussed. In general, however, a simple statement from Kang (1998) summarizes the overall aim of G×E research:

> "The role of a crop improvement program is to develop high-yielding, profitable cultivars for sustainable production in target areas by managing genetic variability and generating new genetic combinations."

Differing priorities on some issues mentioned in the literature required a stand to be taken with respect to the nature of the principal data set. A more detailed history of the Onion Trials Programme is given in Chapter 3, but relevant aspects are mentioned below for each of the following issues:

1. Agronomy versus plant breeding.
2. A regional versus an international focus.
3. Attribute measurement and analysis.
4. The impact of missing data.

Agronomy versus plant breeding

While agronomists and plant breeders often work together, their perspectives differ. The major difference in approach taken is that a plant breeder attempts to improve the genotypes while an agronomist attempts to improve environments.

Plant breeders are interested in the overall improvement of genetic stock from which varieties are selected to create the next generation of varieties. These varieties are then tested and the process continues until the plant breeders are convinced they have attained the current goals. These goals will change over time, however, as disease resistance and other requirements change, and new target environments become available.

On the other hand, the agronomist perspective, chosen in this investigation, involves matching the current pool of varieties (while it may alter from one year to the next) to the best target environments. Introducing new varieties to an environment may highlight possible benefits of further testing each variety so that optimal growing conditions for that variety can be found in each environment.

Plant breeders rely on agronomists, who ensure that the best possible outcome is obtained from any particular variety. They are then able to use the performances of the varieties to decide how to breed a new variety, using heritability aspects including yield. Aspects of heritability are not considered in this investigation, due in part to the make-up of the principal data, but more importantly, because there is a need to get the best out of each variety before its value to the plant breeder is determined. This process of getting the best from each variety in the Onion Trials Programme remains unfinished.

A regional versus an international focus

For experimentation to be of value the target population needs to be known. This population in G×E work is the range of environments where a crop will be grown. The range of environmental conditions (whether they be climatic, geographic, or edaphic) that will be encountered in an international trials programme, will be larger than those found in a regional study. Most G×E work is done in regional studies, while some collaborations may

lead to the sharing of knowledge and resources. Many models presented in Chapter 2 are therefore aimed towards the regional experimentation that dominates the G×E literature. Only the relevant models and measures are considered for use with the principal data.

Successfully taking advantage of specific adaptation will require an understanding of the characteristics of the environments under test, so that extrapolated results are of greater use in the future. This creates a 'catch-22', however, as the best way to gain this understanding is through extensive testing. Chapter 10 further outlines this problem and presents some ideas that may lead to its possible solution. More specifically though, there is a need to be mindful that the principal data has a tropical and sub-tropical focus. When writing about environmental considerations for rice growing, Alagarswamy *et al.* (1996) noted:

> "The genotype by environment (G×E) interaction for grain yield is often high in semi-arid tropics (SAT). This is caused by the unreliability of rainfall and its temporal and spatial variations. ... Furthermore soils are poor and most important nutrients (nitrogen, phosphorus, and potassium) are often deficient. As a result, crop productivity in the SAT is often low and highly variable. These highly variable growing conditions require specific adaptation of genotypes, which can only be identified through extensive multi-environment testing."

Attribute measurement and analysis

Over recent years, many authors have turned to the results from multi-attribute G×E data. While the results from considering attributes other than yield may be of interest, the principal data does not have any other response data available. It has therefore been a simple decision to consider in greater depth the problems and opportunities presented by such data. Many methods mentioned in Chapter 2 are applicable to other data forms, even though they concentrate on the yield performance of the genotypes. The only caveat to consider here is that these methods are suited to continuous rather than discrete, or more specifically, binary data.

The impact of missing data

While many models presented in Chapter 2 are interesting in nature, or have their value to agronomists and statisticians alike, they are less relevant to this investigation. The missing data, let alone sparsity, in the data means that many models cannot be considered unless the data is somehow made complete. Once a complete set of data is obtained the appropriate method to analyse it can be chosen. The relevance and value to this investigation of previous contributions in the G×E literature were assessed using the four issues above. In particular, the impact missing data would have on existing methods

dominated the initial stages of the investigation. Attempts to allow for missing data when employing some of these methods is reported in Section 3.7, but in general proved ineffective. The next section outlines the steps taken to build on existing G×E research in order to seek an answer to the principal research question.

1.4 The path to a solution

In this section, the path taken in seeking an answer to the principal research question is described. In reporting the findings of this investigation the eleven chapters have been grouped in three parts: background to the problem, the development of a solution and the results and implications.

In Part I, this introductory chapter is followed by Chapter 2 which reviews the fundamentals of G×E modelling, and in particular:

1. Provides information on the methods used to recognize the significance of any existing G×E interaction.

2. Describes the joint regression model, its criticisms, and its extensions.

3. Introduces various multiplicative models.

4. Backgrounds the use of cluster analysis in G×E analyses.

5. Considers the use of stability measures.

6. Outlines various other models and methods.

The history of G×E interaction modelling is covered in this chapter, with particular reference to the situation where data are incomplete. It will be shown that there are comparatively few options for dealing with incomplete G×E data available in the literature.

Chapter 3 details the principal data set that inspired this investigation. An account of its history is provided, and initial consideration is given to the impact that missing data will have on the investigation. The incomplete data is analyzed using graphical summaries and the application of some basic models, before construction of the two data sets used throughout the remainder of the investigation is described. Attempts to apply modified versions of existing numerical summaries and the application of a modelling approach for incomplete data are presented in this chapter.

By the end of Chapter 3, it is apparent that existing G×E methodology will not assist in answering the principal research question. Part II describes the process developed to do this.

Chapters 4 through 8 include the new theoretical developments arising over the course of the investigation. The flow chart presented in Figure 1.1 shows the planned path for

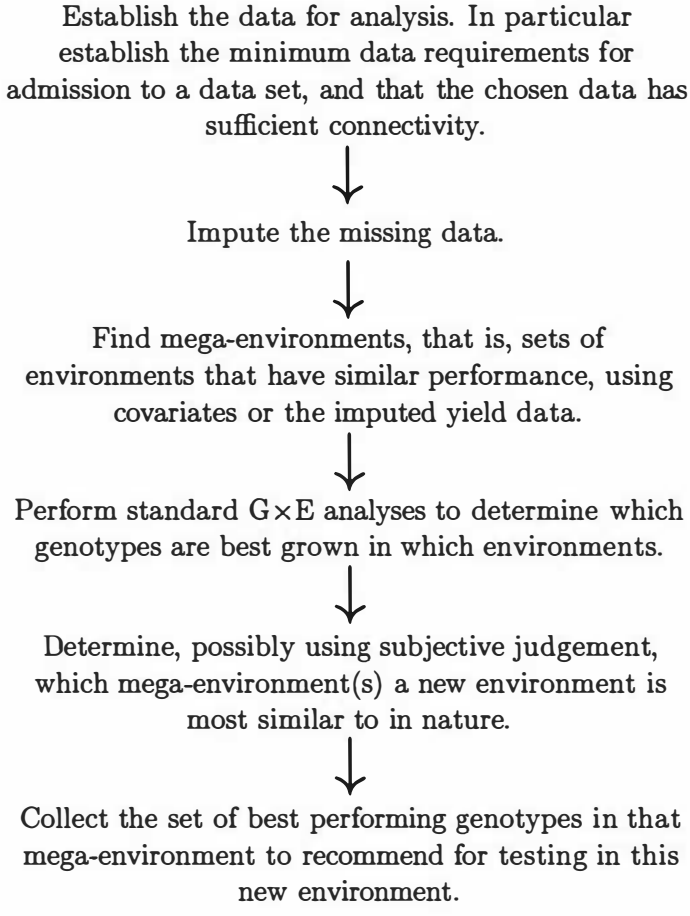


Figure 1.1: A flow chart depiction of the steps followed to find a solution to the principal research question, "Given a certain (possibly new) environment, which onion varieties are most likely to succeed in terms of their edible yield?"

answering the principal research question. The first step of this process is to identify the data to be worked with, presented in Section 3.6.

The second step of the process in Figure 1.1 is to impute the missing data. The new method for this is presented in Chapter 6, but the crucial building blocks for this imputation process are presented in Chapters 4 and 5.

Chapter 4 discusses the development of two new distance measures, appropriate for incomplete data, which compare genotypes from different perspectives. Important features of Chapter 4 include

1. The partition of Euclidean distance into two constituent parts, one based on the difference in a pair of genotypes' mean performances; and the second part based

on the difference in their specific adaptation performances (known as interaction profiles).

2. A method for estimating an upper bound for unobserved distances. This upper bound is subsequently used to adjust observed distances based on a low number of comparisons.

3. Justification for using a new distance measure in place of Euclidean distance as the method for comparing differences when no G×E interaction is present.

Chapter 5 shows how the distance measures developed in Chapter 4 can be used with incomplete data to graphically represent a large amount of information via cluster analysis. The important contribution of this chapter is the new clustering method which involves two clusterings; first using the similarity of G×E interaction or specific adaptation, and second, re-clustering each of these clusters to determine which have similar mean performance.

Chapter 6 then uses the information from clustering genotypes and the theoretical development of Chapter 4 to formulate an imputation method that incorporates the similarity of specific adaptation and the dissimilarity of mean performances. This new imputation method is tested using data from the literature. The important contributions of this chapter include the presentation of:

1. The new imputation method, including an example based on a complete data set that has been made incomplete by randomly deleting some observations.

2. Results from simulation testing using data from the G×E literature.

3. An initial investigation of the consistency of imputed values for the two data sets of the Onion Trials Programme.

The third step of the process in Figure 1.1 is to form 'mega-environments' by clustering environments, and is covered in Chapter 7. This chapter includes results from applying three different methods to create mega-environments. Eight different mega-environment sets were formed.

A consequence of the high number of clusterings in Chapter 7 was the need for a means of comparing the cluster memberships without time-consuming effort. A new graphical tool for comparing two cluster analyses is introduced in Chapter 8 and along with numerical summaries shows the consistency of the clusterings of Chapter 7. A major finding of Chapter 8 is that the data arising from the Onion Trials Programme is too sparse for the principal research question to be answered using the current data. Reasons behind this problem are discussed and an auxiliary research question is proposed.

The last three steps in the process in Figure 1.1 are discussed in Chapter 9, being the first chapter in Part III. The results from applying standard G×E methodology to the

data after imputation are given in this chapter, which includes a number of methods that could be employed for making variety selections for a fictitious new trial.

Lessons learned throughout this investigation for the betterment of the future organization of the Onion Trials Programme are presented in Chapter 10. Application of the recommendations offered in this chapter should enhance future trials programmes by ensuring that wastage of resources is minimized, imputations are improved, and the ultimate aim of answering the principal research question may be realized.

Chapter 11 concludes the investigation by drawing all the findings together in a cohesive summary. Some of the data arising from the Onion Trials Programme are provided in the appendix, while other data sets and electronic resources are to be found on the CD-ROM accompanying this volume.

1.5 Significance of this investigation

The significance of the developments arising from this investigation fall into two categories. In seeking to answer the principal research question, both theoretical and practical contributions towards the field of G×E research have been made.

In spite of a substantial body of G×E research, the problem of dealing with sparsity is not appropriately addressed by the literature. As mentioned in the previous section, a new imputation method was developed that will allow an incomplete G×E data set to be analyzed. This method is transferrable to any two-way data situation where all observations are measured on the same scale and the two factors are expected to have significant interaction.

This new imputation method arose from the development of a clustering method that can be used to graphically represent the similarity of a large number of genotype performances. The new clustering method uses two distance measures that have been adapted to handle incomplete data.

A graphical tool has also been proposed to give statisticians a means of comparing cluster analyses. To date, this had been done using numerical summaries and graphical representations of the summaries, rather than representing the memberships of the clusters.

The application of these theoretical developments will benefit agronomists, in that their efforts collecting data will not be wasted. Pragmatism demanded that data for some genotypes, too sparse to be included in the current analysis, be discarded. The remaining data could be used to form a complete data set to assist with the optimal selection of genotypes for specific environments. If recommendations offered in Chapter 10 for choosing genotypes to be tested in new trials are followed, the scope of the analysis can be widened by recovering previously discarded data. Inclusion of new data and previously discarded data should improve imputations and provide trials programme organizers with a better

data set for answering the principal research question.

1.6 Summary

This chapter introduced the background for this investigation, which sought to answer the principal research question:

> "Given a certain (possibly new) environment, which onion varieties are most likely to succeed in terms of their edible yield?"

The principal data that inspired this investigation arose from an international programme of onion trials. The process by which a solution was developed to meet the somewhat daunting challenge of analyzing a complex data set, with over 95% of possible genotype-by-environment combinations unobserved, was outlined in this chapter.

Chapter 2

Fundamentals of G×E analysis

2.1 Introduction

Creation of the foundations from which to build new methodology plays an important role in any investigation. This chapter reviews the history of G×E analyses, with special attention given to those methods applicable when data are incomplete. On the whole the methods presented here provide decision criteria to answer the fundamental question of G×E analyses: "Which varieties should be grown where?"

Multi-environment trials (METs) are conducted for crop improvement and selection to examine genotype-by-environment (G×E) interaction. The importance of interaction effects in G×E analyses has been well documented over the last thirty years. Review papers by Freeman (1973, 1985), Lin *et al.* (1986), Crossa (1990), Cooper and DeLacy (1994), Piepho (1994a), and Flores *et al.* (1998) identify many of the models and methods used in G×E analyses and explain their inter-relationships.

G×E research is made challenging by the existence of G×E interaction. Necessary conditions for its existence were given by Yan and Hunt (1998), namely:

> "A significant GE interaction for yield in a MET depends on three aspects: genotypes that are different enough; environments that are different enough; and genotypes that respond to the environments differentially. Genotypic and environmental differences are necessary for the existence of a significant GE interaction. If either or both are small, no significant GE interaction can be expected."

Preliminary investigation of the 400 genotypes and 123 environments of the principal data suggested that the first two conditions for the existence of G×E interaction will be met. Gaining an understanding of the ability of these genotypes to respond differentially to the environments required further analysis.

There are three options for dealing with G×E interaction. It can be ignored, avoided, or exploited (Kang, 1998). Ignoring a significant interaction effect can lead to the wrong

decisions being made as the main effects are incorrectly estimated. Avoiding interaction is not always an option in practical terms, but if sets of environments can be found that have little interaction with tested genotypes, the analysis can continue with little need to be concerned with the interaction. Once any G×E interaction has been determined significant, using techniques introduced in Section 2.2, the methods highlighted in subsequent sections can be used to exploit the interaction.

The joint regression model of Yates and Cochran (1938) was one of the first attempts to model the G×E interaction, and is discussed in Section 2.3. As will be noted, the model has been widely criticized but has been extensively used nonetheless. Variations of joint regression have been proposed and are reviewed in Section 2.4. The joint regression model is an example of a multiplicative model. Many other multiplicative models are examined in Section 2.5, notably including the additive main effects and multiplicative interaction (AMMI) model.

Cluster analysis techniques have been used in many ways in G×E research. Section 2.6 presents many of these, and should also be used as introductory material for the theoretical developments of Chapters 4 and 5. Stability analysis is concerned with the ability to reliably select genotypes for use in a given environment. Many of the stability parameters that have been proposed are reviewed in Section 2.7.

Other suggested methods, some quite novel, for analyzing G×E data are considered in Section 2.8. Throughout the chapter, the impact of missing data has been considered. Where appropriate, available methods for handling missing data are discussed, and in other instances, suggestions for adjustments are made.

There is a plethora of analyses which come under the G×E research umbrella. This was noted by Lin and Binns (1994) when they said:

> "The basic problem is that 'GE interaction' covers a wide range of objectives, each having several methods to attain it, and the concepts and assumptions underlying the methods are not the same."

Each of the review papers cited above has a particular perspective, as does this chapter. In reviewing the G×E literature, it is recognized that the selection of which contributions are the most significant for this investigation is ultimately subjective.

2.2 Establishing the significance of G×E interaction

The existence of G×E interaction is common in G×E analysis. It is, however, important to recognize when its existence is likely, and to establish its relative importance. This section outlines the tests for the existence of G×E interaction and the history behind the methods used to deal with existing G×E interaction so that comparisons can be made among tested genotypes and environments.

It would be pointless to continue with the current examination if interaction is an insignificant factor in the testing of genotypes over environments. From experience, agronomists expect G×E interaction as a matter of course. The testing that continues worldwide recognizes the importance of G×E interaction. An understanding of the nature of existing G×E interaction and its causes enables agronomists to improve the output of planting.

From a statistician's viewpoint, interaction is important because there is a risk in making the 'wrong' decision if the interaction is ignored. Yan and Hunt (1998) stated, "In a typical MET, the GE interaction effect is usually equal to or larger than the genotypic effect.". Modelling the interaction appropriately will assist in making the 'right' decision. Of course, statisticians base their decisions on probabilities and models. They are not therefore 100% correct in their decisions. The salient point here is that both agronomists and statisticians wish to make the best decision possible at a given time, using the available information to the best of their ability. Understanding the importance of interaction and recognizing its value is paramount if statisticians are to assist agronomists with G×E analyses.

Prior to the Finlay and Wilkinson (1963) modelling of G×E interaction (discussed further in Section 2.3), the interested reader would need to look beyond G×E literature to find much work regarding the significance of the interaction component of crop performance. Although the work of Finlay and Wilkinson (1963) re-invented the much earlier Yates and Cochran (1938) model, it came at a time when demand for such a model's use was increasingly evident. The work of Tukey (1949) in a general two-way context starts the story of non-additivity in two-way tables, and has led to the work of others in the area.

Terminology

The term non-additivity comes from the fact that interaction is a component part of the residual left over when the additive two-way model

$$Y_{ik} = \mu + G_i + E_k + \epsilon_{ik} \tag{2.1}$$

is fitted to data. This model expresses the yield Y_{ik} of the ith genotype in the kth environment as the sum of the grand mean μ, the main effects for genotype G_i and environment E_k, and an error ϵ_{ik}. It therefore contains all main effects and forces any other effects to be subsumed into the error term. When replicate data are available, this model more fully expresses the yield of the rth replicate Y_{ikr} as

$$Y_{ikr} = \mu + G_i + E_k + f(i,k) + \epsilon_{ikr} \tag{2.2}$$

or more commonly,

$$Y_{ikr} = \mu + G_i + E_k + GE_{ik} + \epsilon_{ikr} \tag{2.3}$$

with each replicate having error ϵ_{ikr}. The remaining component $f(i,k)$, which reflects the effect of combining the ith genotype and the kth environment, is non-additive. The term for this component is more simply expressed as an interaction term GE_{ik}, and in later sections is known as the multiplicative part of the model. 'Multiplicative models' is used to describe the family of models concentrating on the non-additive component of G×E performance.

As introduced in Section 1.2, interaction of genotype and environment comes about when the difference between the performances of two genotypes changes from one environment to another. A non-constant advantage of one genotype over another can be of two forms; absolute or comparative advantage. These are termed 'wide adaptation' and 'specific adaptation' respectively. A comparative advantage is often found by a change in rank performance of genotypes across environments. 'Crossover' interaction occurs when the difference between the yields changes from positive to negative across environments. While the existence of quantitative interaction is important, it is this qualitative interaction that is truly important to many decision-makers.

Transformations

Transformation has been suggested in G×E analyses, principally to remove heterogeneity of variance (Finlay and Wilkinson, 1963), but also as a means of removing G×E interaction. It has been noted that generally researchers should find a suitable transformation of the data, if it exists, to allow an additive examination of the data (Johnson and Graybill, 1972). Transforming the data is not a guaranteed approach, however, as crossover interaction cannot be removed from data by many transformations because they do not alter the ordering of performances (Freeman, 1985).

Johnson and Graybill (1972) qualify their advice by noting that it is possibly a dangerous venture when the G×E interaction is not a function of the main effects. Transformation to remove any G×E interaction that is a function of the main effects is also unnecessary because many models presented in this chapter allow for non-additivity that is related to the main effects. In fact, some use this relationship explicitly (Finlay and Wilkinson, 1963).

A common option in statistical analyses is to transform data to rank measurements. This has been done in G×E analyses, as addressed in Section 2.7, but is warned against by such authors as Moro and Denis (1997) who stated:

> "Analysis of genotype by environment interactions by ranks, indeed deals only with qualitative interactions but it seems to us that the loss of information implied by transforming yields into ranks is too drastic to be the only way of

analyzing such experiments."

Tests for non-additivity

Tests for non-additivity initially worked with the null hypothesis that a two-way table contained no interaction effect. Various structures of interaction were offered as possible alternative hypotheses. Tukey (1949) used the alternative model

$$Y_{ik} = \mu + G_i + E_k + \kappa G_i E_k + \epsilon_{ik} \tag{2.4}$$

with one additional parameter κ being assigned a single degree of freedom. This then gave rise to use of names like 'Tukey's one degree of freedom test for non-additivity'. This model suggests that the interaction effect is a scalar multiple of the product of main effects. As observed by Milliken and Johnson (1989), this test does not prove that the two-way table is additive because it has a specific structure in the alternative hypothesis.

An alternative to this was provided by Johnson and Graybill (1972) based on a likelihood ratio test. Hegemann and Johnson (1976) compared this test with the Tukey one degree of freedom test for non-additivity, and found that when the model in (2.4) holds, Tukey's one degree of freedom test is the best. In general, however, the test proposed by Johnson and Graybill (1972) is preferable for the bulk of possible interaction structures. They recommended that testing for non-additivity should therefore be conducted using both tests as there is uncertainty about the G×E interaction structure before testing. Williams and Wood (1993) used simulated data to conclude that the test of Johnson and Graybill (1972) should be used.

Milliken and Johnson (1989) presented these and other tests that have been proposed for detecting the existence of interaction in two-way tables. They use interaction plots to illustrate differences between the various approaches which construct an alternative hypothesis to the additive model. The above tests do not distinguish between an interaction that changes the ranking of genotype performances and those interactions that do not result in a rank change, but this problem has been resolved by Baker (1988) who provided two tests for detecting interactions leading to changes in the genotype rankings.

Application to the principal data

The methods for determining the existence of interaction implicitly rely on having complete data, which is not the case for the principal data of this investigation. The one degree of freedom test could be performed on the data by fitting the additive two-way model in (2.1) to the data and estimating the row and column main effects. The residuals from this additive model could then be regressed (without a constant term) against the product of the main effects. As demonstrated by Milliken and Johnson (1989), for complete data, the degrees of freedom in the residual part of this regression would need to be adjusted

to allow for the previously fitted main effect terms. The limitation is then the inability of the test to recognize all forms of interaction.

Following this idea, Gauch (1988) suggested that the only way to be completely sure that interaction is insignificant is to assign one degree of freedom (without fitting a model) to the total sums of squares for interaction, and testing this against the term for replicates. This simple but effective method is not able to be employed in relation to the principal data as replicate data are not available. Snee (1982) contended that there will be difficulty establishing the difference between interaction and heterogeneity of variance in an unreplicated two-way table.

Establishing the existence of G×E interaction in the principal data, using formal testing techniques outlined above, is currently unachievable. Due to the wide variation of the genotypes and environments of the principal data, it is assumed that there is a significant G×E interaction component that must be addressed in this investigation. The fitting of multiple regression models in Section 3.5 confirms this assumption.

2.3 Joint regression

The first real success in finding a model to explain G×E interaction was presented by Yates and Cochran (1938). This model was subsequently overlooked, however, until it was 're-invented' by Finlay and Wilkinson (1963). The foremost value of this joint regression model is that it uses an environmental index in place of covariates to explain the G×E interaction. It is relatively simple to fit using a general linear model (GLM), and provides a different response pattern for each genotype.

The reason that this model is given its own section in this chapter is because its development promoted thought by statisticians and agronomists into the way that G×E interaction was viewed. The troublesome interaction was now successfully modelled and could therefore be used advantageously to make decisions on crop selection (with recognition of the interaction) for the first time. Its use brought out the criticism of the theorists, and the praise of the pragmatists. Its presentation has at times left a lot to be desired. As an example, it has been suggested (Freeman, 1973) that some papers have used inferior, let alone incorrect, ANOVA tables.

While the limitations and interpretations have been brought into focus by many authors over the years, their investigations have meant that all subsequent modelling approaches appear to have been given the same level of scrutiny. A complete understanding of joint regression's use and limitations therefore, gives the foundations of thought that will assist in the future modelling of G×E interaction, from both a user's and a developer's perspective.

The model

The joint regression model attempts to fit a linear function of the environment mean for each genotype to explain the G×E interaction. The model appears in the form

$$Y_{ik} = \mu + G_i + E_k + b_i E_k + \epsilon_{ik} \tag{2.5}$$

Eberhart and Russell (1966) included replicates in their joint regression model, giving

$$Y_{ikr} = \mu + G_i + E_k + b_i E_k + \eta_{ik} + \epsilon_{ikr} \tag{2.6}$$

where η_{ik} is unexplained G×E interaction and ϵ_{ikr} is the error of the rth replicate. The mean of these replicates is zero for every combination of genotype and environment, and when added to η_{ik} gives ϵ_{ik} in (2.5).

The underlying idea supporting the use of the environmental main effect as an index is that a higher level of nutrients, better weather, etc. will result in a higher mean yield in an environment. The environment can therefore be summarized in terms of its mean yield, or main effect, when inadequate covariate information is available (Finlay and Wilkinson, 1963). The capability of the joint regression model to summarize other covariate data was shown by Hardwick and Wood (1972), who compared a multiple regression on covariate information and a joint regression model.

Freeman and Dowker (1973) investigated joint regression's limited ability to summarize covariates by using principal component analysis to determine the number of independent factors causing the G×E interaction. They found that when only one component exists for the G×E interaction, the joint regression model may well be appropriate. When two or more components are significant it is less likely that the joint regression model will serve as a useful tool.

A result of the joint regression model's application has been the way authors have interpreted the outcome. For example, Wright (1976) noted that when the Finlay and Wilkinson joint regression model holds, the environments can be grouped by their main effects, so that groups of homogeneous environments can be analysed to find the best performing genotypes for each group. He proved that in such situations the correlation of two environments will be relative to the difference in their environmental main effects.

Significance and validity of the joint regression model

The correct ANOVA table for a standard joint regression model was provided by Freeman and Perkins (1971). At times, various authors have combined some rows of the ANOVA, notably the rows modelling the environmental main effect and G×E interaction (Eberhart and Russell, 1966). The significance of the explained G×E term, $b_i E_k$ in (2.6), would normally be tested against the unexplained G×E component η_{ik}, while the whole G×E

interaction sum of squares would be tested against the replicate sum of squares. Freeman (1973) however, recommended testing each slope parameter against its own portion of the error component, to ensure that heterogeneity of variance is not masking a problem. Eberhart and Russell (1966) indicate that the use of the F-test for the significance of the joint regression parameters is not valid if the errors display heterogeneity. From a statistician's perspective however, heterogeneity suggests that the model is inappropriate, and another model should be found that circumvents this problem. Regardless of the error term used to test significance, it is imperative that the various terms, especially environmental main effect and G×E interaction, are kept as two separate components when testing the significance of the model (Baker, 1969).

Of course, statisticians know these basic facts of model fitting; at times however, the application of joint regression, and related comments have served to educate the less statistically minded readers of G×E literature. The work of Westcott (1986), for example, showed the effect of deleting the results from the highest and lowest yielding sites on two separate analyses of the data from Yates and Cochran (1938). He showed that the coefficients for the linear regression model were altered by the change, by virtue of the high leverage the deleted values had imparted on the model. These findings supported the Hardwick and Wood assertion that joint regression results are potentially biased when there are low numbers of genotypes or environments.

Criticism of the joint regression model

One of the first criticisms of joint regression was that the environmental index is not independent of the response variable, and is therefore, a violation of the assumptions of regression analysis (Freeman and Perkins 1971). Freeman (1973) later withdrew much of the impact of this criticism by stating that the assumption is not violated if inferences are made on marginal means which can be regarded as fixed.

Another criticism of joint regression is the inability to compare results between analyses. Joint regression parameters are dependent on the genotypes and environments under consideration (Knight, 1970). (This concern applies to the majority of G×E analysis methods.) The joint regression model might be useful for description of the inter-relationships among the genotypes and environments, but cannot be used for prediction of results for a new environment as the environmental index is not calculable (Lin *et al.*, 1986).

Some criticism of the joint regression model has proved unfounded. For example, Hill (1975) noted that there is no *a priori* reason to suggest that joint regression is appropriate, and stated "Indeed, the null hypothesis being tested by the joint regression analysis is that no relationship exists between the G×E interactions and the additive environmental component apart from that due to chance variation." Westcott (1986) identified the error in this statement. He asserted that the joint regression model is searching for a linear response, and that any other response pattern would be rejected under joint regression.

Joint regression as a measure of stability

Finlay and Wilkinson (1963) noted "The simple linear regressions used to describe various types of variety adaptability to a range of environments can also be used as a quantitative measure of phenotypic stability." In a joint regression context, stability is reflected by a genotype having b_i near zero. A stable genotype is therefore a genotype with performances that are parallel to the average performance in each environment. The only real embellishment of the Yates and Cochran (1938) model given by Finlay and Wilkinson (1963) was to plot the joint regression coefficient b_i for each genotype against its mean yield $\bar{y}_{i\cdot}$ in all environments (seen in Figure 2.1). This provided a triangular shaped scatter plot of genotypes whose extremities have different interpretations. Points with significantly positive slopes ($b_i > 0$) are said to be well adapted to high yielding environments, while negative regression slopes ($b_i < 0$) are said to be well adapted to low yielding environments. Genotypes with regression slope of zero are said to be equally adapted across environments and have 'dynamic' stability. If these genotypes have high mean yields they are said to be well adapted and if they have low mean yields they are said to be poorly adapted across all environments. It is therefore desirable to have either extreme b_i and low mean yield or high mean yield and $b_i = 0$. An 'efficient' set of cultivars can be found using this approach. Cultivars around the outside of the triangle pattern are those that should be recommended, with the exact recommendation depending on the environment's mean yield. It should be noted that genotypes with b_i near negative one are considered stable under a different definition of stability. 'Static' stability is defined as constant performance across environments, irrespective of any change in the quality of inputs. Note that the notation and parameterization used in this section differ from that of Finlay and Wilkinson (1963) for reasons of consistency throughout this investigation.

Finlay and Wilkinson (1963) noted that the triangular shape existed for their data but did not attempt to explain this for all data sets. This phenomenon will exist as the result of using the mean environmental yield as the explanatory variable in the model. For example, a high yielding and heavy interaction genotype added to the data would increase the explanatory variable for each environment and affect the parameters found when the model is fitted.

Eberhart and Russell (1966) used the sum of squares that are left unexplained by joint regression attributable to each genotype as a second stability parameter. A desirable genotype would in their opinion have a regression slope near zero (tracking the environmental index), and a low deviation sum of squares. Breese (1969) advocated the use of the second Eberhart and Russell (1966) parameter as it is a measure of the unpredictable component of G×E interaction, but the validity of this second parameter has been questioned extensively. Hardwick and Wood (1972) state that in terms of their underlying model the deviations are not independent of the regression on the environmental mean, so that the second stability parameter of Eberhart and Russell (1966) is not meaningful.

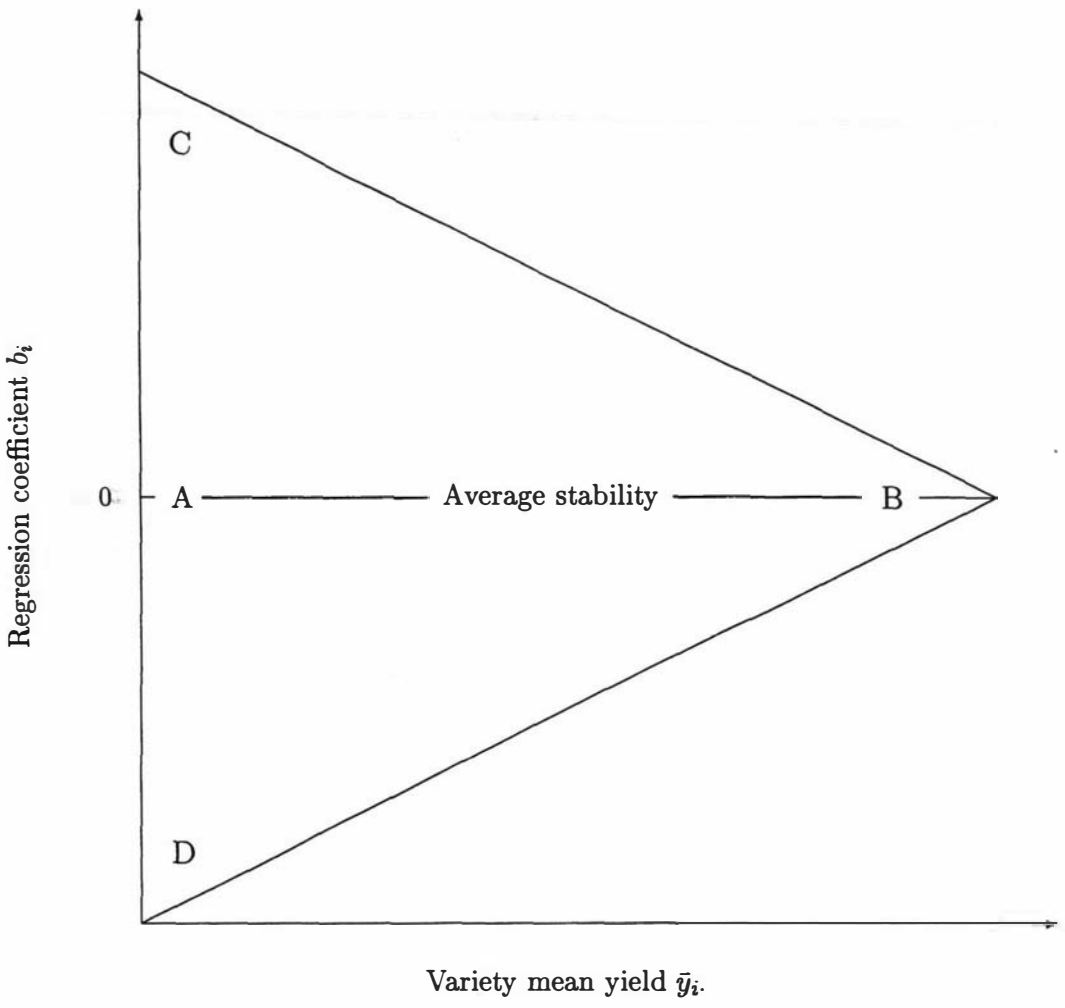


Figure 2.1: Alternative representation of the Finlay and Wilkinson (1963) plot that shows genotype stability. The model in (2.5) is used in preference to that proposed by Finlay and Wilkinson (1963) to ensure consistency throughout the discussion. The plot of genotype joint regression coefficients b_i versus genotype mean yield $\bar{y}_{i\cdot}$ will show a triangular shape in general. Genotypes plotted near 'A' in the figure are said to be poorly adapted in general, while genotypes plotted near 'B' are said to be well adapted in all environments. Genotypes plotted near 'C' are defined as unstable, but specifically adapted to higher yielding environments. Genotypes plotted near 'D' are said to be stable and specifically adapted to the lower yielding environments.

Westcott (1986) identified the potential of Eberhart and Russell's (1966) second stability parameter to label a genotype as unstable when it in fact has a different response pattern to the others in the analysis. The second Eberhart and Russell (1966) parameter is therefore useful as an indicator of the poor fit of the joint regression model for some genotypes.

The next section presents the variations that have been proposed for joint regression. A discussion of the relevance of joint regression to this investigation is left until the end of that section.

2.4 Variations on the joint regression model

The standard joint regression model attempts to explain G×E interaction for each genotype by linearly scaling the environmental main effect. As described in Section 2.3, the slope of this line (b_i) for each cultivar can be plotted against the mean yield for each cultivar. This section outlines how joint regression has been varied to suit different research interests. These variations are both graphical and numeric.

Some standard regression techniques have been incorporated into the joint regression framework, including for example, the use of weighted regression (Johnson, 1977) and the use of a quadratic function of the environmental main effect (Ng, 2001).

Joint regression on genotypes rather than environments

Wright (1971) proposed using the genotype main effect as the explanatory variable in a joint regression. The model

$$Y_{ik} = \mu + G_i + E_k + b_k G_i + \epsilon_{ik} \tag{2.7}$$

which has parameters b_k for each environment to respond in a different linear way to the genotypes under investigation. He noted that there was no *a priori* reason to suggest that the joint regression model should use the environment main effects as an index, but that if both models (2.5) and (2.7) yield significant relationships then a model that considers both sets of parameters together should be investigated. This model was originally presented by Tukey (1949), given in (2.4), and allows a single parameter κ to scale the product of the two main effect vectors. Wright (1971) suggested fitting this parameter using

$$\hat{\kappa} = \frac{\sum_{i=1}^{I}\sum_{k=1}^{K} y_{ik} G_i E_k}{\sum_{i=1}^{I} G_i^2 \sum_{k=1}^{K} E_k^2} \tag{2.8}$$

This parameter describes the situation where the G_i's and b_i's in (2.5) are completely correlated, that is, all joint regression lines have a common point of intersection. The existence of concurrent regression lines in practice is rather debatable.

Ng (2001) converted a quadratic joint regression model so that it used genotype main effects instead of environmental main effects as explanatory variables. The model using environmental main effects was presented as

$$Y_{ik} = \mu + G_i + E_k + b_{i1}E_k + b_{i2}E_k^2 + \epsilon_{ik} \tag{2.9}$$

and her alternative model as

$$Y_{ik} = \mu + G_i + E_k + b_{i1}G_i + b_{i2}G_i^2 + \epsilon_{ik} \tag{2.10}$$

She found this model unsuitable for the data set under consideration.

Trimmed joint regression

Gusmão *et al.* (1992) provided a method that would fit the joint regression model with the outliers in the residuals removed so that a better model could be found. They replaced an observation with an extreme residual (high specific adaptation) with the fitted value for that G×E combination. Iterating until no observations needed to be removed would in their opinion lead to an improvement in the use of this model.

It is debatable whether this technique will be of any use, unless the underlying phenomenon can be successfully modelled using the joint regression model in the first place. Trimmed joint regression will not be applied further in this work, and has received little attention in the G×E literature.

Fitting non-linear joint regression

Non-linear regression is an iterative process for fitting a model using nonlinear optimization techniques. It allows one of the explanatory variables to be dependent on the fitted values for the data. Using the mean of the fitted values in each environment as the explanatory variable in a joint regression can therefore be achieved using a non-linear regression model.

When data are complete the difference between the linear and non-linear models is negligible as the parameters and their standard errors will be the same. This is not necessarily true when data are incomplete. The non-linear approach has been found superior when the joint regression model is fitted to incomplete data (Ng and Williams, 2001).

Joint regression and missing data

Digby (1979) used an iterative approach to find joint regression parameters that account for the missing data in his ten genotype by seventeen environment table of spring wheat. This data is notable for its similarity to the principal data, introduced in greater detail in Chapter 3, as the missing data is a result of insufficient resources to test all varieties in all locations. The data is less sparse and somewhat smaller in size, however, having data for 134 of the 170 combinations of the ten genotypes and seventeen environments.

The iterative approach of Digby (1979) has been re-expressed to use the same notation as the joint regression model found in (2.5), and follows the process:

1. Set all joint regression coefficients b_i to zero.
2. Estimate μ, G_i, and E_k using current estimates of b_i.
3. Re-estimate b_i for each genotype using current estimates of μ, G_i, and E_k.
4. Repeat steps 2 and 3 until sufficient convergence is achieved.

This approach is an improvement on that offered by Freeman (1975), who analysed a G×E matrix with over 50% of its data missing. Digby (1979) noted the problem of using an additive model to impute missing data, as done by Freeman (1975), and then fitting a multiplicative model to the complete data. Such an approach biases the parameters and reduces their likelihood of being deemed significant.

On the other hand, Ng and Williams (2001) criticized the approach of Digby (1979) on the grounds that it does not give the correct standard errors for the parameter estimates. This occurs as the parameters are not fitted concurrently; some parameters are held constant, and therefore assumed known, while others are estimated. They compared the method of Digby (1979) with a non-linear regression method proposed by Ng and Grunwald (1997) which was found superior for two reasons:

1. It has better efficiency, as its expected rate of convergence is more consistent.
2. It is also a stronger theoretical method for fitting parameters, as the correct standard errors are used for inference.

The Ng and Grunwald (1997) method uses linear estimates of the non-linear objective function being minimized, and iterates until convergence is reached. At this point a local optimum has been found that has parameter estimates that can be used for inference and comparison.

Application to the principal data

While the joint regression model has its criticisms (both valid and invalid), in situations where it is appropriate it can be used in conjunction with the knowledge of its limitations

to provide a suitable means of describing the relationship between genotypes and their environmental performances.

The major drawback to joint regression analysis is that the environments are assumed to be similar in all respects other than those that linearly impact on mean yield. Many authors suggest that this is not the case. In some situations it may be possible to take subsets of the environments so that there are groups of environments that are homogeneous within each subset (with different mean yields) but heterogeneous between subsets (not necessarily in terms of mean yield). Eberhart and Russell (1966) noted

> "Stratification of environments has been used effectively to reduce the genotype-environment interaction. The region for which a breeder is developing improved varieties can often be so subdivided that all environments in the sub-region are somewhat similar. This stratification usually is based on such macro-environmental differences as temperature gradients, rainfall distribution, and soil types."

Crossa (1988) split the 37 environments of his data into two groups for low and high yielding environments (negative or positive environmental main effect respectively), following the approach of Verma *et al.* (1978). Verma *et al.* (1978) fitted the joint regression model to the two environmental groups separately, and compared the responses for each genotype. According to Verma *et al.* (1978), a desirable genotype would have positive b_i in high yielding environments and negative b_i in low yielding environments. This therefore provided a logical mix of dynamic and static stability.

When the approach of Verma *et al.* (1978) and Crossa (1988) is taken, there are four parameters to compare for each genotype; two mean yields $\bar{y}_{i\cdot}$ and two joint regression coefficients b_i. Plotting the pair of points from each joint regression, as in Figure 2.1, and then joining them with a line segment will show:

1. How the genotype is affected by the change from one subset to the other.
2. The genotype's relative improvement in mean yield.
3. The magnitude of the comparative advantage of choosing the right environment for each genotype, as measured by the length of the line segment.

This suggestion could be used effectively if a factor other than mean yield was used to determine the environment grouping, such as tropical versus non-tropical locations. The two joint regression coefficients b_i could be plotted against each other to establish the need to subdivide environments. Highly positively correlated sets of b_i values would indicate that there is no need to subdivide environments.

The need to subdivide environments suggests that there is more than one factor affecting the G×E interaction. As noted by Freeman and Dowker (1973), the joint regression framework is likely to be most effective when only one factor, or a the combination of two

or more factors, affects the interaction in a linear manner. There is no guarantee that this is the case for the principal data of this investigation. The data for which the joint regression model has proved effective have in general arisen from regional trials programmes where environmental differences are lower in magnitude. The next section introduces modelling approaches that allow for a larger number of factors to describe the G×E interaction exhibited by the data.

2.5 Multiplicative models

Multiplicative models are those models that have a component based on both genotype and environment effects or factors in combination. The joint regression model given in (2.5) is a simple example, and many others have been used in G×E analyses. In this section, the more notable multiplicative models are introduced including the most well known additive main effects and multiplicative interaction (AMMI) model.

Factorial regression models

Baril *et al.* (1995) presented factorial regression models of the form

$$Y_{ik} = \mu + G_i + E_k + \sum_{h=1}^{H} \upsilon_{hi} z_{hk} + \sum_{m=1}^{M} x_{mi} \tau_{mk} + \sum_{m=1}^{M} \sum_{h=1}^{H} \phi_{mh} x_{mi} z_{hk} + \epsilon_{ik} \qquad (2.11)$$

where M genotype covariates $x_{m\cdot}$ and H environmental covariates $z_{h\cdot}$ are available as explanatory variables for inclusion in a linear model.

Fitting this model in full will use up many degrees of freedom as the parameters ϕ_{mh}, υ_{hi}, and τ_{mk} will use $H\,(I-1)+M\,(K-1)+HM\,(I-1)\,(K-1)$ degrees of freedom in total for the terms used to explain the G×E interaction. If there are too many covariates available, the factorial regression model could quickly become over-parameterized and unwieldy.

Reduced models can be formed using the above general expression if no genotype or environment related covariates exist. These reduced models would be multiple regression models which include terms for interaction between genotypes (environments) and environment (genotype) covariates. Factorial regression models can provide a means of interpreting G×E interaction (Baril *et al.*, 1995), but their ability to provide meaningful interpretations relies heavily on the appropriateness of covariate information collected over the course of a trials programme.

Additive main effects and multiplicative interaction (AMMI) models

Principal components are often used in multivariate data analysis to gain an understanding of the differences among a sample of observations in terms of the variables over which

they have been measured. In G×E analysis, the environments are used as variables and genotypes as observations. At times, covariate information that would be used in factorial regression models is not available, so principal component terms are substituted for the factorial regression terms in (2.11) to give the additive main effects and multiplicative (AMMI) model

$$Y_{ik} = \mu + G_i + E_k + \sum_{n=1}^{N} \lambda_n u_{in} v_{kn} + \epsilon_{ik} \tag{2.12}$$

The eigenvalues λ_n, and corresponding eigenvectors $u_{.n}$ and $v_{.n}$ for genotypes and environments respectively are generated from a singular value decomposition of the doubly centred G×E matrix. The model in (2.12), actually represents a family of models as the value of N can range from zero to the lower of I and K. The number of multiplicative terms in the AMMI model cannot be determined *a priori* (Gauch, 1988), but is chosen for each data set on the grounds of parsimony and the significance of terms (discussed further below). The unexplained G×E interaction and error are included in the term ϵ_{ik}. The ϵ_{ik}'s are assumed independent and normally distributed. The AMMI model has been generalized to deal with counts, proportions, and non-normally distributed errors, thus creating a class of models named GAMMI (van Eeuwijk, 1995b). There is little need in the current context to cover this material in depth, but its usefulness should be noted in dealing with such response variables as pestilence resistance and storage potential.

Although AMMI models were in existence in areas such as psychology etc., they were not used widely in G×E analyses until the late 80s and early 90s. There has been widespread use since that time as they provide convenient means of interpreting a large amount of information via the use of biplots (Kempton, 1984; Zobel *et al.*, 1988; and Gauch, 1992). It has often been observed that the AMMI model with one multiplicative term is superior to the joint regression model, especially given that it explains a greater portion of the G×E interaction sum of squares (Freeman, 1975; Gauch, 1990; Yau, 1995).

Another major advantage of the AMMI model is its ability to overcome problem effects of non-homogeneous variance caused by either row or column effects. Snee (1982) asserted that a model capable of distinguishing interaction from heterogeneous variance should be used, especially when only one observation is available for each combination of the row and column effects. The AMMI model is one such model that can cater for many forms of non-additivity.

Predictive ability of the AMMI model

The ability of AMMI models to summarize G×E data with interaction that arise from a number of sources is well documented. Gauch (1990) found that a reduced AMMI model was actually better than the treatment or cell means model ($N = \min\{I, K\}$). He noted the influence of the Stein effect in stating

> "Insofar as a model differs from its data, according to this conventional perspective, it is worse than the data. But the Stein effect says the exact opposite - a reduced model can be better than the full model. Since the individual replicates have error the treatment means have noise."

He asserted that the benefits of using a reduced model arise:

> "Because the AMMI model misspecifies the true model to some degree, this use of indirect information introduces some bias. However, in many cases the problem with variance outweighs this problem with bias, particularly given a small number of replications and a sizeable EMS. Consequently the Stein effect occurs. But AMMI's improvement does not come from nowhere or from sheer computation, but rather from using more data in making each yield estimate."

In keeping with this notion, Crossa *et al.* (1990) found that the reduced models chosen for two data sets were better than the treatment means models. They used cross-validation techniques to show that the predictions based on the reduced models are better than those predicted by the cell means.

Cross-validation relies on the existence of replicate data. The data used in the analysis presented by Gauch (1988) had differing numbers of replicates at some sites, lending itself to this approach. The problem of determining the validity of the AMMI model when only one replicate is available for each G×E combination remains unsolved.

As a consequence of its predictive success, it has been claimed that use of AMMI can save resources (Crossa *et al.*, 1990; Gauch, 1992). These authors report the predictive success of their models in terms of the number of additional replicates that would have been required to achieve the same level of accuracy. To state *a posteriori* , that resources were saved is irrelevant, but to be able to safely predict the amount of resources required in any new investigation is a laudable outcome (considered further in Chapter 10).

Gauging the significance of models

Zobel *et al.* (1988), as well as Gauch (1988), noted the ability of the AMMI modelling process to suggest use of a different model. Zobel *et al.* (1988) looked at the significance of the main effect terms G_i and E_k together or the multiplicative component to suggest a different model, while Gauch (1988) looked at selection of the 'right' number of multiplicative terms.

In either of these situations, the correct number of degrees of freedom needs to be assigned to each term in the model. The degrees of freedom for genotype and environment main effects have not been disputed and are $(I-1)$ and $(K-1)$ respectively. In many analyses using the AMMI model, the method for counting the degrees of freedom for the

nth multiplicative term has been determined using the formula

$$df(n) = I + K - 1 - 2n \tag{2.13}$$

due to Gollob (1968). This gives the nth multiplicative term degrees of freedom relating to the total number of genotypes I, environments K, and n the number of the multiplicative term under consideration. Note that successive multiplicative terms will have fewer degrees of freedom associated with them using this formula.

Piepho (1995) compared the F-test using the degrees of freedom formula of Gollob (1968), with three other tests of significance of multiplicative terms in an AMMI model when the residuals are either non-normal or heteroscedastic. One of the tests he proposed is superior to the more commonplace F-test using the Gollob formula for degrees of freedom, when either of the assumptions based on normal and homoscedastic residuals is violated.

When no error term for replicates can be estimated, it is recommended that the unexplained G×E interaction term be used as a 'surrogate' error term (Gauch, 1992). If AMMI models are to be used with the principal data of this investigation this use of unexplained G×E interaction will need to be employed in lieu of a replicate error term.

One problem of AMMI models, especially when there is no replicate error term available, is the tendency to add too many terms into the model on the basis of their significance. Cornelius (1993) and Piepho (1995) both found that the tests for significant terms in the AMMI model would quite likely suggest adding more terms than were necessary thus leading to over parameterization. Some subjective judgement is therefore required when selecting an AMMI model for data, and should always encompass the notion of parsimony.

Linking AMMI to factorial regression

Principal component based models have their place, but the need to develop good covariate based models cannot be overstated. Their application must be the ultimate aim in any modelling procedure if real world phenomena are to be determined as the cause of differences in genotype performances. Criticism of the use of the environmental main effect in a joint regression model can be extended to the AMMI model. When an AMMI model is fitted, any number of multiplicative terms may be determined but these may remain abstract. Attempts have been made to link the principal terms of AMMI models to environmental covariates (Voltas *et al.*, 1999).

Having found the suitable AMMI model or other multiplicative model based on principal components, the researcher could check for correlation between a real climatic or edaphic factor that closely matches the particular component. A combination of principal components and specific factors can therefore be applied to elucidate the nature of the G×E interaction.

Brancourt-Hulmel *et al.* (1997) noted that multiple regression methods using covariate information should be preferred to AMMI models because the covariates for a new environment can be estimated, and therefore predictions for the new environment can be made. When working with two complex data sets, Vargas *et al.* (1999) found that the use of factorial regression models proved superior to AMMI models, for several reasons:

1. They used real covariate information.
2. The parameters could be tested for their significance.
3. The factorial regression models found the appropriate concomitant variables to explain a significant portion of the G×E interaction, as did AMMI models.

Unfortunately, they also found that several factorial regression models explained the relationships in the G×E interaction differently, thus leading to some confusion of the significant factors.

A more direct link between factorial regression and AMMI models was proposed by van Eeuwijk (1995a). The model, known as 'reduced rank regression', replaces the environmental principle component terms in (2.12) with linear combinations of the environmental concomitant variables $z_1, \ldots, z_H$, for each of the n AMMI axes used. The reduced rank regression model given as

$$Y_{ik} = \mu + G_i + E_k + \sum_{n=1}^{N} \lambda_n u_{in} \left[\sum_{h=1}^{H} \psi_{hn} z_{hk} \right] + \epsilon_{ik} \tag{2.14}$$

The ψ_{hn}'s are coefficients of the environmental covariates $z_1, \ldots, z_H$ in the nth reduced rank factorial regression axis.

Relationships to other models

Removal of terms from the AMMI model, given in (2.12), results in a series of related models (Zobel *et al.*, 1988). In particular, removal of the multiplicative terms results in the additive model of (2.1). Removal of G_i or E_k from (2.12), results in models that were termed column regression and row regression respectively (Bradu and Gabriel, 1978), while removal of the grand mean and all main effects for genotype and environment results in a model termed the 'completely multiplicative model' by Seyedsadr and Cornelius (1992). This model, denoted COMM(N),

$$Y_{ik} = \sum_{n=1}^{N} \lambda_n u_{in} v_{kn} + \epsilon_{ik} \tag{2.15}$$

is therefore based entirely on principal components. Its chief problem is that it does not separate the genotypic and environmental main effects from the G×E interaction, and this confounding makes interpretation difficult. If either genotype or environment main effects

are significant, they are likely to dominate the first PC axes. Fitting an AMMI model therefore gives the researcher an opportunity to determine that one of these other models is equally effective and more parsimonious (Zobel *et al.*, 1988).

Adding a constant β to the completely multiplicative model of (2.15) gives

$$Y_{ik} = \beta + \sum_{n=1}^{N} \lambda_n u_{in} v_{kn} + \epsilon_{ik} \tag{2.16}$$

which has been named the 'shifted multiplicative model' (SHMM) by Seyedsadr and Cornelius (1992) who showed how it could be fitted to data.

Unlike the additive, AMMI, row and column regression models, the shifted multiplicative model with N multiplicative terms, here denoted SHMM(N), cannot be fitted using a combination of ANOVA and singular value decomposition. Instead, Seyedsadr and Cornelius (1992) recommend use of the Newton-Raphson iterative method for finding an estimate $\hat{\beta}$ of the constant term in the shifted multiplicative model. This method is suitable if the initial estimate of $\hat{\beta}$ is close enough to the optimal value in the model (Seyedsadr and Cornelius, 1992). They recommend various initial values for $\hat{\beta}$ in the SHMM(N) model including:

1. Zero, thus fitting the completely multiplicative term on the first iteration.
2. The grand mean of the data $\bar{y}_{..}$, which is considered to be a poor choice.
3. Using $\hat{\beta}$ found for SHMM(N-1) or SHMM(N+1).
4. The parameter κ in the model of Tukey (1949) given in (2.4).

In some instances the estimate of $\hat{\beta}$ can indicate use of an alternate model. If $\hat{\beta}$ is well outside the range of the data then SHMM(N) is equivalent to AMMI(N-1) (Seyedsadr and Cornelius, 1992), and more specifically, if $\hat{\beta}$ in the SHMM(N) model is near zero then the best model for the data is probably COMM(N).

The SHMM model can be interpreted as providing a set of concurrent regression lines (Crossa *et al.*, 1993). The fitted values of a SHMM(1) model would appear as a set of lines all passing through a single point, and if this point is outside the range of the data, the model suggests that the G×E interaction is proportional interaction, rather than crossover interaction (Crossa *et al.*, 1993). This fact has been used to identify groups of environments that show no significant G×E interaction with genotypes (Crossa *et al.*, 1993).

Graphical interpretation of results

The most common way to interpret the findings of an AMMI or other multiplicative model is to use biplots. Biplots mark points for both genotypes and environments on the same axes, but usually use arrows emanating from the origin to indicate either genotypes or environments.

The first two principal component (PC) axes are normally plotted, but Crossa et al (1990) interpreted a biplot with main effects and the first interaction axis. Gauch and Zobel (1997), however, note the need to determine the difference between the plot of the first two PC axes and the plot of the first PC axis and the mean effects of genotypes and environments. The second plot will assist with the selection of the best genotypes for each environment, while the first will give a greater insight to the nature of the interaction (Gauch and Zobel, 1997).

When a biplot is created using the first two PC axes, the genotype arrows point in the same direction as the points for environments to which they are most specifically adapted. Projecting the points for environments onto these arrows will give an indication of the strength of the adaptation, which can be measured using the length of the projections. Extending the biplot to include a third dimension has been investigated by Gower (1990), and to allow for confidence regions in two-dimensional biplots (Denis and Gower, 1996).

Multiplicative models and missing data

In Section 2.4, it was noted that Freeman (1975) imputed missing G×E data using the additive model of (2.1) and subsequently fitted a multiplicative model to the complete data. The imputed values are therefore not consistent with the fitted model, and actually work against the model in that the significance of terms may be understated.

The Healy-Westmacott algorithm, described by Little and Rubin (1987) and McLachlan and Krishnan (1997), is a self-consistent method whose fitted values suit the model and in part determine that model. The algorithm:

1. Imputes missing values.
2. Fits the proposed model (assuming it to be the correct model).
3. Uses the outcome of this model to predict the missing values.
4. Iterates until some convergence criterion is satisfied.

If the errors from the model are normally distributed, the Healy-Westmacott algorithm is an implementation of the EM algorithm (McLachlan and Krishnan, 1997).

The EM algorithm and the AMMI model were combined by Gauch and Zobel (1990). Their approach works through the following process:

1. Estimate missing yields using the additive model given in (2.1).
2. Find parameters for the N multiplicative terms in the AMMI model.
3. Re-estimate missing yields using the current set of parameters.
4. Continue steps 2 and 3 until some convergence criterion is attained.

Use of the unweighted two-way additive means as initial estimates of missing entries was recommended by Gauch and Zobel (1990), but other sets of initial values were also offered, The choice of starting values has an effect on the number of iterations that the EM-AMMI model takes to converge.

In standard AMMI model fitting the first interaction term fitted will remain the same when other terms are added. That is, the first interaction axis will be identical in two different AMMI models for the same data. This is not the case for the EM-AMMI model as the flow-on effects of the estimation of missing values will impact on the parameter estimates found for different EM-AMMI models. It is therefore necessary to fit each EM-AMMI model separately for a given set of data (Gauch and Zobel, 1990).

In theory it is possible to fit the EM-AMMI(N) model to data that has at least $N+1$ observed values in every row and column of the G×E matrix. There is also a need to ensure that the data are connected, but this was not explicitly mentioned by Gauch and Zobel (1990) who instead gave minimum criteria based on having at least $N+1$ complete rows and $N+1$ complete columns within the data. Gauch and Zobel (1990) claim that the EM-AMMI model can be used on data sets as sparse as the principal data of this investigation, as long as it meets their specified minimum criteria.

Application to the principal data

The AMMI model has been recommended for the analysis of G×E data, especially for large regional or international trials (Yau, 1995) and should be used in favour of the joint regression model (Gauch, 1990; Yau, 1995). Yau (1995) observed that the joint regression model may, when appropriate, provide a simpler means of describing the G×E interaction without the loss of too much information.

It seems quite reasonable to expect that the Healy-Westmacott algorithm and any of the models presented in this chapter could be applied to the principal data of this investigation. An obvious starting point was to use the EM-AMMI model of Gauch and Zobel (1990), and these results are presented in Section 3.8. The chief limitation on use of factorial regression and reduced rank regression models is the unavailability of sufficient genotypic and environmental covariates.

2.6 Cluster analysis

Cluster analysis is used in G×E analyses to identify distinct groups of homogeneous genotypes or environments. As a data-driven descriptive tool, cluster analysis is most suited to specific adaptation problems rather than wide adaptation problems (Kang, 1998). To this end, clustering of genotypes is performed on the basis of both genotype main effect and G×E interaction (Mungomery *et al.*, 1974), or more commonly the G×E interaction alone (Lin, 1982).

The outcome from a cluster analysis, usually a dendrogram, does not immediately assist in determining the underlying phenomena that cause differences in genotype performances. A dendrogram also does not give an indication of the performance, but suitable averaging has been used within clusters to overcome this problem (Byth *et al.*, 1976; Ivory *et al.*, 1991). Such averaging may remove the effects of errors that have impacted on the observed data, as a clustering is usually based on raw data rather than expected values.

In this section, the use of hierarchical agglomerative clustering in G×E research is discussed. To complete clustering, decisions need to be made over which distance measure and cluster linkage method will be applied. The application of data transformations and other distance measure construction is briefly introduced below.

Another feature of cluster analyses is the need to choose the level at which clustering is truncated. This has usually been performed manually, often upon inspection of the resulting dendrogram. Few publications use sophisticated criteria for determining the truncation level or gauging the significance of the resulting clustering. Use of cluster analyses is therefore very different to use of the models presented in the previous sections.

Aims of cluster analysis

The first point to consider in this discussion is why clustering is used in G×E research. A good clustering will explain a large portion of the sums of squares that exist in a data set in the most parsimonious manner. Cluster analysis techniques have been employed by G×E researchers to meet wide-ranging objectives. These include:

1. Clustering genotypes when there were many to compare across a smaller number of environments, so that the average or best genotype from each cluster could be compared (Mungomery *et al.*, 1974; Lin, 1982).

2. Clustering genotypes to ensure they are compared to the most similar check cultivars (Lin and Binns, 1985).

3. Clustering environments when there were a smaller number of genotypes to compare, so that genotypes could be compared to the average environment from each cluster (Ivory *et al.*, 1991).

4. Clustering environments to subdivide the G×E matrix into matrices with little G×E interaction in order to better understand the nature of the environments (Abou-El-Fittouh *et al.*, 1969). Researchers can then establish which environments are so similar that they give the same results, and therefore do not all need testing in future thus leading to resource efficiency (Lin and Morrison, 1992; Baril *et al.*, 1994; May and Kozub, 1995).

5. Clustering both genotypes and environments to reduce the number of comparisons necessary (Byth *et al.*, 1976; Corsten and Denis, 1990; Baril *et al.*, 1994; Cooper and

DeLacy, 1994).

Many of these approaches aim to separate the influences of main effect and G×E interaction. It is well-known that the impact of G×E interaction makes comparison of averages or main effects invalid if it is ignored. Finding groups of genotypes or environments so that the G×E matrix can be subdivided into submatrices that exhibit no G×E interaction will allow comparisons based on means (Lin, 1982). In a similar vein, Ivory *et al.* (1991) used bar graphs to compare genotype means for each group of environments, which has the advantage of displaying the best environmental grouping to which each genotype is specifically adapted. Rather than use a comparison of simple means, Lin *et al.* (1986) suggests using multiple range tests within groups found by clustering with a distance based on G×E interaction.

Yau (1991) criticized the approach of Lin (1982) for clustering genotypes on the basis of G×E interaction similarity because "the clustering of widely adapted with non-adapted lines is not acceptable to most plant breeders". The flaw in this statement is that it assumes that all genotypes grouped together are interchangeable, but this is not the case when the means are then compared. Lin (1982) advocated the comparison of genotypes within clusters on the basis of their mean performance, and in this instance, non-adapted genotypes would be exposed as inferior.

Ivory *et al.* (1991) used a 6×19 G×E matrix, while the matrix of Cooper and DeLacy (1994) was 15×10 in size. These authors used cluster analysis to classify the environments, but only in the case of Ivory *et al.* (1991) does this seem logical. It seems to have proved more effective to reduce the number of comparisons by reducing the longer dimension of the G×E matrix, such as Mungomery *et al.* (1974), who clustered the 58 genotypes in their scenario. Subsequent use of ordination to show how the different clusters differ in a two-dimensional display has been named 'pattern analysis".

Cooper and DeLacy (1994) have reduced their 15×10 G×E matrix to a 5×5 matrix and subsequently plotted the means on an interaction plot. Their problem scenario could probably have been managed without clustering both genotypes and environments, but Byth *et al.* (1976) needed to group both genotypes and environments to reduce their 49×63 G×E matrix. They clustered genotypes and environments separately to get a 10×10 matrix, which they further reduced when plotting results on an interaction plot. These authors noted that greater benefit could be gained by combining the initial classification of genotype and environments. This was later achieved by Corsten and Denis (1990) and then Baril *et al.* (1994).

As discussed, the performance of genotypes has commonly been investigated (after cluster analyses) using graphical techniques. Lin *et al.* (1986), when writing in support of cluster analysis, stated: "The advantage of a non-parametric approach is that a cultivar's response characteristics can be assessed qualitatively, without the need for a mathematical characterization." Byth *et al.* (1976) were able to define a model for their two-way classi-

fication, that included terms for main effects, genotype grouping, environment grouping, differences for individual genotypes from genotype group effects, differences for individual environments from environment grouping and the four interaction terms that were then available. Parameterization of the outcome of clustering has been addressed further in Section 5.5.

Distance measures

All cluster analyses use a distance measure when forming clusters. The particular measure used differs according to the need of the researcher. For example, Abou-El-Fittouh *et al.* (1969) grouped environments in the U.S. cotton belt by grouping on the basis of the correlation coefficient, and Lin (1982) used a distance measure that allowed the cluster analysis to be compared to a two-way ANOVA.

Corsten and Denis (1990), in their simultaneous clustering of genotypes and environments, adjusted observed inter genotypes and inter environment distances by the number of comparisons being made. They did, however, assume the matrix to be complete.

Ouyang *et al.* (1995) clustered locations using an incomplete G×E matrix. As with Corsten and Denis (1990), they used an average squared difference to form their distance matrix. Some inter-location distances were not available, and Ouyang *et al.* (1995) estimated these unobserved distances using the maximum of the observed distances. This could be done in situations like their data, but is not appropriate when the data is structured differently. The Ouyang *et al.* (1995) approach is considered further in Chapter 4 when the need to estimate unobserved distances in the principal data is discussed.

Standardization and scaling

Standardization of variables is commonly performed in cluster analyses to remove any excessive weighting caused by changes in the variance of variables (Manly, 1994). Although the problem is less noticeable in G×E research, many transformations have been employed to ensure that clusters are based on the desired biological basis.

DeLacy *et al.* (1990) compared four transformations that could be used within environments to scale the yields of genotypes. These four transformations were:

1. Centring within an environment by subtracting the environment mean. Ivory *et al.* (1991) used this transformation, later termed 'coding' by Cooper *et al.* (1993).

2. Standardizing the yields within an environment. Squared Euclidean distance measured using this data would result in a measure of the phenotypic correlation between environments (Cooper *et al.*, 1993).

3. Ranking the yields within each environment, and then subtracting the average rank so that results are comparable across environments.

4. Scaling the ranks of within-environment yields by the within-environment standard deviation of the ranks, which is necessary when different numbers of genotypes are used in each environment across years.

Cooper *et al.* (1993) discussed a further rank based transformation which does not use the centring of ranks within environments before the scaling by the standard deviation of ranks. The centring of ranks is, however, not important when data are complete.

All of these transformations are aimed at determining which environments order genotype performances in a similar way. If environments are to be classified on these grounds, then the transformations most advocated would be standardization of yields (Fox and Rosielle, 1982, Cooper *et al.*, 1993) or within-environment standardized ranks (Cooper *et al.*, 1993). These suggestions amount to standardization of the observations being clustered, whereas standardizing variables when clustering observations is more common outside G×E research. Yau (1991) clearly stated the need to use within-environment standardization when clustering genotypes.

Yau (1991) also advocated the use of range transformation as an alternative to standardization, but noted its lack of availability in standard statistical packages. This transformation divides the yields of each G×E combination by the within-environment range and results in the adjusted data for every environment having a range of one.

Cluster formation strategies

The vast majority of cluster analyses in the G×E literature are based on hierarchical agglomerative cluster formation strategies. There are many methods for forming clusters that fall under this broad umbrella, and little comparison of their worth to G×E analyses has been published. A notable exception to this is the work of Ramey and Rosielle (1983) who suggested a better method for forming genotype clusters than the method proposed in Lin (1982).

Other strategies have been devised for identifying homogeneous groups of observations, and have been reviewed in Anderberg (1973) and Everitt (1993). Lefkovitch (1985) described an approach for clustering genotypes on the basis of similarity of both mean and across environment variation for each genotype using a conditional clustering method presented by Lefkovitch (1980). McLachlan and Basford (1988) show how mixture models can be employed to determine the memberships of a pre-determined number of clusters. Discussion of the clustering method presented by Moro and Denis (1997) is left until Section 2.8 because it is based on the dominance of one genotype over others, rather than on symmetric relationships like most cluster analyses.

Stopping criteria

Other than manual inspection of dendrograms, little attention has been given to the point at which clustering should be truncated. In some studies, the number of clusters has been chosen before clustering was commenced (Byth *et al.*, 1976), but two other broad strategies have appeared in the G×E literature.

Ghaderi (1980) suggested that the correct truncation point for clustering could be found using the among-cluster to within-cluster variance ratio of ANOVA, but this was criticized by Lin and Butler (1990) as laborious. The speed and power of modern computers makes this criticism somewhat redundant, so the technique merits further consideration. Corsten and Denis (1990) have also employed a distribution based test in their simultaneous two-way clustering.

A simpler method was given by Baril *et al.* (1994) who introduced the term 'mean-square decreasing method' in their two-way classification. If this strategy was converted to a one-way clustering scenario, it would lead to continuation of clustering while the change in the explained sum of squares exceeds the overall mean-square of the entire data set. Baril *et al.* (1994) use a plot of explained sum of squares versus degrees of freedom to show the value of each step in the clustering process. Clustering was stopped when the tangent of a pair of consecutive points on this plot was parallel to the line joining the first and last points. This idea works well because a clustering process generally explains the sum of squares faster than it expends degrees of freedom at first, but this benefit diminishes throughout the process.

Application to the principal data

The principal data of this investigation is so large (400×123) that some effort to reduce the number of comparisons that need to be made to answer the principal research question should be given priority. The ability of cluster analysis to determine groups of genotypes with similar specific adaptation can assist in meeting the ultimate objective of finding genotypes that suit tropical and subtropical locations. Cluster analysis is therefore potentially useful, despite the current paucity of methodology appropriate for handling incomplete data. An opportunity exists to develop new cluster analysis methodology capable of handling incomplete data.

Clustering of either (or both) genotypes and environments needs to be performed using a distance measure based on similarity of G×E interaction, so that interaction-free subsets of the original data can be found, thereby allowing simple comparisons based on mean performance. The distance measure of Ouyang *et al.* (1995), while capable of handling incomplete data, does not achieve this aim as it confounds main effect and interaction. Distance measures capable of handling missing data are introduced in Chapter 4.

Finding an appropriate distance measure for incomplete data will make cluster analysis

a viable option for analysing the principal data. Cluster analysis was felt to show the greatest potential of the options presented in this chapter for analysing an incomplete G×E matrix. Many new developments in the use of distance measures for clustering of incomplete data, first presented in Godfrey *et al.* (2001), are discussed in greater detail in Chapters 4 and 5.

2.7 Stability measures

Stability measures have been employed in G×E analyses to rank genotypes according to different criteria. The various stability parameters have been repeatedly examined with the most notable contribution being that of Lin *et al.* (1986). In that paper many stability parameters were classified into stability types, and the current examination is guided by that framework.

Lin *et al.* (1986) determined that a genotype was stable if:

1. It had small among-environment variance.
2. Its response was parallel to the mean of the genotypes in each trial.
3. The residual MS attributable to that genotype in a joint regression was small.

Lin and Binns (1988b) presented a fourth type of stability that considers the difference between predictable and unpredictable environment variation and the effect that this has on each genotype.

Freeman (1973) provided the first real summary of the stability parameters, noting the Wricke ecovalence, Shukla's parameter, and those found from the various joint regression models. All of these and other stability parameters were classified into the three types of stability by Lin *et al.* (1986) and can be found in Table 2.1. The discussion that follows focuses on the type of stability and includes other stability parameters where appropriate. These additional parameters are listed in Table 2.2.

Type 1 stability — Static performance

A genotype with type 1 stability has constant performance across environments and therefore has a low genotype variance. Type 1 stability is synonymous with the notion of static performance (Becker and Leon, 1988), and was discussed in relation to the joint regression model of Section 2.3. A genotype deemed stable under this type of stability does not benefit from the addition of environmental conditions.

This type of stability does not explicitly consider any differences between environmental main effect and G×E interaction. Lin *et al.* (1986) noted that

> "..., the usefulness of type 1 stability depends very much on the range of environments under which the experiment is conducted. If the range is very

Type	Equation	Authors or Users
1	$s_i^2 = \sum_{k=1}^{K}(y_{ik} - \bar{y}_{i\cdot})^2/(K-1)$	
1	$CV_i = s_i/\bar{y}_{i\cdot} \times 100$	Francis and Kannenberg, 1978
2	$\theta_i = \frac{I}{2(I-1)(K-1)}\sum_{k=1}^{K}(y_{ik} - \bar{y}_{i\cdot} - \bar{y}_{\cdot k} + \bar{y}_{\cdot\cdot})^2 + \frac{SS(GE)\dagger}{2(I-1)(K-1)}$	Plaisted and Peterson, 1959
	where, $SS(GE) = \sum_{i=1}^{I}\sum_{k=1}^{K}(y_{ik} - \bar{y}_{i\cdot} - \bar{y}_{\cdot k} + \bar{y}_{\cdot\cdot})^2$	
2	$\theta_{(i)} = \frac{-I}{(I-1)(I-2)(K-1)}\sum_{k=1}^{K}(y_{ik} - \bar{y}_{i\cdot} - \bar{y}_{\cdot k} + \bar{y}_{\cdot\cdot})^2 + \frac{SS(GE)}{(I-2)(K-1)}$	Plaisted, 1960
2	$\hat{W}_i = \sum_{k=1}^{K}(y_{ik} - \bar{y}_{i\cdot} - \bar{y}_{\cdot k} + \bar{y}_{\cdot\cdot})^2$	Wricke, 1962
2	$\hat{\sigma}_i^2 = \frac{I}{(I-2)(K-1)}\sum_{k=1}^{K}(y_{ik} - \bar{y}_{i\cdot} - \bar{y}_{\cdot k} + \bar{y}_{\cdot\cdot})^2 - \frac{SS(GE)}{(I-1)(I-2)(K-1)}$	Shukla, 1972
1 or 2	$b_i = \sum_{k=1}^{K}(y_{ik} - \bar{y}_{i\cdot})(\bar{y}_{\cdot k} - \bar{y}_{\cdot\cdot})/\sum_{k=1}^{K}(\bar{y}_{\cdot k} - \bar{y}_{\cdot\cdot})^2$	Finlay and Wilkinson, 1963
1 or 2	$\beta_i = \frac{\sum_{k=1}^{K}(y_{ik} - \bar{y}_{i\cdot} - \bar{y}_{\cdot k} + \bar{y}_{\cdot\cdot})}{\sum_{k=1}^{K}(\bar{y}_{\cdot k} - \bar{y}_{\cdot\cdot})^2}$	Perkins and Jinks, 1968
3	$\delta_i^2 = \frac{1}{(K-2)}\left[\sum_{k=1}^{K}(y_{ik} - \bar{y}_{i\cdot})^2 - \beta_i^2\sum_{k=1}^{K}(\bar{y}_{\cdot k} - \bar{y}_{\cdot\cdot})^2\right]$	Eberhart and Russel, 1966
3	$\delta_i^2 = \frac{1}{(K-2)}\left[\sum_{k=1}^{K}(y_{ik} - \bar{y}_{i\cdot} - \bar{y}_{\cdot k} + \bar{y}_{\cdot\cdot})^2 - \beta_i^2\sum_{k=1}^{K}(\bar{y}_{\cdot k} - \bar{y}_{\cdot\cdot})^2\right]$	Perkins and Jinks, 1968
	$\dagger\ SS(GE) = \sum_{i=1}^{I}\sum_{k=1}^{K}(y_{ik} - \bar{y}_{i\cdot} - \bar{y}_{\cdot k} + \bar{y}_{\cdot\cdot})^2$	

Table 2.1: Summary of stability statistics adapted from Lin *et al.* (1986). References for these measures can be found in the accompanying discussion or from Lin *et al.* (1986).

Type	Equation	Authors or Users
2	$P\left(y_{ik} > \bar{y}_{\cdot k}\right) \times 100\%$	St-Pierre *et al.*, 1967
2	$p_i = \sum_{k=1}^{K} \frac{\left(y_{ik} - m_k\right)^2}{2K}$	Lin and Binns, 1988
1	$s_i^{(1)} = \frac{2}{K\left(K-1\right)} \sum_{k=1}^{K-1} \sum_{k'=k+1}^{K} \lvert r_{ik} - r_{ik'} \rvert$	Hühn and Nassar, 1989
1	$s_i^{(2)} = \sum_{k=1}^{K} \frac{\left(r_{ik} - \bar{r}_{i\cdot}\right)^2}{K-1}$	Hühn and Nassar, 1989
2	$s_i^{(3)} = \frac{\sum_{k=1}^{K} \left(r_{ik} - \bar{r}_{i\cdot}\right)^2}{\bar{r}_{i\cdot}}$	Hühn, 1979
2	$s_i^{(6)} = \frac{\sum_{k=1}^{K} \lvert r_{ik} - \bar{r}_{i\cdot} \rvert}{\bar{r}_{i\cdot}}$	Hühn, 1979

Table 2.2: Summary of some stability statistics not presented by Lin *et al.* (1986). References for these measures can be found in the accompanying discussion.

> large, such as a collection of sites from a continental area, type 1 stability may not be very meaningful, but if the geographical range can be restricted, then it could be important."

It is likely therefore that type 1 stability will be useful in finding stable genotypes in the principal data.

Francis and Kannenberg (1978) gauged genotypes by their mean yield and the coefficient of variation CV_i for genotypes. Desirable genotypes had high mean yield and low coefficient of variation.

Although this measure was considered to be type 1 by Lin *et al.* (1986), empirical evidence suggests that the CV_i does not provide a stability parameter meeting the definition of type 1 (Flores *et al.*, 1998, Jalaluddin and Harrison, 1993). Lin *et al.* (1986) claimed that the use of the coefficient of variation was effectively the same as investigating the within genotype variance except for some minor data transformations before calculations were done, as analysis of proportional changes is equivalent to analysis of logarithms. Pham and Kang (1988) refuted this with an example.

Type 2 stability — Dynamic performance

A genotype with type 2 stability has an expected performance across environments that is parallel to the environment mean, and therefore has a low G×E interaction. Type 2 stability is synonymous with the notion of dynamic performance (Becker and Leon, 1988),

and was also discussed in relation to the joint regression model of Section 2.3.

The limitation on the use of type 2 stability measures was addressed by Lin *et al.* (1986) who stated:

> "Type 2 stability is a relative measure depending on the genotypes included in the test so its scope of inference is confined to the test set and should not be generalized. ... A genotype stable by this definition is so only with respect to the other genotypes in the test, without any assurance that it will appear stable if assessed with another set of genotypes."

Recognition of this concern allows the application of type 2 stability to the principal data of this investigation.

St-Pierre *et al.* (1967) used the proportion of the environments in which the genotype exceeded the mean performance of all genotypes in an environment as a measure of wide adaptation. It can be expressed as

$$P(y_{ik} > \bar{y}_{\cdot k}) \times 100\% \tag{2.17}$$

where $P(\cdot)$ is the probability operator, and the stability parameter is a percentage.

Wricke's ecovalence was presented by Lin *et al.* (1986) and Kang *et al.* (1987) as:

$$\hat{W}_i = \sum_{k=1}^{K} (y_{ik} - \bar{y}_{i\cdot} - \bar{y}_{\cdot k} + \bar{y}_{\cdot\cdot})^2. \tag{2.18}$$

It measures the contribution of each genotype to the overall sum of squares for the G×E interaction. It was defined by Lin *et al.* (1986) as the sum of squared G×E interaction effects across all environments. It can therefore be re-defined as the squared residuals for each genotype, from fitting the additive model in (2.1).

Shukla (1972) created a stability variance parameter for comparing genotypes by separating the G×E interaction into components for each genotype, and expressed as

$$\hat{\sigma}_i^2 = \frac{I}{(I-2)(K-1)} \sum_{k=1}^{K} (y_{ik} - \bar{y}_{i\cdot} - \bar{y}_{\cdot k} + \bar{y}_{\cdot\cdot})^2 - \frac{\sum_{i=1}^{I}\sum_{k=1}^{K} (y_{ik} - \bar{y}_{i\cdot} - \bar{y}_{\cdot k} + \bar{y}_{\cdot\cdot})^2}{(I-1)(I-2)(K-1)} \tag{2.19}$$

This stability measure was defined by Lin *et al.* (1986) as "the variance of a genotype across environments".

The precision of this parameter has been questioned when few environments have been used in the analysis, and in particular, when there are less than ten environments (Kang and Pham, 1991). Piepho (1994b) showed that it is possible to estimate Shukla's stability parameter in data sets with missing cells. Unfortunately he also noted that the two methods shown may fail if there are too many missing observations.

Piepho (1994b) suggests splitting the data to get subsets of the original data where parameters are estimable. Piepho cautions using either of the approaches if the data is sparse.

Lin *et al.* (1986) and Kang *et al.* (1987) noted that Wricke's ecovalence and Shukla's stability variance were rank correlated, while Lin *et al.* (1986) added that these were also rank correlated with the earlier work of Plaisted and Peterson in 1959 and Plaisted in 1960. Kang *et al.* (1987) found that Shukla's parameter $\hat{\sigma}_i^2$ was a coding of, and should be used in preference to Wricke's ecovalence $\hat{W}_i$ when investigating stability. They categorically stated that only one of $\hat{\sigma}_i^2$ and $\hat{W}_i$ need be calculated when partitioning G×E interactions.

Lin *et al.* (1986) put forward the idea that joint regression coefficients are more preferable to use than the stability variance of Shukla and others, as they give an indication of "the shape of the response as well as its variation." These parameters can only be used when the joint regression model is appropriate, as discussed in Section 2.3.

Lin and Binns (1988a) introduced the stability measure

$$p_i = \sum_{k=1}^{K} \frac{(y_{ik} - m_k)^2}{2K} \tag{2.20}$$

where y_{ik} is the yield of the ith genotype in the kth location and m_k is the maximum yield in the kth location. The lower this value the more superior the genotype is supposed to be in terms of stability for general adaptability. They state

> "If the selection is based solely on p_i, a narrowly adapted cultivar, i.e., poor in general adaptability but good in specific adaptability, may be discarded. This is avoided by computing a pair-wise GE interaction mean square between the maximum response and each test cultivar. If the mean square is not substantially larger than the estimated error, it implies parallelism of both responses (i.e. the differences from the maximum responses are about the same for all locations). Under such circumstances, the p_i value is an appropriate indicator of superiority. On the other hand, a large GE interaction implies differences in the response patterns and comparison by p_i values becomes meaningless. Then, a breeder should examine the specific adaptability of that particular cultivar individually by plotting the observed values of both the maximum response and the candidate cultivar on the location mean. The closeness of the observed values of the candidate cultivar and the maximum response indicates areas of specific adaptability."

The relevance of this measure to analyzing the principal data of this investigation was summarized by Lin and Binns (1991a) who noted:

> "..., the merit of the present method becomes more apparent as the geographical area covered by the test sites increases in scope. For example, if

> the test sites are chosen from a county within a state, selection by cultivar means or by type 1 stability may be adequate, but if the test sites are chosen from a large area such as a national or international trials, selection by these conventional parameters are not meaningful. Under such circumstances, the utility of the present method becomes significant."

The drawback of the method is that it does not explicitly cater for incomplete data. This has been addressed in Section 3.7 so that it can be applied to the principal data.

Type 3 stability — predictable performance

A genotype with type 3 stability is expected to have small deviations from its predicted yield in each environment. The problem with the use of this stability measure is that it indicates a lack of fit of the model chosen to extract the predicted performances. It has been used in conjunction with the joint regression model by Eberhart and Russell (1966) and Perkins and Jinks (1968), but will not be considered for use in this investigation.

Type 4 stability — interaction of genotype with unpredictable environment variation

A genotype with type 4 stability has predictable performance at each location across seasons. Kang (1998) termed this type, 'temporal' stability, and noted that it "is desired by and beneficial to growers". Kang (1998) compared this notion to 'stability' (Type 1 stability) which is "beneficial to seed companies and breeders". Lin and Binns (1989) noted that "In contrast to the three types of conventional stability parameters discussed by Lin *et al.* (1986), type 4 stability has the advantage of being independent of other cultivars included in the test and being a homeostatic characteristic of the cultivar coping with unpredictable variation." Lin and Binns (1991b) remarked

> "The difference between type 1 and type 4 is that the former (a simple variance estimate or CV of the genotype across locations) measures the homeostatic property in terms of overall environmental variation, while the latter (the year within location MS averaged over locations) measures it only with respect to unpredictable variation, excluding the part (predictable) that is controllable. Thus, type 4 resolves one of the problems of type 1, namely, its impracticality ... Indeed, the strength of type 4 is that this parameter is not limited to the range of sites included in the test."

Lin and Binns (1988b) denote the interaction of location and year (defined as an environment) in earlier examples as able to be broken into a fixed component (location) and a random component (year within location) so that type 4 stability is the result. Type 4 stability can therefore be found for each genotype in the set and compared.

Piepho (1994d) provided a method for breaking the G×E interaction into three components for G×Y, G×L, and G×L×Y combinations. His aim was to weight these three components equally for analysis. The unbalanced nature of the principal data of this investigation precludes use of this type of stability.

Nonparametric stability measures

Nonparametric stability measures use ranks instead of raw data to gauge genotype stability.

Hühn and Nassar (1989) presented two non-parametric measures of stability and provided statistical tests for their significance. The first measure, found using

$$s_i^{(1)} = \frac{2}{K(K-1)} \sum_{k=1}^{K-1} \sum_{k'=k+1}^{K} |r_{ik} - r_{ik'}| \tag{2.21}$$

is the average absolute rank difference over the K environments. Their second measure

$$s_i^{(2)} = \sum_{k=1}^{K} \frac{(r_{ik} - \bar{r}_{i\cdot})^2}{K-1} \tag{2.22}$$

where $\bar{r}_{i\cdot}$ is the average rank of genotype i, gives the variance of ranks among environments. For maximum stability, both of these measures would be near zero. Becker and Leon (1988) note that these two parameters are likely to be correlated, while Flores *et al.* (1998) stated, using empirical evidence, that they are based on the static notion of stability. They have therefore been classified as type 1 stability measures.

Becker and Leon (1988) reported two nonparametric measures due to Hühn in 1979. These were

$$s_i^{(3)} = \frac{\sum_{k=1}^{K} (r_{ik} - \bar{r}_{i\cdot})^2}{\bar{r}_{i\cdot}} \tag{2.23}$$

and

$$s_i^{(6)} = \frac{\sum_{k=1}^{K} |r_{ik} - \bar{r}_{i\cdot}|}{\bar{r}_{i\cdot}} \tag{2.24}$$

where $\bar{r}_{i\cdot}$ is the mean of the ranks for genotype i across all environments. Kang and Pham (1991) noted that these statistics have not been used a great deal in the G×E literature. Flores *et al.* (1998) asserted that $s_i^{(3)}$ should only be used when the expected heterogeneity of genotypes and/or environments is large, and only when studying the stability of performances.

Kang and Pham (1991) noted that both of these statistics showed significant correlation with Wricke's ecovalence, the deviation from regression (Eberhart and Russell, 1966), and p_i (Lin and Binns, 1988a). On this basis they have been classified as type 2 stability

measures.

Mixing yield and stability

Flores *et al.* (1998) determined empirically that many stability measures actually mix yield performances with stability, including $s_i^{(6)}$. Kang *et al.* (1991) presented the rank sum approach for equally weighted selection on both yield and stability. This measure was created by adding the rank of the mean yield (highest mean yield has rank one) to the rank for Shukla's stability variance (ranked lowest to highest). Thus the lowest possible rank sum of two would arise if the highest mean yielding genotype was also the most stable according to the Shukla parameter. Although this measure combines both yield and stability, Kang and Pham (1991) showed that this was not the case for two of the five data sets they used.

Application to the principal data

The structure of the principal data does not lend itself to use with stability measures due to its incompleteness. It should be possible, however, to adjust some stability measures so that the incompleteness is taken into account. Type 4 stability measures are dependent on data being balanced over a genotype×location× year structure, and if these data were available type 4 stability would be employed. The type 3 stability measures will not be used in this investigation for reasons explained, but the type 1 and 2 measures will be given further consideration in Section 3.7.

Because Wricke's ecovalence is easier to calculate and is equivalent to Shukla's stability parameter, it is preferred for use with the principal data. Appropriate averaging is considered for adjusting Wricke's ecovalence. The Lin and Binns (1988a) superiority score can also be adjusted using appropriate averaging. Both these measures are considered in Section 3.7.

The incompleteness of the principal data makes comparison of within-environment ranks meaningless. If a suitable means of making the principal data complete is employed, the nonparametric stability measures will become applicable to the principal data.

2.8 Other models and methods

This section describes methods that do not fall into the categories presented in other sections of this chapter. A mixture of graphical and numeric approaches are reviewed, which range in complexity from simple categorization to diagrams requiring more detailed inspection.

Flores *et al.* (1998) reported use of a simple approach that counted the number of environments in which a genotype performed in each third of the environmental ranges. This extends the stability measure of St-Pierre *et al.* (1967) to give three numeric summaries

for each genotype. Genotypes usually placed in the top third are considered relatively well adapted and stable. This method would be of greater use in regional trials rather than an international set of trials where the environments are more diverse.

Many standard graphical techniques for analyzing multivariate data have been employed in G×E research, including minimal spanning trees (Westcott, 1987) and multidimensional scaling (Basford, 1982). Goodchild and Boyd (1975) used principle components analysis to model climatic and edaphic conditions for a number of wheat growing areas in Western Australia. They were able to reify the first two PC axes geographically with use of maps and shading of four groupings of shires (small administrative units). A similar analysis was presented by Seif *et al.* (1979) for the New South Wales wheat belt finding similar results, but a different cause (namely differing altitudes) was deemed to be significant.

Manly (1994) presented the use of 'star' diagrams to show multi-dimensional data on a two-dimensional plot. Each variable is given a ray emanating from a central point for each observation. The length of the rays indicates the magnitude of the value of the observation in the variable. Flores *et al.* (1998) reported that this method has been employed in G×E research. The largest and most regular shaped polygon of the endpoints of the rays is the best performing genotype in terms of both yield and stability.

Flores *et al.* (1998) also reported the use of plots of means and standard deviations of genotype yields as well as the within-environment ranks of yields. Genotype selections could then be based on maximizing the mean yield and minimizing the genotype standard deviation. The use of rank data rather than the original yields would reduce the influence of heterogeneous environment variance.

The 'safety first' approach of Eskridge (1990) found the αth percentile of the distribution of a genotype's expected response. They considered different criteria for genotype comparisons, including: Shukla's stability variance, the joint regression slope parameter, the deviation sum of squares parameter of Eberhart and Russell (1966), and the general genotype variance across environments. As its name suggests, this is a very conservative approach, that is dependent on the value of α chosen by the researcher.

Another extension of the stability parameter of St-Pierre *et al.* (1967) was presented by Eskridge *et al.* (1993) who compared a genotype's performance to a check cultivar across environments using a reliability function. The principal aim is to find the function for each genotype that describes the probability of exceeding the check cultivar's performance by a certain level d. This function is expressed as

$$R_i(d) = P((y_i - y_c) > d) \tag{2.25}$$

where the R_i is a function of d, and is found for each of the i genotypes. The yield of a check variety is denoted y_c in this expression. These functions can be graphed as

step functions, resulting in stable genotypes having steeper step functions than less stable genotypes. The approach is limited to comparisons between the test genotypes and the check cultivars, and thus only between genotypes in the programme. There is no real need however for all genotypes to be grown in exactly the same environments if environments are relatively homogeneous.

Menz (1980) used stochastic dominance as a means of determining which genotypes were risk efficient. Such genotypes would perform well in low yielding environments and have the possibility of good yields in high yielding environments. Use of this method in conjunction with the testing of mean yields could provide a means of selecting genotypes. This author asserted that the high mean yield and risk efficiency would almost certainly mean that the genotype had good wide adaptability, and that this criterion would best suit risk averse growers. Menz (1980) expressed the rules of stochastic dominance as:

1. A grower prefers a higher return.

2. A grower is risk averse.

3. A grower's risk aversion increases with risk.

As with the safety first method of Eskridge (1990), this method is conservative and is best suited to regional trials rather than a set of international trials.

The notion of finding genotypes whose performances dominate the performance of other genotypes was put into a clustering framework by Moro and Denis (1997). They used a clustering method based on the inferiority of genotypes to one another. The distance measure employed was asymmetric and can be expressed as

$$D_{ij}^{(M\&D)} = \frac{\sum_{k=1}^{K}(\min\{0,(y_{ik}-y_{jk})\})^2}{\sigma^2} \tag{2.26}$$

where σ^2 is the variance of y_{ik} which was assumed to be $N(\mu_{ik},\sigma^2)$. The ith genotype in the pairing is known as the 'leader' and the jth genotype is the 'dominated' genotype.

In simple terms, genotypes were clustered by selecting the pair whose $D_{ij}^{(M\&D)}$ was the minimum of those considered. Once a pair of genotypes were clustered, all $D_{ij}^{(M\&D)}$'s relating to the jth (inferior) genotype were discarded, and the clustering process continued until all genotypes were in the same cluster. When clustering is truncated, genotypes are not necessarily similar to the other genotypes clustered with it. All the genotypes that are deemed inferior are to be ignored, and it is this fact that separates this method from the standard cluster analyses based on symmetric inter-genotype distances. This method can be adapted for incomplete data, and will not be limited by having an incomplete distance matrix. Unlike other clustering methods, this method does not necessarily find groups of genotypes that perform similarly across environments. The procedure should, however,

find a reduced set of genotypes that need to be compared.

When the full AMMI model is fitted, that is where N is equal to the lower of I or K in (2.12), the total G×E interaction effect is explained by the model. This model has been termed the 'cell means' or 'treatment means' model, and is effectively the mean over replicates of each G×E combination. Piepho (1994c) differentiated these values which he called 'best linear unbiased estimator' or 'BLUE' for short, from 'best linear unbiased predictors' (BLUP). BLUPs are found by choosing genotype, environmental or both main effects as random rather than fixed terms in a model.

Piepho (1994c) asserted that in most cases the genotypes can be considered a random effect, as they form a random set of existing genotypes. Some discussion of the fixed or random nature of the effects being modelled in G×E analyses has been presented (Cochran and Cox 1957; Lin and Binns, 1994). The principal data of this investigation cover a wide range of international locations and it was decided that in keeping with the advice of Lin and Binns (1994), the environmental main effect would be fixed if this technique was applied. The model below is for fixed environmental effects, but Piepho (1994c) also gave the models for all combinations of fixed and random effects.

When comparing the BLUP results to those found using AMMI models, Piepho (1994c) fitted the BLUP of the 'true' yields using the model

$$\text{BLUP}\,(y_{ik}) \;=\; \bar{y}_{\cdot k} + w_G^2\,(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}) + w_{GE}^2\,(y_{ik} - \bar{y}_{i\cdot} - \bar{y}_{\cdot k} + \bar{y}_{\cdot\cdot}) \tag{2.27}$$

where w_G^2 was defined as the 'repeatability of estimated genetic effect' and w_{GE}^2 as the 'repeatability of estimated interaction'. These weights are based on the variances of residuals σ^2, the genotype main effect parameters σ_G^2, and G×E interaction parameters σ_{GE}^2, and are found using

$$w_G^2 \;=\; \frac{\sigma_{GE}^2 + K\sigma_G^2}{\sigma_{GE}^2 + \sigma^2 + K\sigma_G^2} \tag{2.28}$$

and

$$w_{GE}^2 \;=\; \frac{\sigma_{GE}^2}{\sigma_{GE}^2 + \sigma^2} \tag{2.29}$$

respectively. When the variance of residuals is zero ($\sigma^2 = 0$) the BLUP is identical to the cell mean, because the weights w_G and w_{GE} reduce to one. These weights have the effect of reducing the deviation from the environmental mean of each G×E combination because $1 \geqslant w_G^2 \geqslant w_{GE}^2 \geqslant 0$. Piepho (1994c) used the term 'shrinkage' to describe the impact these weights had in moving "estimated genotypic and interaction effects towards their zero mean whenever $\sigma^2 > 0$".

As a result of the comparisons between BLUP and AMMI, Piepho (1994c) commented that:

1. The choice of whether to use BLUP or AMMI should be considered on a case-by-

case basis. Cross-validation should be used in both BLUP and AMMI analyses, and should be appropriate for the design of the trials being analysed.

2. AMMI provides a better means of interpreting G×E interaction patterns than BLUP.

3. AMMI also allows the imputation of missing values, while BLUP provides predictors for observed G×E combinations only, but consideration could be given to use imputation via EM-AMMI prior to application of BLUP.

Piepho (1994c) also commented on the applicability of the BLUP procedure when data is incomplete, and noted that the procedure can find BLUPs for only the G×E combinations that have data. This was disputed by van Eeuwijk (1995a) who noted that the BLUP procedure could be used to provide estimates of unobserved G×E combinations. He asserted that the unfitted parameter for a particular G×E interaction parameter could be replaced by its expected value (zero), so allowing missing observations to be predicted as the combination of the main effects found using BLUP.

Westcott (1987) provided a graphical approach for gauging genotype performances. The best genotypes were to be found around the edges of an ordination. He used a dissimilarity measure designed specifically to achieve this aim. The dissimilarity between two genotypes was found using

$$\sum_{k=1}^{K} \frac{1}{K} \frac{\max\{y_{\cdot k}\} - (y_{ik} + y_{jk})/2}{\max\{y_{\cdot k}\} - \min\{y_{\cdot k}\}} \tag{2.30}$$

for any pair of genotypes $i \neq j$. The dissimilarity matrix was completed using $1/K$ when $i = j$.

The major embellishment of this procedure given by Westcott (1987) was to calculate the dissimilarity over subsets of the K environments. The graphics were then created for a series of environment groupings. Flores *et al.* (1996) used the AMMI model and the data subsetting approach of Westcott (1987) in conjunction to establish which genotypes were stable in low performing environments and good performing in high yielding environments. This data subsetting was applied in two different ways: environments were sorted by yield and by exposure to a parasite.

Application to the principal data

Most of the methods described in this section are not applicable to the principal data of this investigation. The two limitations are the lack of replicate data, and the scope of the trials covered. The approaches that are based on stability assume environments to be more homogeneous than is expected for an international programme of trials.

The BLUP procedure of Piepho (1994c) requires replicate data, but the suggestion of van Eeuwijk (1995a) for imputing missing data in the BLUP procedure should be

considered further. The concern raised in Section 2.4 in relation to fitting a multiplicative model using additive imputed values does not matter in this instance. The predictions of missing G×E combinations would not be used until fitting of the model has concluded. The weights used for the shrinkage of genotype main effect and G×E interaction parameters cannot be estimated without an estimate of the error variance, so this method will also not be used in the analysis of the principal data.

2.9 Summary

This chapter presented the various modelling methods for G×E analyses. Their ability to be used when data are incomplete was investigated, and on the whole shown to be extremely limited.

Methods for determining the significance of G×E interaction were discussed in Section 2.2. The assumption was made that the heterogeneity of the environments used in the Onion Trials Programme means that G×E interaction exists and must be considered for any modelling approaches used.

Until such time as the principal data set is made complete, most of the methods in this chapter will not be considered any further. The three options that will be investigated are:

1. The EM-AMMI model (Gauch and Zobel, 1990).
2. Various stability measures.
3. Cluster analysis.

The Healy-Westmacott algorithm has been applied to fit a linear model to incomplete data. Gauch and Zobel (1990) used this algorithm with the additive main effects and multiplicative interaction (AMMI) model. Results from its application to the data arising from the Onion Trials Programme are presented in Section 3.8.

It is possible to adjust some of the stability measures introduced in Section 2.7 to allow for the incompleteness of the data. This was undertaken and is presented in Section 3.7 for the data arising from the Onion Trials Programme.

The other area of modelling that already has strategies for dealing with incomplete data is cluster analysis. New cluster analysis methods are developed in Chapters 4 and 5. These new developments are applied to the data arising from the Onion Trials Programme in Section 5.4. The next chapter presents a more detailed description of the Onion Trials Programme's history, and its introductory data analysis.

Chapter 3

The principal data

3.1 Introduction

This chapter completes the first part of the investigation by describing the Onion Trials Programme in greater detail. Its background is explained, and problems that arose in its analysis are exposed. In the next section, the manner in which the Trials Programme started and the way it has developed since 1990 are explained. Details of the data collected by collaborators are given, and the differences between the principal data and that arising from regional trials programmes are also discussed in Section 3.2.

The main focus of this investigation is towards analysing incomplete data caused by not testing all genotypes in all environments. The problems of missing data and sparsity are considered in Section 3.3. Two key definitions are also presented in this section that are used by researchers working with incomplete data. In Section 3.4, the entire data set is graphically analysed, while attempts to model onion yields using regression are covered in Section 3.5.

It was determined that some minimum criteria needed to be established to allow admission of results for genotypes and environments to be included in the G×E matrices to be analysed. The creation of the two data sets used throughout this investigation is described in Section 3.6, while Section 3.7 uses these data sets to show how current methodology can be adapted for use with incomplete data. The EM algorithm in conjunction with the additive main effect and multiplicative interaction (AMMI) model was applied to these two data sets and results are presented in Section 3.8

3.2 The trials programme

In 1990, Dr Lesley Currah was employed by the Natural Resources Institute (UK) to evaluate short-day onions in climates of the tropics. The Onion Trials Programme started when leftover seed from an initial trial in Zimbabwe was sent to other collaborative researchers with similar aims. As noted by Currah *et al.* (2001), "The trials programme

therefore began in rather a casual way, without any long-term plan, nor any idea of how it would increase in scale and duration." This conference paper, presented by Dr Currah, gave a short history of the Onion Trials Programme, and has proved invaluable for the following description of the programme, while Currah (2002a) provided a synopsis of the current state of development and research on short-day onions in the tropics. This synopsis showed that a project of the trials programme's magnitude had never previously been attempted for onions in the tropics.

The Onion Trials Programme started with a simple aim, "To gain a better understanding of the reactions of onions to as many types of climate as possible, through informal long-distance collaboration." (Currah *et al.*, 2001). The assistance of the many collaborators around the world over the years 1990 to 2000 was a crucial factor in the success of the trials programme. Without their support and commitment this project would not have continued for so long. Most collaborators were not able to commit resources in more than a couple of years, so there was little continuity of location use over years. For this reason, environments have been defined as the combination of location and year throughout this work.

The list of cultivars available for testing also varied over years, because it was dependent on donations from seed companies. The ongoing development and success of the programme can in part be attributed to this co-operation of seed distributors. For many collaborators the opportunity to select seed from a long list of previously inaccessible seed stock was appreciated. Over time some varieties became obsolete, while others (including hybrids) were developed. Collaborators were given at most thirty different varieties to test, while some were unable to test even ten varieties. The quantity of seed varied from variety to variety, and some stayed on the list for fewer seasons. It was little surprise to find that some of the more than 300 varieties were grown in only a handful of environments.

Apart from the costs of postage for delivery of seed, no direct financial commitment was incurred in seed distribution. No formal agreements were therefore needed between organizers and the individuals doing the work. Collaborators were asked to provide some minimal information, including:

1. Yield, both total and marketable,
2. Weather data,
3. Plot sizes,
4. Dates of sowing, transplanting, and harvest.

Collaborators with greater research experience provided organizers with more information than was requested, including:

1. Proportion of bulbs that bolted, split, or doubled.

2. Dry matter content and size of bulbs (sometimes important for marketing).

3. More detailed meteorological data.

On the other hand, approximately half of seed parcels distributed were wasted, as yield data were not provided by collaborators (Currah *et al.*, 2001). This wastage is from the organizers' perspective, as collaborators are assumed to have gained by extending their research experience (Currah, 2003).

Individual trial results for the years 1990 to 1995 were compiled into one volume by Currah *et al.* (1997); results from subsequent years are in preparation for publication in a similar volume. In this publication each trial is described individually, including trials based on total marketable yield, and related storage trials. No attempt was made to link the individual trials in this publication, but an initial investigation, for data from the years 1990 to 1998, was presented by Currah *et al.* (1999).

It must be remembered that the focus of this study was yield of onions, while in many circumstances cultivar selections have been and will be based on factors that are not directly related to yield (Currah, 2002b). Aspects of interest when selecting cultivars include:

1. Proportion of bulbs that bolt.

2. Colour and taste preferences for local consumption.

3. Export opportunities in crops that are not generally consumed locally.

4. Storage potential of varieties. "Factors such as high dry matter, good skin quality, and number of outer dry skins come into these decisions." (Currah, 2002b).

5. Size preference.

6. Disease and pest resistance.

7. Natural dormancy.

Currah (2002b) noted that these characteristics are often at odds with yield maximization; for example, higher yielding varieties tend to be low in dry matter content which limits their storage potential. Final variety selections will therefore not always be based on the yield performance alone. It is hoped by those involved in this project that storage trial results and other data can be incorporated into recommendations in the future.

Data from 123 admissible trials were amalgamated into one file by Dr Currah. Each G×E combination used one row of a large spreadsheet and had the following data recorded.

1. A code for each trial. This combination represents:

   (a) The name of the site where the trial took place.

(b) The name of the country in which this site is located.

(c) The initial sowing date of the trial.

2. Fixed covariate information for each environment.

   (a) The altitude of the location, measured in metres above sea level.

   (b) The latitude of the location, measured in degrees away from the equator; that is, the hemisphere is not recorded.

3. The commercial cultivar name of the genotype tested. This includes a two letter code indicating the company that supplied the seed.

4. The average yield per square metre of each genotype grown in a trial.

5. Covariate information recorded for each G×E combination:

   (a) The length of the growing period (in weeks).

   (b) The weekly minimum and maximum temperatures for the environment, recorded until the genotype was harvested. On many occasions these temperatures were estimated from monthly data.

   (c) The photoperiod, recorded at weekly intervals until the genotype was harvested.

Covariate information was not available for all G×E combinations. The temperature and photoperiod was only available in full for 101 of the 123 environments. Two reasons accounted for this: First, some trials were so recent that the covariate information had not become available, and second, because it was not accurately recorded by the collaborator. While G×E yields, geographic covariates, and environmental temperature data were collected by collaborators, the photoperiods were calculated using a formula based on the latitude and sowing date of each trial (Keisling, 1982), discussed in more detail below.

In most regional trials programmes it is common for the set of tested varieties to change over seasons. In each year of such programmes the current set of varieties is usually tested in all environments. The *ad hoc* nature of the Onion Trials Programme was a considerable departure from this type of programme.

Standard practice in such regional programmes is to test last year's winners against the current year's new varieties; the best of these varieties go on to the set that will be used the following year. There is usually a reasonable amount of overlap from one year to another which leads to a series of complete G×E analyses that can be linked over time. If these data sets are combined, the amount of sparsity contained would be moderate compared to the data from the Onion Trials Programme. Whereas regionally focused programme trials performed many years ago usually have less value in the decision making process than those conducted in recent years, this was not the case for the Onion Trials Programme.

Standard methodology for analysing the smaller, regionally focused programme that has near perfect consistency of experimental practice is therefore inappropriate for the Onion Trials Programme. The sparsity in this programme's data is at a much less manageable level, and more importantly, the amount of G×E interaction that occurs will be more significant in determining the best varieties.

Calculation of photoperiod

Keisling (1982) noted the necessity for a formula based method, as against interpolation of tabulated results, to calculate a location's photoperiod. The formula developed by Keisling (1982) takes account of the location's latitude, the time of year (measured as a Julian date), and the impact of indirect light from the sun below the horizon. The formula appears as,

$$\text{Photoperiod} = \frac{2}{15}\cos^{-1}\left[\cos\alpha\sec\phi\sec\delta - \tan\phi\tan\delta\right] \tag{3.1}$$

where α is the zenithal distance in degrees of the sun at the event of interest and directly accounts for the impact of indirect light, ϕ is the latitude in degrees (with the southern hemisphere latitudes recorded as negative values), and δ is the declination of the sun in degrees which is dependent on the Julian date.

Calculation of photoperiods is not a simple task; calculation of Julian dates requires the researcher to consider the impact of leap years for example. The task is time consuming, but use of an internet interface such as those provided at www.currah.co.uk, or the University of Hawaii's site, makes it less arduous. As an example, the photoperiods for the week from the 24th October 2003 to the 30th in Palmerston North New Zealand are 13.6, 13.7, 13.7, 13.7, 13.8, 13.8, and 13.9 hours. A trial located at Massey University would therefore have an average photoperiod over that week of 13.76 hours.

Further covariate information

Further covariate information has been made available in a handbook of the trials (Currah *et al.*, 1997), but is not included as part of the data set in this study. This additional environmental and genotypic information is not yet comprehensive or complete, but includes:

1. The colour of onion varieties.
2. The size and shape of the average onion of each variety.
3. The taste aspects of varieties.
4. The experimental design used at sites.
5. The spacing and arrangement of plants at sites.
6. The plot size at the sites.

7. The method and amount of irrigation at sites.

8. Details of fertilizer and pesticide administration.

9. Weather detail, such as monthly rainfall, temperatures, and relative humidity.

10. Weed management techniques.

11. Soil types.

A complete list of the 123 environments with their fixed covariate information and data for the 400 genotypes used in this study are provided on the CD-ROM accompanying this volume.

3.3 The problem of missing data and sparsity

As identified in the previous section, collaborators were in general able to test less than 10% of onion varieties included in the Onion Trials Programme. When the G×E matrix of yields in kilograms per square metre was formed, less than 2,400 of the possible 49,200 G×E combinations had been tested. Reasons for these G×E combinations not being tested included:

1. Inability of collaborators to grow all varieties through lack of time and physical space.

2. Restricted quantities of donated seed.

3. Specific selection of test varieties by collaborators to meet their local market needs; for example, a collaborator that chose to grow only red onion varieties.

4. Unavailability of varieties in certain years, whether by virtue of their obsolescence or their lack of development.

5. The belief that a certain combination would not result in marketable onions.

Regardless of the cause, it was clear that there were many untested combinations. The data is therefore called 'sparse'.

Statisticians need to define the nature of any missing data so that its impact can be appropriately incorporated into population parameter estimation. Little and Rubin (1987) presented the two most commonly used concepts:

Missing completely at random (MCAR)	If the probability that an observation is missing does not depend on the variables under study, they are said to be 'missing completely at random'. In this instance the missingness is independent of the values of both response and explanatory variables. The missing data is then assumed to be a simple random sample from the entire population.
Missing at random (MAR)	If the probability that an observation is missing does not depend on the expected value of the observation, then it is said to be 'missing at random'. In this instance the missingness is independent of the values of response variables, but can be dependent on the values of explanatory variables. Missing completely at random is therefore a special case of missing at random.

These definitions (constructed for more common levels of missing data) provide an initial framework for how the missingness will affect the development of solutions to overcome the sparsity problem.

The data collected over the Onion Trials Programme is not a simple random sample of the total possible data because of the constraints described above. It therefore breaks the 'missing completely at random' condition, but may be classified as 'missing at random' because there is no reason to suggest that the cause of missing data is related to the expected value of observed data.

While variety selections for each location were made by the organizers and collaborators in conjunction, there is no *a priori* reason to suggest that the cause for them was their probable success. The omission of some tested G×E combinations which showed a lack of success (i.e. varieties failed, or produced negligible marketable yield), would have caused the data to break the missing at random condition. In fact, if it is known with certainty, possibly from previous experience, that selection of a particular variety at a particular environment would result in a failure, this knowledge should be included in the data.

Rubin (1976) observed that the missing at random data problem can be easily dealt with. However if the process causing the data to be missing is definable, he suggested the parameter estimates found using present data, and therefore the inferences made, would be unsound unless this process is accounted for in the hypothesis. In this study, the nature of the missingness can be identified and therefore the effects of variety selection on parameter estimates need to be considered. For this reason a simple examination of the mean performance of either genotypes or environments would be erroneous. Methods for dealing with incomplete block designs could be applied to G×E data provided no G×E interaction exists, in which case, genotypes and environments can be used as treatments and blocks respectively in an incomplete block analysis. These methods would provide a set of means that take account of the missingness, and thus be comparable.

The existence of G×E interaction complicates the analysis of sparse data. When few entries in the G×E matrix are missing, and a model that assumes normally distributed errors is used, the G×E interaction component for these entries can be estimated to be zero (the expected value of interaction terms in the model). The performance of a missing G×E combination would be estimated using only the main effects of genotype and environment under an incomplete block analysis. The danger of doing this is that the main effects are dependent on the subset of environments in which genotypes were grown, and parameter estimates are therefore prone to error; a genotype tested in environments that favoured it, would appear to be successful everywhere if such a model was used.

In contexts where the amount of G×E interaction is low compared to the main effects for genotype and environment, this practice will not be too risky in terms of making the wrong decision. In a context like the Onion Trials Programme, however, the interaction is likely to play a significant role in determining which varieties are recommended for certain environments. In this study, a solution to overcome the effects of the sparsity must therefore consider the impact of G×E interaction.

The sparse data from the Onion Trials Programme can be analyzed in two ways. Either current methodology is adapted to take account of the missing data, or some way of imputing missing data is found to allow standard analyses to occur. The remainder of this chapter shows the limitations of the first of these options, while the second option is considered in detail in subsequent chapters.

3.4 Initial data analysis.

One of the first tasks to undertake in any project, especially one of this size, is an initial data analysis. This exploration uncovered irregularities in the data and led to some data-cleaning. Because the G×E matrix is sparse, and available covariate information is incomplete, the figures presented in this section often use different subsets of the entire data on file. In the majority of these figures, the maximum amount of data usable was used. Each figure should be considered independently, although broad generalizations are possible. It is clear that some determination of what data should be used in the remainder of this work is necessary; this is presented in Section 3.6. The initial data analysis presented in this section is broken into two parts. First the yield data specific to the G×E matrix is considered, and second the covariate information is linked to the yield data available.

Looking at yields from the G×E matrix

Mean yields, standard deviations and the number of times genotypes and environments were used in the Onion Trials Programme are plotted against one another in Figures 3.1 and 3.2 respectively. The top panels of these figures show evidence that the standard

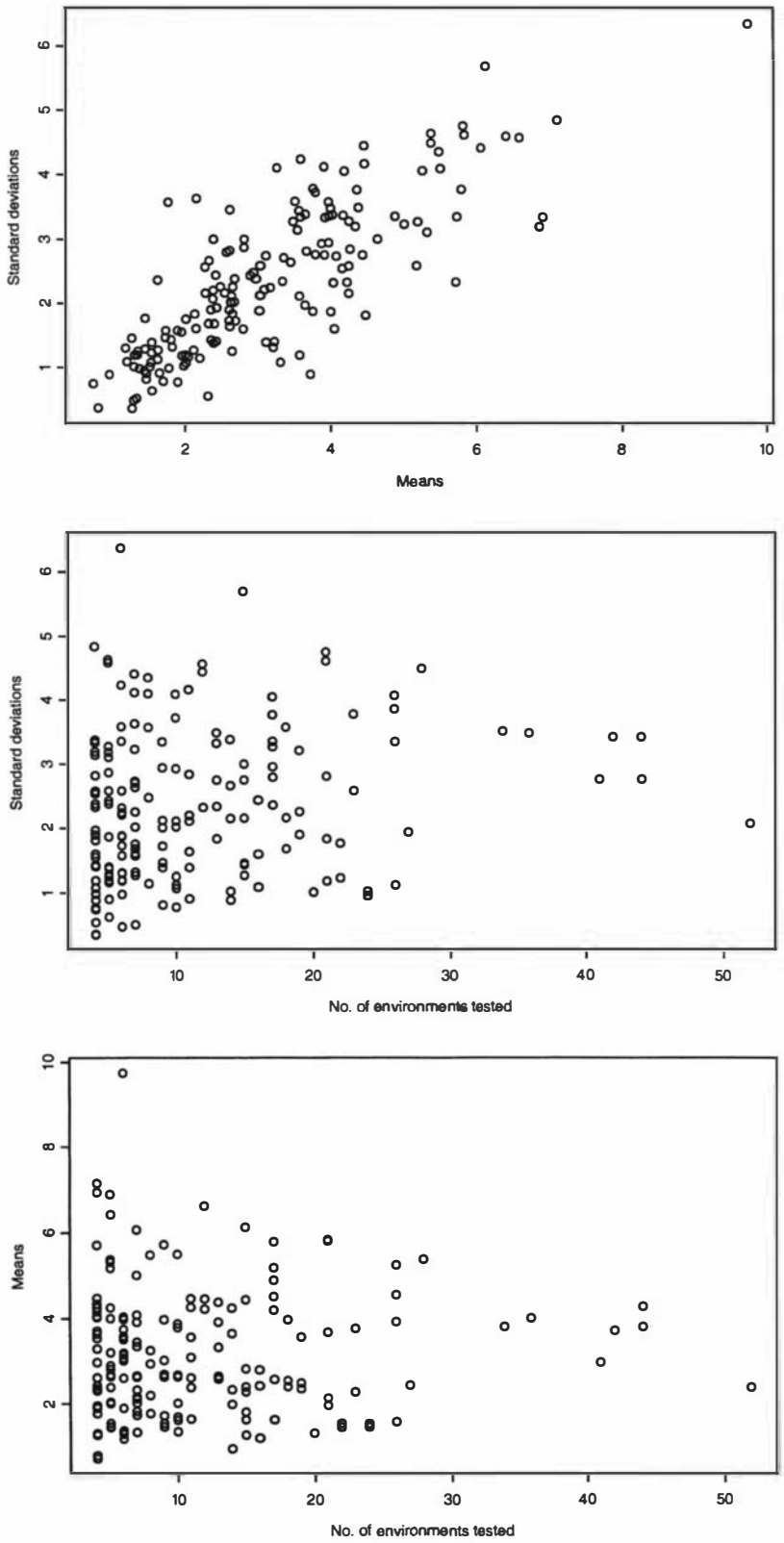


Figure 3.1: Genotype means, standard deviations, and the number of times each genotype was used plotted against one another. Means and standard deviations are measured in kilograms per square metre.

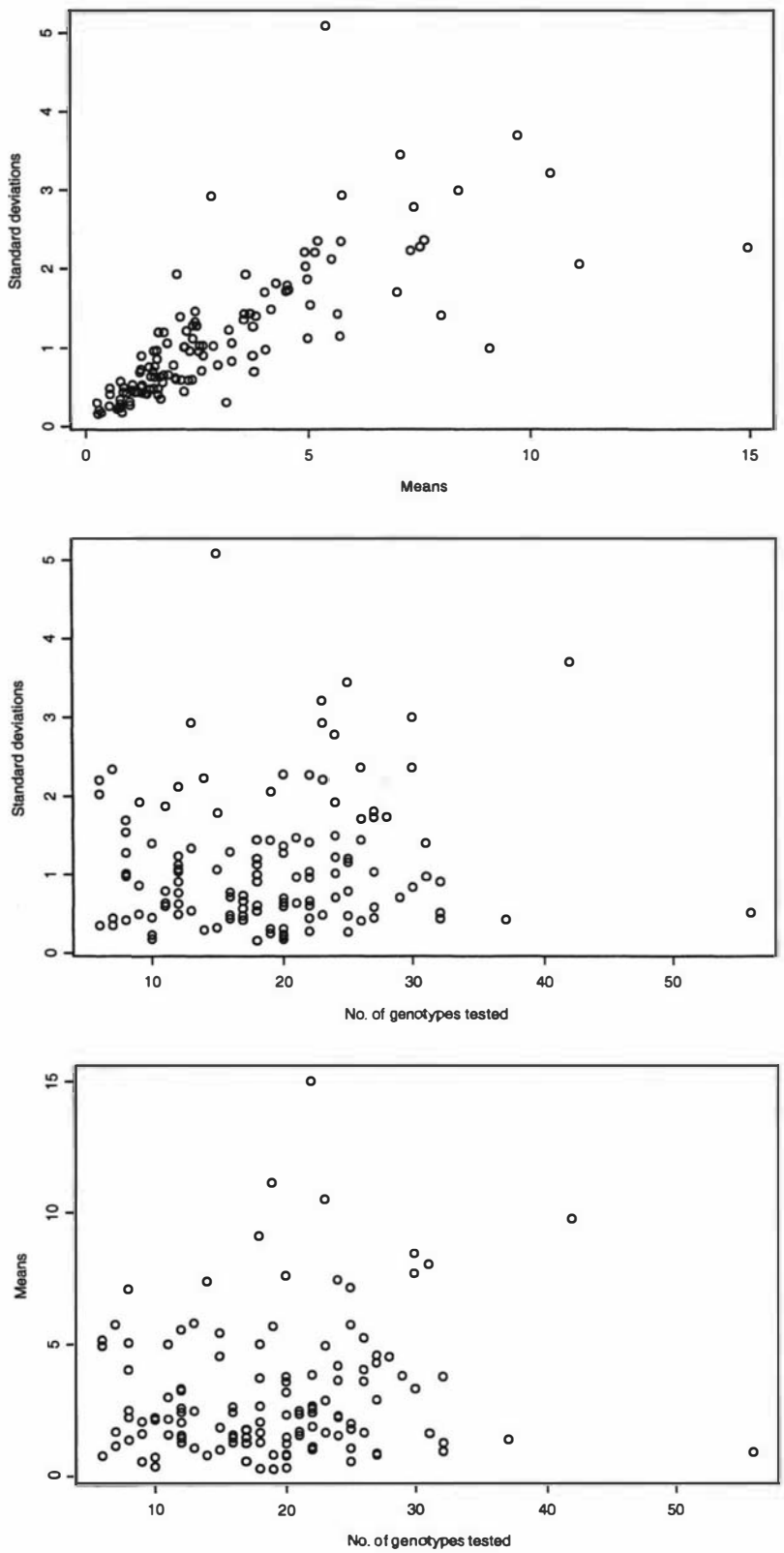


Figure 3.2: Environment means, standard deviations, and the number of times each environment was used plotted against one another. Means and standard deviations are measured in kilograms per square metre.

deviations of both genotypes and environments increase with their respective means. This is common in yield data as low genotype or environment means require all data to be low within these genotypes or environments. Higher means can result from a broader range of yields and therefore are capable of having greater variation within genotypes or environments. The genotype means and standard deviations are not dependent on the number of times each was used, but the middle panels of these figures show that the variation of yields becomes more consistent as genotypes were used more often. The same is true for the data from environments.

In many statistical analyses it is convenient to assume normally distributed response data. It is common for yield data to be positively skewed as the phenomenon of increasing variance with increasing mean comes into effect. Normal probability plots presented in Figure 3.3 show how using square and cube root transformations removes some of the positive skewness in the data. In the normal probability plots for transformed yields there is a group of observations in the lower left corner which do not follow the pattern of the majority of yields. This occurred as a direct result of the many crop failures that led to very few or no marketable onions. The scale of the first normal probability plot masks this phenomenon.

Of the three options presented, the square roots of yields appear to give the closest approximation to normally distributed data. Choosing to use this data instead of the raw yields also removes the relationship between mean and variance, as seen in Figures 3.4 and 3.5. In future analyses of the Onion Trials Programme data, taking the square roots of the G×E yields will be done as the preferred transformation. This transformation does not completely remove heterogeneity of variance for either genotypes or environments. The standard deviations of square roots of yield remain comparatively high for some environments that tested few genotypes.

Linking covariates with yield data

A parametric model that explains yield in terms of variety response to certain environmental factors would preferably use all covariate information collected. Only the latitude and altitude are constant for all tested G×E combinations within each environment; these have been used in Section 3.5 to form a parametric model that aims to explain yield performance. Figure 3.6 shows the range of environmental altitudes and latitudes used in the Onion Trials Programme by presenting histograms of each covariate, and a scatter plot of these covariates plotted against one another. As the main focus of the Onion Trials Programme is on short day onions in the tropics and sub-tropics, the latitudes covered are mainly low (near the equator); exceptions to this include environments from Mediterranean locations such as Italy and Greece which experience climatic conditions over the winter which justify their inclusion in the programme. Altitudes are wide-ranging; sites used in Nepal and the highlands of Kenya are examples of high altitude environments. Generally

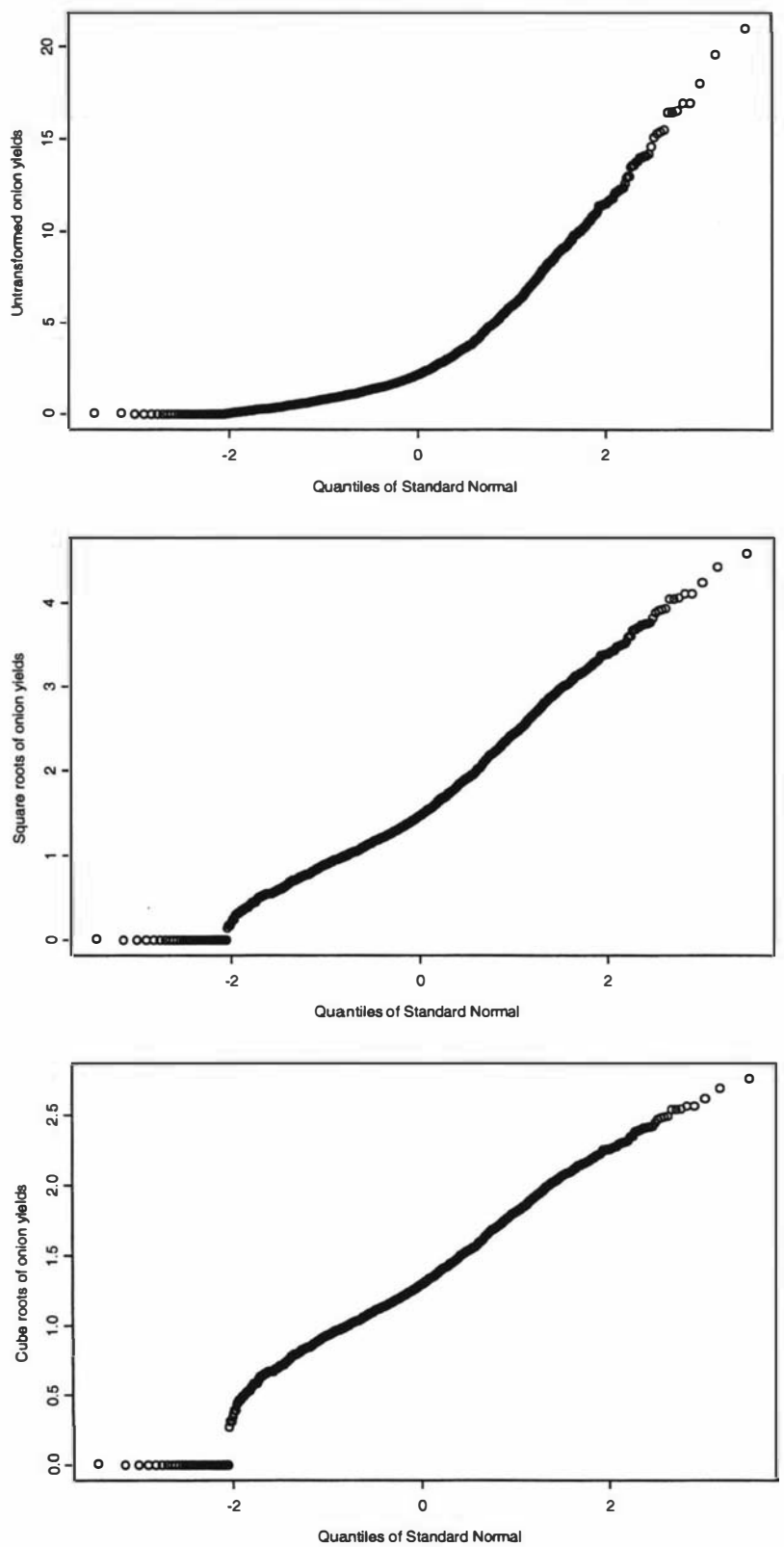


Figure 3.3: Normal probability plots of untransformed onion yields (top), then using the square roots (middle) and cube roots (bottom) of yields.

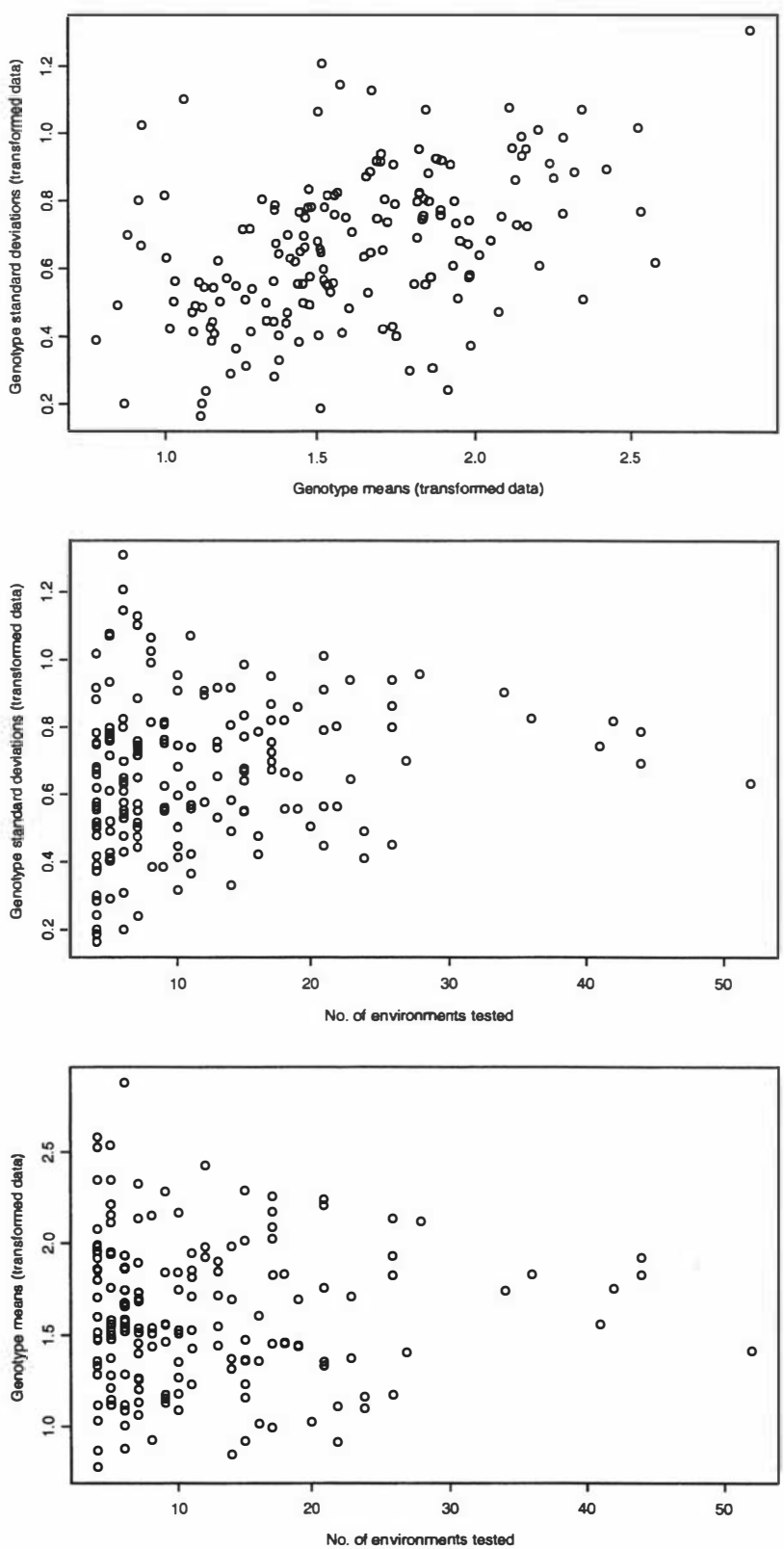


Figure 3.4: Genotype means, standard deviations, and the number of times each genotype was used plotted against one another. Means and standard deviations are calculated for the square roots of yields in kilograms per square metre.

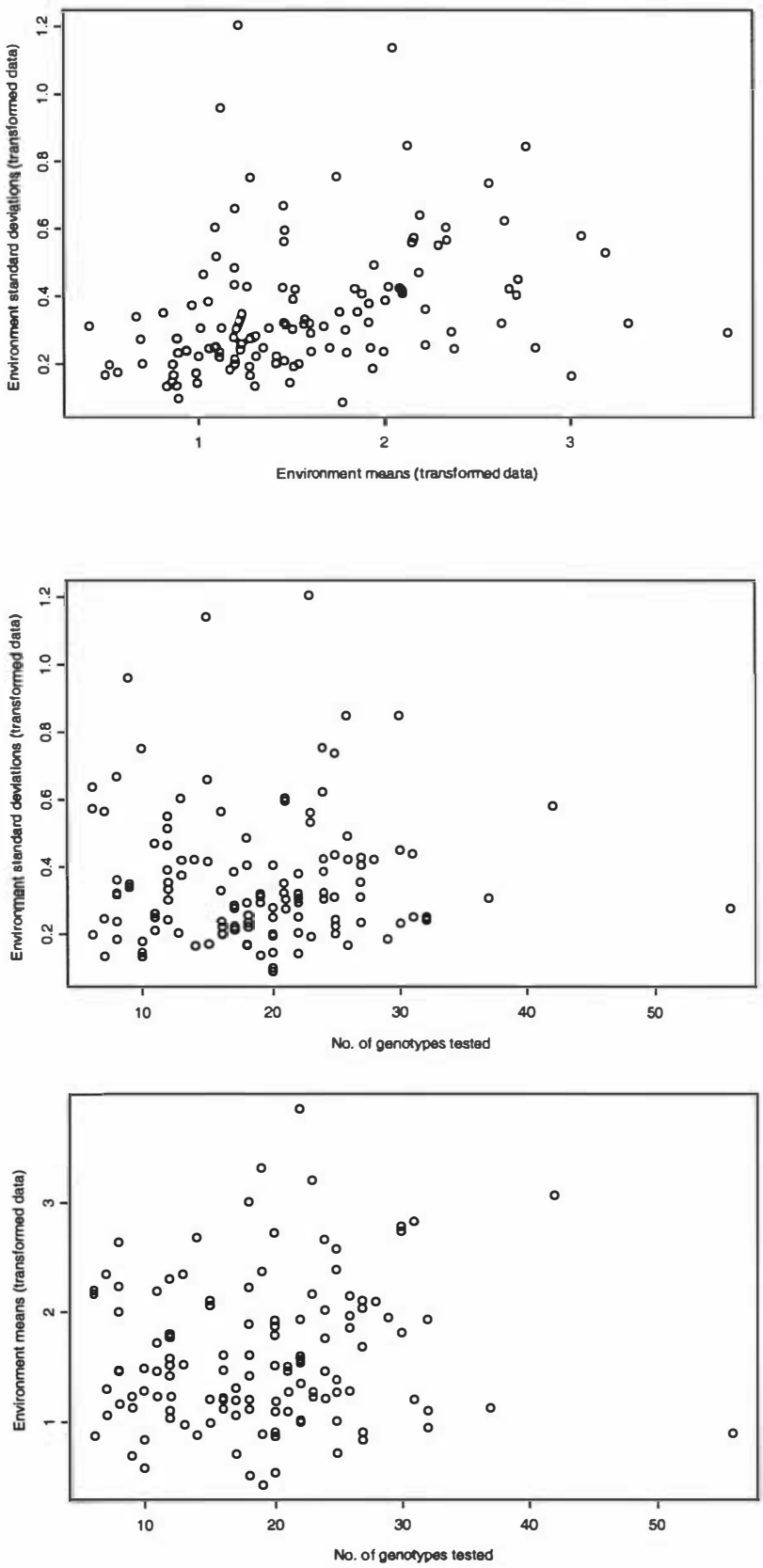


Figure 3.5: Environment means, standard deviations, and the number of times each environment was used plotted against one another. Means and standard deviations are calculated for square roots of yields in kilograms per square metre.

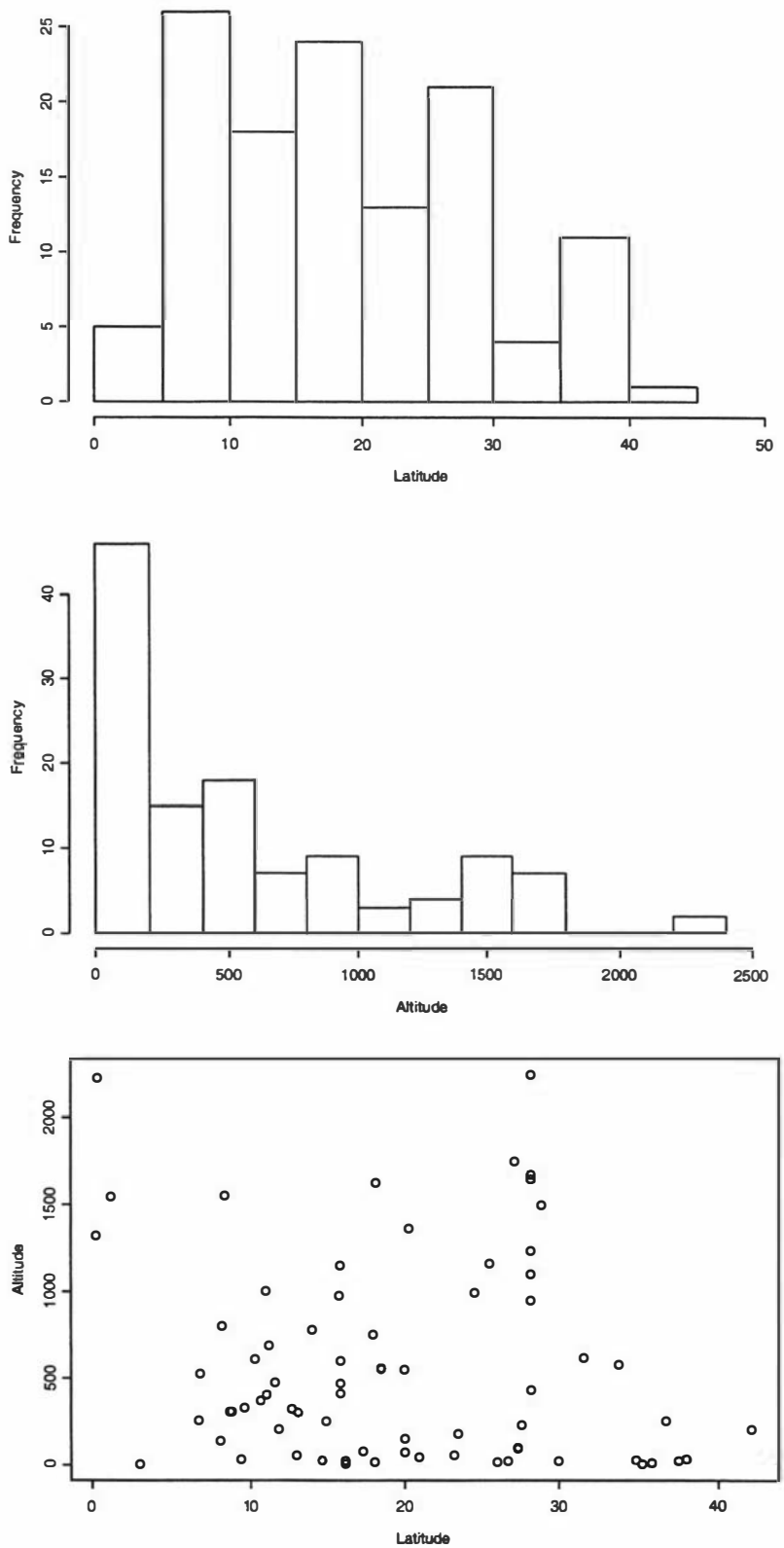


Figure 3.6: Histograms of latitude(top) and altitude (middle) for the 123 environments of the Onion Trials Programme. The scatter plot (bottom) shows the joint distribution of these variables.

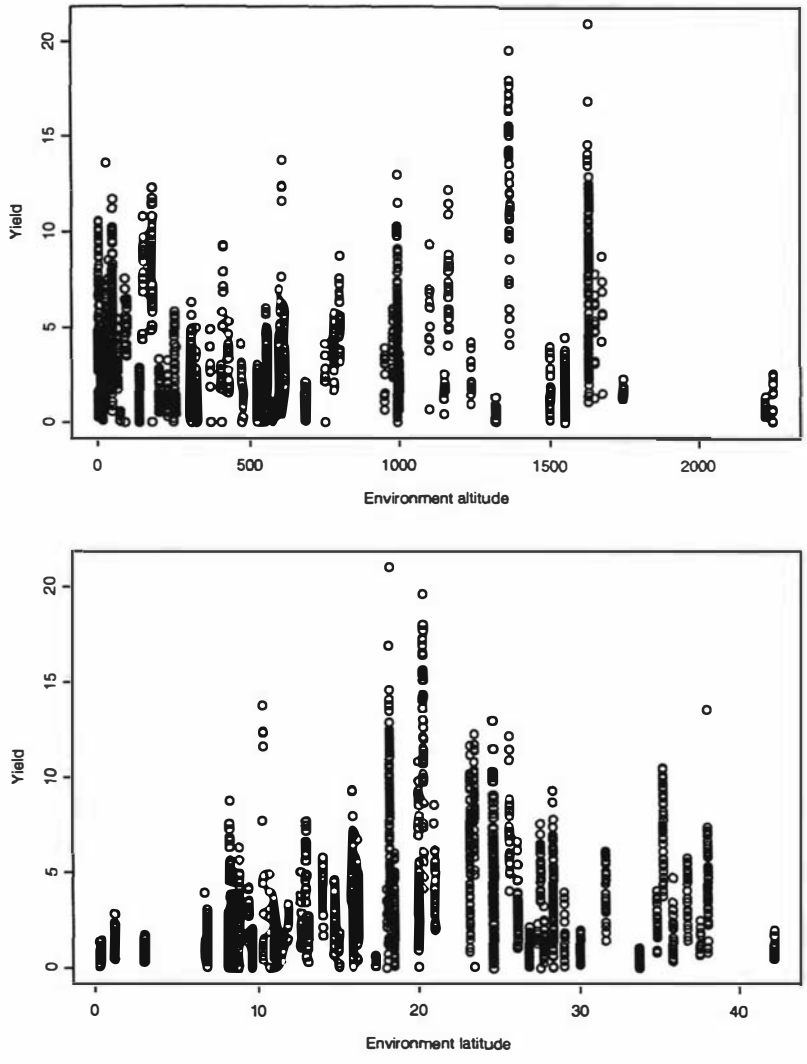


Figure 3.7: Onion yields per square metre plotted against altitude (top) and latitude (bottom).

the widest range of altitudes comes from locations between 20° and 30° of latitude.

Figure 3.7 shows how yields respond to the stresses of altitude and latitude using raw data, while Figure 3.8 uses the square roots of yields for this comparison. These plots are hard to read because there are many points overlaid; environment mean yields and standard deviations have been plotted against altitude and latitude in Figure 3.9 to clarify the relationship between yield and these two covariates. It shows that environmental means and standard deviations are not linearly related to latitude. Environments at high altitude however tend to have higher mean yields and show lower variability than those at lower altitudes.

Of the 400 varieties in the Trials Programme 'Red Synthetic HZ' is the most frequently tested, having been used in 52 of the 123 trials. Figure 3.10 shows a histogram of its yields and plots them against environment altitudes and latitudes. These plots do not show the relative strength of this genotype in terms of its suitability for use in certain latitudes or altitudes; rather they show the absolute value of growing this variety in such conditions. At first glance, Red Synthetic HZ performed outstandingly well in a Zimbabwean trial

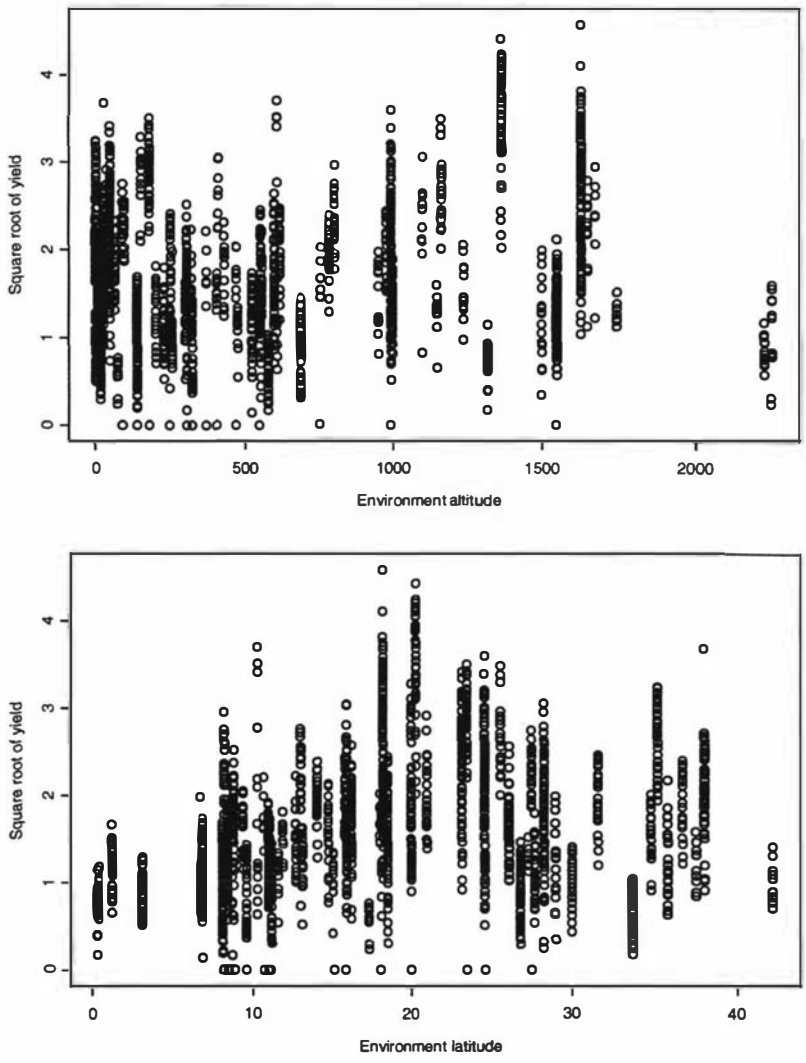


Figure 3.8: Square root of yield per square metre plotted against altitude (top) and latitude (bottom).

(D03603) where it yielded 10.7kg/m^2, but this must be compared to the average yield for this trial which was 10.47kg/m^2. The plots given in Figure 3.11 have been created by subtracting the environment mean from each yield of Red Synthetic HZ, and therefore present its relative effectiveness to the other genotypes grown in each environment. Red Synthetic HZ generally performed below average, but in certain environments performed well. This illustrates the specific adaptation of this variety and its contribution to the G×E interaction of the data set. The most notable relative yield of Red Synthetic comes from a Zambian trial (X04804) where it yielded 5.63kg/m^2 compared to the average of 3.77kg/m^2. When considering the performance of a variety using relative yields, there is some danger in saying that it is specifically adapted to the environment where it has a high relative yield. Relative yields are dependent on the subset of genotypes that were tested at each environment so relative yield plots should not be used to focus on particular environments, but could be used to look for patterns over sets of environments.

The amount of time between sowing and harvesting of onion varieties differs from environment to environment. Figure 3.12 shows the range of growing periods used in each

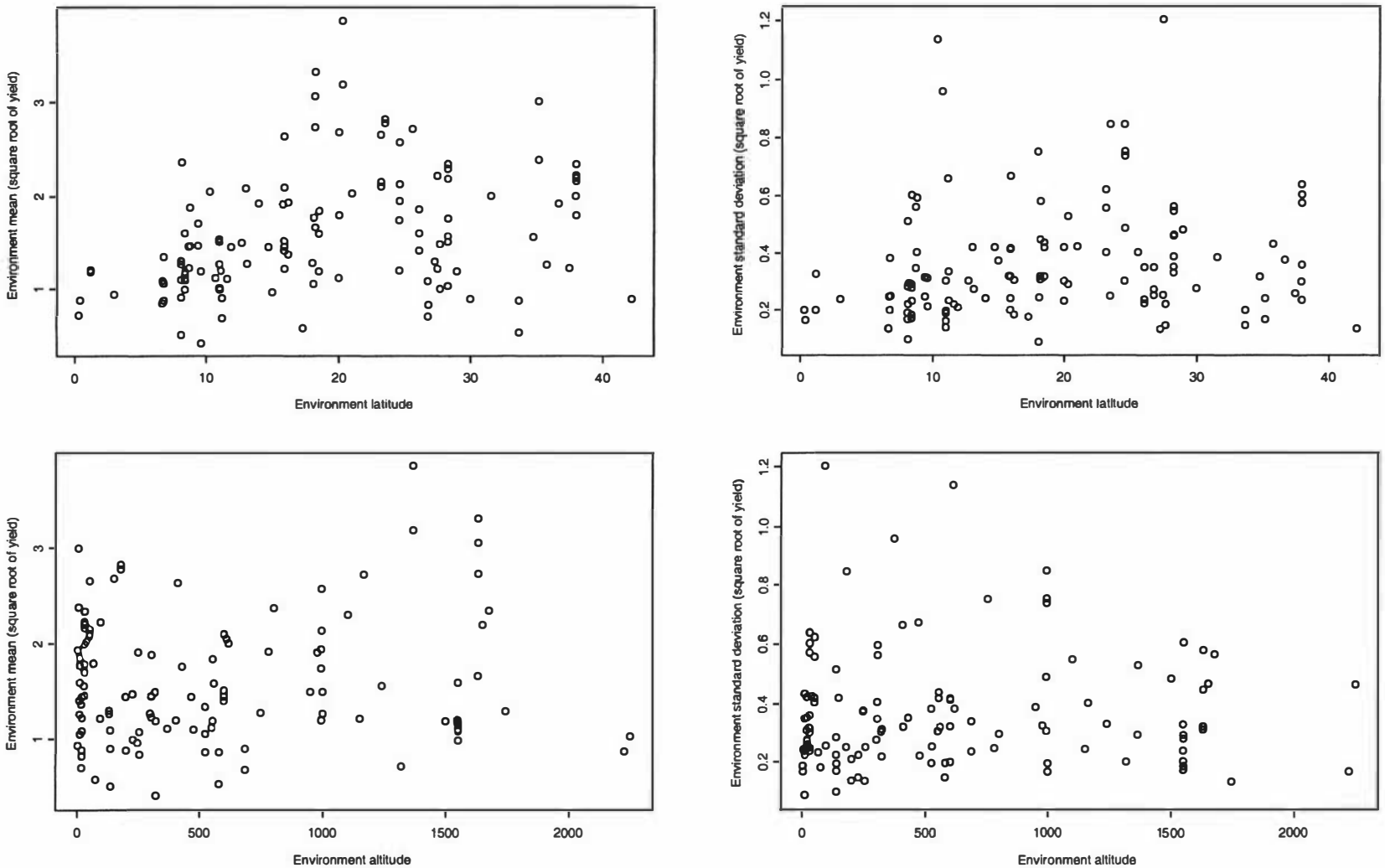


Figure 3.9: Environmental means (left) and standard deviations (right) of square roots of yields plotted against latitudes (top) and altitudes (bottom).

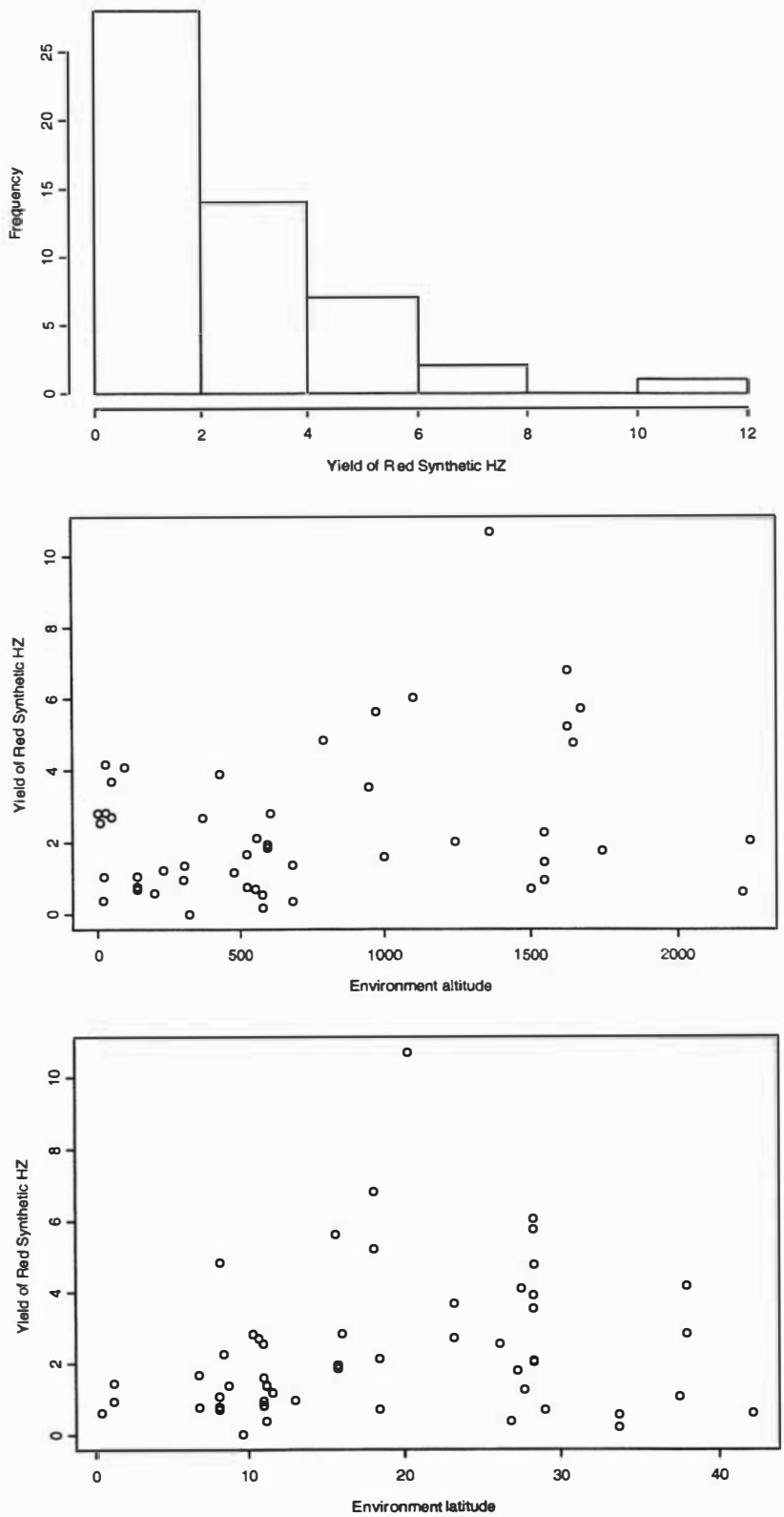


Figure 3.10: Yields of the most frequently used genotype of the trials programme 'Red Synthetic HZ' are presented in a histogram (top), and scatter plots against environment altitudes (middle) and latitudes (bottom)

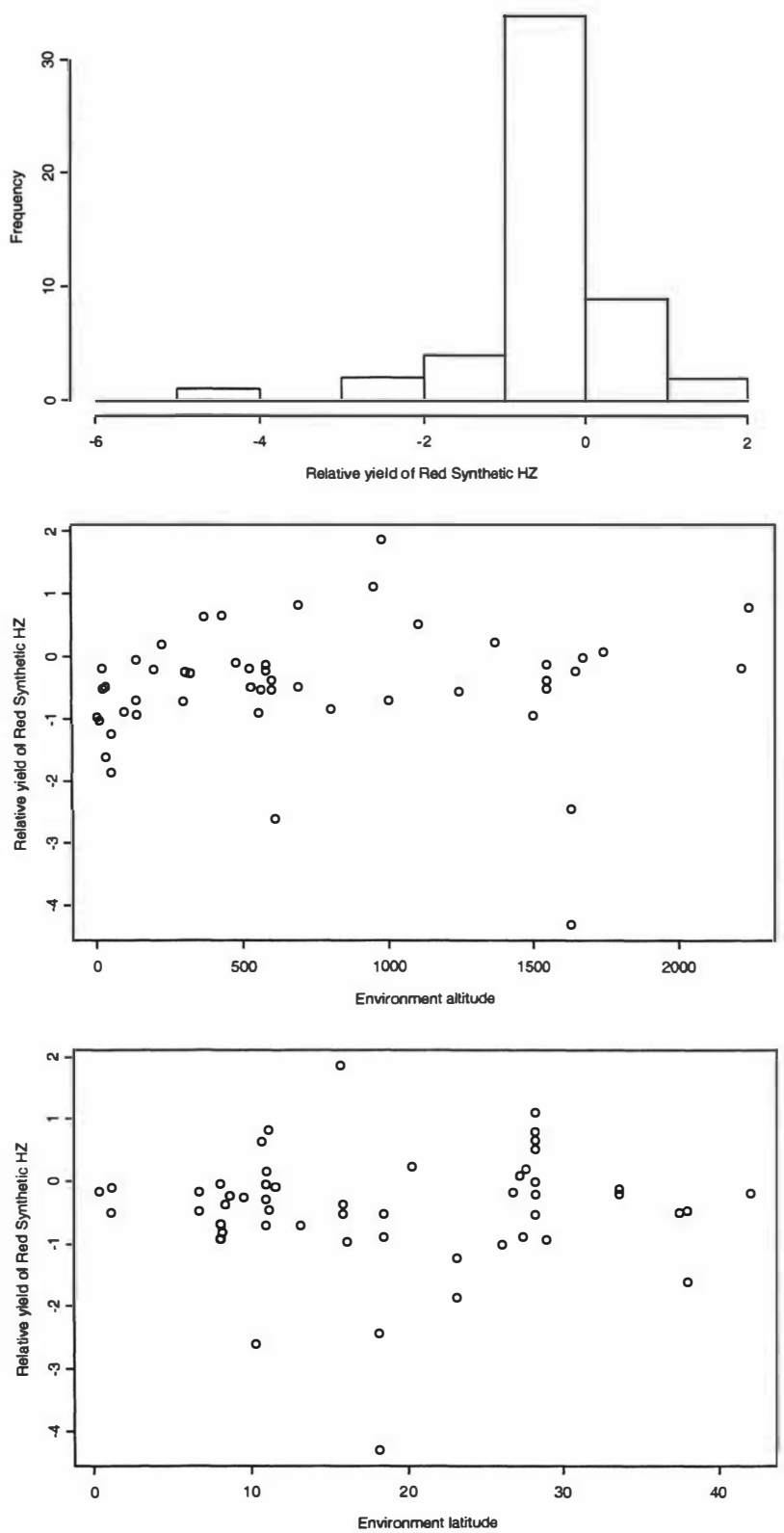


Figure 3.11: Relative yields of the most frequently used genotype of the trials programme 'Red Synthetic HZ' are presented in a histogram (top), and scatter plots against environment altitudes (middle) and latitudes (bottom). Relative yields were found by subtracting the environment mean yield from that of Red Synthetic HZ.

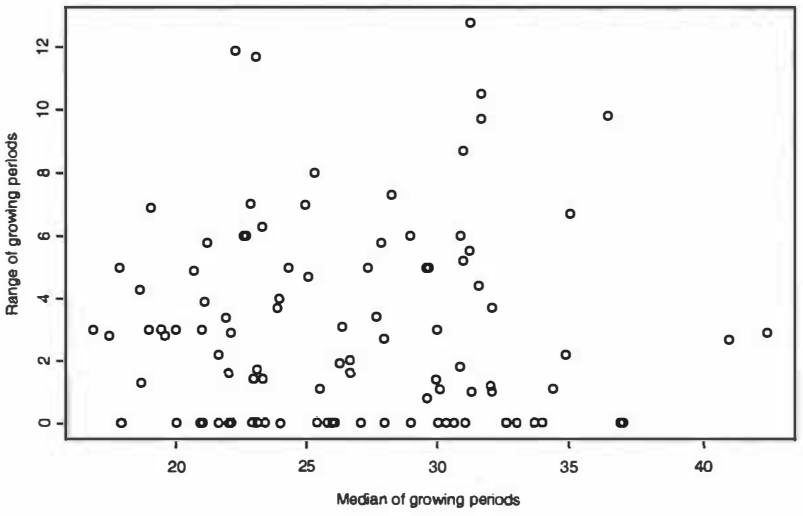


Figure 3.12: Range of growing periods plotted against corresponding median of the growing periods for each environment.

environment plotted against the median growing periods of the environments. There are a number of environments in which the growing periods are wide-ranging, with harvesting dates spanning more than two months in some environments. There is clearly a need to understand the cause of such differences as they may impact on future recommendations. This figure shows that genotypes will determine how long they need to reach maturation in the tested conditions. In some environments, collaborators harvested all varieties at the same time, giving rise to many points at the bottom of this scatter plot. If the growing period is to be used as a response variable in future analyses, data from these environments will need to be discarded. These wide-ranging growing periods also indicate the need to consider sowing dates that are different to the standard local practice if optimal results are to be gained from varieties that are not in common use.

In Figure 3.13, the median and range of growing periods of environments are plotted against the corresponding latitudes. The environments nearer the equator tend to have shorter growing periods than those at higher latitudes. There appears to be no relationship between the range of growing periods within an environment and the latitude of the environment. The collaborators who chose to harvest all varieties at the same time came from environments at wide ranging latitudes. Currah (2003) noted that this practice was occasionally undertaken to avoid the possible theft of mature onions.

3.5 Initial modelling using regression analysis

Preliminary modelling attempts to relate the yield performance of genotypes to environmental factors were made by Dr Currah prior to this current study. She used data arising from the earlier years of the Onion Trials Programme to develop simple regression models. In this section an enlarged data set from the Onion Trials Programme has been used to extend Dr Currah's preliminary modelling. Following Dr Currah's example, linear regression

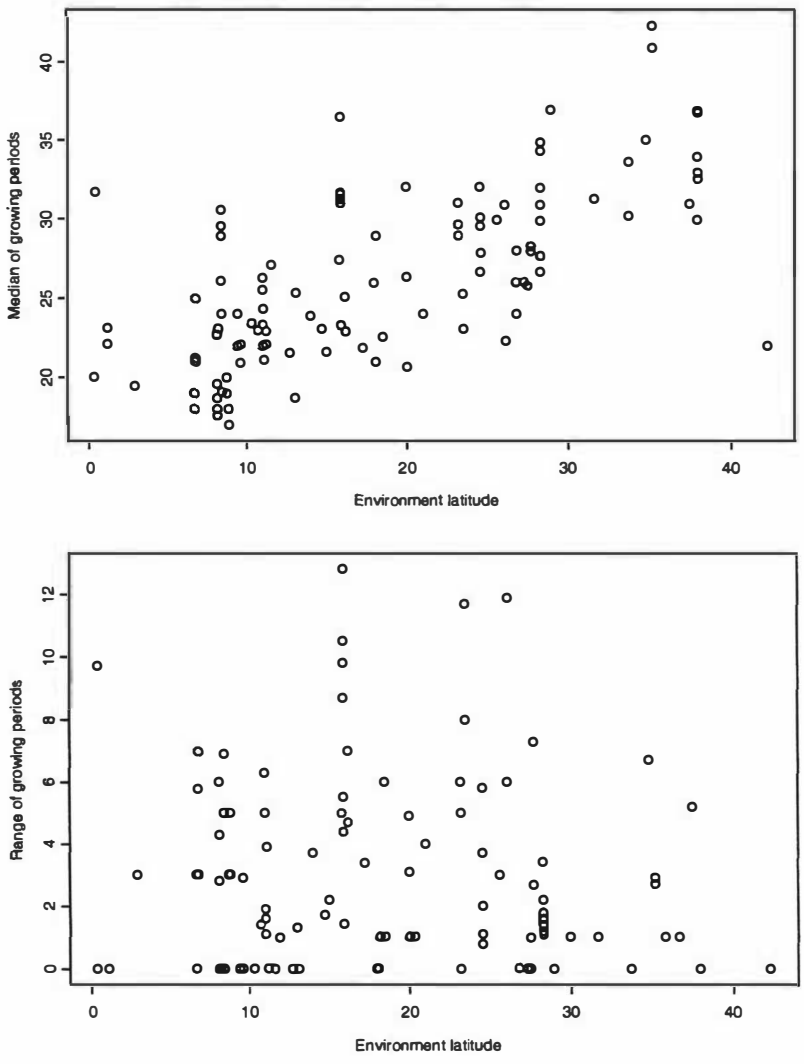


Figure 3.13: Medians (top) and ranges (bottom) of growing periods plotted against latitude. Periods are measured in weeks.

modelling of yield performance has been based primarily on latitude as the explanatory variable.

Due to the sparsity of the data arising from the Onion Trials Programme, the full two-way model that includes row (genotype) and column (environment) factors, as well as their interaction, cannot be fitted because there is insufficient data to estimate all parameters in this model. This model, often referred to as the 'cell means' model (Gauch, 1992), cannot be tested for its validity due to the unavailability of replicate data. A simpler model that explains the G×E interaction is therefore required. This section also highlights the need to develop better methodology for sparse G×E data.

For reasons described in Section 3.4 the square roots of yields were used as the response variable in this modelling rather than the raw yields. The explanatory variables able to be used in these models are latitude and altitude, because the remaining covariate information is inconsistent due to the differing growth periods between G×E combinations. Brewster (1990, 1994, 1997) argued that the amount of daylight onion plants receive affects their performance, and that daylength (photoperiod) is one of the most important factors

Model	Equation
A	$\sqrt{Y_{ik}} = \mu + G_i + E_k + \epsilon_{ik}$
B	$\sqrt{Y_{ik}} = \mu + G_i + E_k + \beta_{1i}\text{Lat}_k + \epsilon_{ik}$
C	$\sqrt{Y_{ik}} = \mu + G_i + E_k + \beta_{1i}\text{Lat}_k + \beta_{2i}\text{Lat}_k^2 + \epsilon_{ik}$
D	$\sqrt{Y_{ik}} = \mu + G_i + E_k + \beta_{1i}\text{Lat}_k + \beta_{2i}\text{Lat}_k^2 + \beta_{3i}\text{Alt}_k\epsilon_{ik}$

Table 3.1: Four models used to explain the performance of genotypes. Model A is the additive model of (2.1) fitted to the square roots of yields, while Models B, C, and D include successive interaction terms for genotype with linear and quadratic latitude effects, and a linear altitude effect.

controlling the timing of bulbing for an individual cultivar. Given that daylength is related to latitude, this covariate has been focused on as the explanatory variable. Allowances have also been made to relate yield and latitude using polynomial regression because there is no assumption that the relationship is linear. The four models are presented in Table 3.1, while Table 3.2 shows the results of applying them to as much of the data as possible for each model; each genotype must have enough data to estimate all relevant parameters in each model. Genotypes with insufficient data have been discarded as more complicated models have been fitted; this criterion has also been extended to environments. To allow comparison of different models, simpler models have been fitted to the same data as the more detailed models. Results from a total of eight combinations of data and model are presented.

Terms for genotype and environment have been included in every model; environments are reflected by two components, one for the latitude, and another to allow each trial's mean performance to be removed. Note that when the quadratic of latitude and altitude are added to the model, their main effects are also confounded with the environment main effect. The aim is to explain the G×E interaction of the genotypes and the environments, using latitude and altitude as the dominant causes of specific adaptation to environments.

Using the lowest Residual MS as the decision criterion, the model formed by relating the square root of yield to a quadratic of latitude is the best. The exact model (noted as 'C' in Tables 3.1 and 3.2) is:

$$\sqrt{Y_{ik}} = \mu + G_i + E_k + \beta_{1i}\text{Lat}_k + \beta_{2i}\text{Lat}_k^2 + \epsilon_{ik} \tag{3.2}$$

The ANOVA table for this model is presented in Table 3.3, and its validity can be checked using the diagnostic plots presented in Figure 3.14. The upper panel of this figure shows the residuals from this model plotted against the fitted values. The linear series of points in the lower left part of this graphic are the results for the zero yields that were added into the data. Residuals from this model are homoscedastic across the range of fitted values, but appear to be non-normally distributed. The Kolmogorov-Smirnov test statistic for the residuals from this model was 0.0700, which results in rejection of the

	Model							
	A				B		C	D
No. of observations used	2343	2197	1668	1466	1668	1466	1466	1419
Discard if not more than	0	1	6	8	6	8	8	8
Source								
Cultivar	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Latitude	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Tricode	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Cultivar*Latitude					0.000	0.000	0.000	0.000
Latitude^2							0.000	0.000
Altitude								0.000
Cultivar*Altitude								0.005
Cultivar*Latitude^2							0.000	0.000
Error MS	0.125	0.125	0.128	0.122	0.121	0.117	0.11	0.111

Table 3.2: Results of applying various regression models to the sparse Onion Trials Programme data. The number of observations used in each model is dependent on the number of parameters fitted. The second row of the table shows how many observations for each genotype and environment were needed to be included in the analysis. For comparative purposes, the models requiring less data in each row of the G×E matrix were also fitted using the same data constraints as more complicated models. An asterisk indicates an interaction term, while empty cells reflect the absence of that term from the model. Values presented are individual p-values for terms in each model. The models, labelled A to D, are specified in Table 3.1.

Source	Df	SS	MS	F	p
Cultivar	85	198.47	2.335	20.95	0.000
Tricode	110	580.56	5.278	47.36	0.000
Cultivar*Lat	85	17.30	0.204	1.83	0.000
Cultivar*Lat^2	85	15.52	0.183	1.64	0.000
Residuals	1100	122.58	0.111		

Table 3.3: ANOVA for the regression of yield on cultivar and tricode factors, a quadratic form of latitude, and the interaction of the quadratic of latitude with cultivar. The asterisk indicates this interaction in keeping with the expression in (3.2). The model is shown in Table 3.2 as model 'C'.

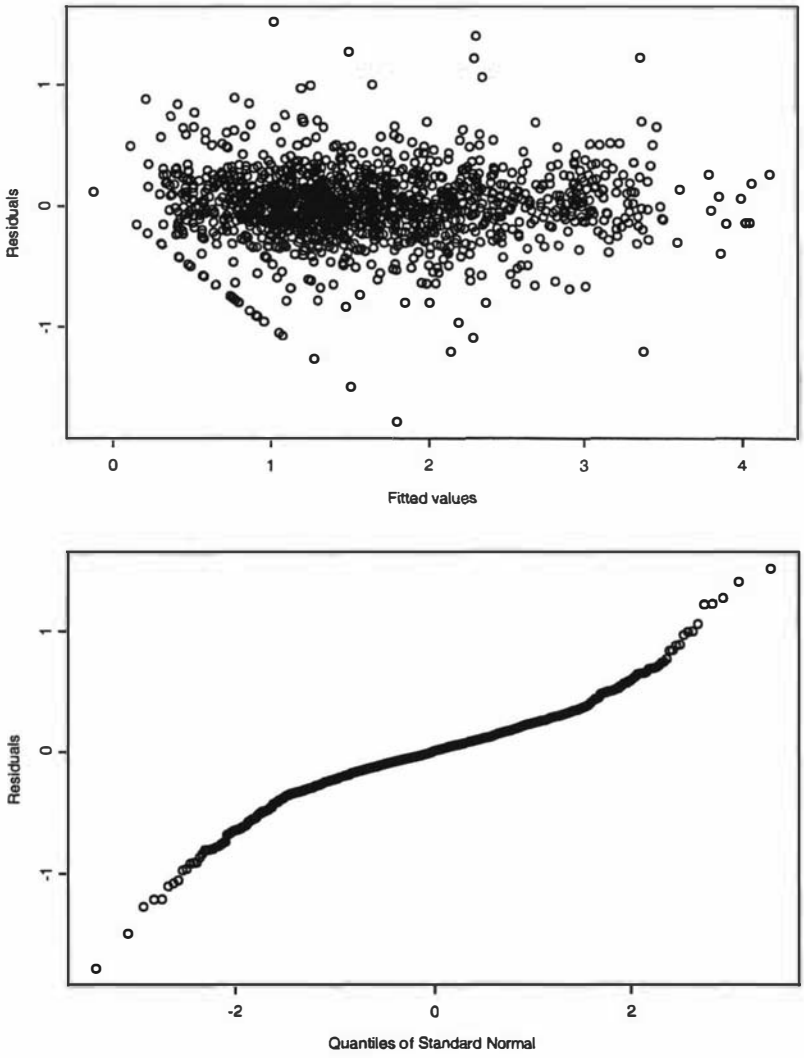


Figure 3.14: Diagnostic plots for the regression model presented in (3.2), whose ANOVA is presented in Table 3.3. The upper panel shows the plot of residuals against the fitted values from the model, while the lower panel shows the normal probability plot for the residuals.

Source	Df	SS	MS	F	p
Latitude	1	0.0041	0.0041	0.074	0.787
Latitude^2	1	0.0035	0.0035	0.062	0.804
Residuals	47	2.6055	0.0554		

Table 3.4: The portion of the ANOVA table specific to the most commonly tested genotype (Red Synthetic HZ) extracted from the ANOVA, summarized in Table 3.3, for the model relating genotype performances to environmental latitude.

null hypothesis of normally distributed residuals. It should be noted that one of the fitted values was negative; this would be of greater concern if more values represented impossible results.

If having non-normally distributed residuals is acceptable, then the model in (3.2) will assist trials programme organizers to answer the principal research question. Another concern exists however; parameters found using this model are subject to bias, arising as observations are deleted from the analysis (Schafer, 1997). This concern remains for the data arising from the Onion Trials Programme because G×E combinations are not 'missing completely at random'.

This model allows for a quadratic relationship between the square roots of yields of each genotype and the latitudes of environments. Considering the scatter plot given in the bottom panel of Figure 3.11, it is unlikely that the model sufficiently explains the specific adaptation of the most frequently used genotype (Red Synthetic HZ). The portion of the ANOVA specific to this variety is presented in Table 3.4, and clearly shows that neither the linear nor quadratic terms are significantly different from zero for this genotype and are therefore redundant. The model given in (3.2) is therefore unsatisfactory. In spite of its shortcomings, it can be used as a standard that must be surpassed if improvements are to be made.

3.6 Finding a suitable subset of data

Reduction of data via deletion of rows and/or columns from the G×E matrix is wasteful of information, but is necessary for the following reasons:

1. A minimum amount of data is needed for model-fitting, as in Section 3.5. If there are insufficient data in rows or columns there will be no error term against which to gauge significance of results.

2. Under-represented genotypes (environments) cannot be compared to any other genotypes (environments) and should automatically be discarded. This is crucial when distance measures are calculated in cluster analyses, considered in Chapter 4.

3. The under-represented genotypes and environments contribute heavily to the spar-

sity problems.

A balance between reducing sparsity and model-fitting problems, and the desire to waste as little data as possible, must be found.

Removing the genotypes that were grown in only one or two environments was an obvious first step. These were often 'local' varieties that collaborators included for personal interest. While they have been discarded from the G×E matrix, it should be possible to include their results at the recommendation stage, so that there is still value to be gained from the efforts of collaborators in respect of their individual research goals.

As more genotypes were deleted, benefits to reduction in sparsity diminished. This can be seen in Table 3.5. Reduction of the G×E matrix, through deletion of under-represented genotypes was satisfactory until the minimum representation of genotypes P_m, as well as environments Q_m, were set at five. At this point, two environments became inadmissible as they no longer had results from five or more genotypes. Following this process further resulted in the reduced G×E matrices outlined in Table 3.5.

An unfortunate effect of genotype deletion was the removal of environments. By implication, the efforts of collaborators in these deleted environments were then deemed to be of less value to the overall programme in its current form. If ways could be found of ensuring that more of these deleted genotypes (tested in deleted environments) were selected for future trials, their inclusion in a future analysis of the overall programme may be possible. Methods for ensuring this are discussed further in Chapter 10.

The ability to compare genotypes (environments) will be discussed throughout this work. The notation P_{ij} will indicate the number of environments the ith and jth genotypes have in common, and P_{ii} indicates the number of environments in which the ith genotype was tested. Correspondingly, Q_{kk} is used to show how many genotypes were tested in

Minimum representation	No. of genotypes remaining	No. of environments remaining	Available data as %	Sufficient P_{ij} as %
1	400	123	4.74	1.84
2	254	123	7.00	4.57
3	208	123	8.18	6.81
4	169	123	9.51	10.3
5	143	121	10.8	14.0
6	123	118	12.1	18.3
7	104	109	14.0	23.8
8	87	98	16.3	29.9

Table 3.5: The effects of row and column deletion to ensure minimum representation of both genotypes and environments is passed. The sparsity of the resulting G×E matrices is expressed in terms of the amount of data present in the matrix and is found by dividing the number of observations in the reduced matrix by the product of its dimensions. The last column presents a measure of inter-genotype linkage that exists in the G×E matrices.

combinations that have sufficient linkage. Sufficient linkage is defined, in this instance, as having four or more environments in common, although this threshold can be adjusted as necessary. These percentages are calculated as the number of available pairwise links of sufficient quality ($P_{ij} \geqslant 4$) divided by the number of genotype pairs $I(I-1)/2$, where I is the number of genotypes. It provides a useful gauge on the inter-connectivity of genotype information in the data. Once again this value is increasing as rows and columns were deleted from the G×E matrix. Inter-connectivity is important as it measures how well genotypes can be linked to one another. If inter-connectivity is low, few options for comparison will exist for some genotypes which is clearly undesirable. If cluster analysis is to be used, distance measures will need to be calculated between pairs of genotypes. The last two rows of Table 3.5 show that approximately one quarter of genotype pairs are sufficiently linked. A strategy for determining distances for genotype pairs with insufficient linkage will be presented in Chapter 4.

While the sparsity of the G×E matrix was decreasing, and the sufficient linkage criterion was increasing, more data was thrown away. When $P_m = Q_m = 5$, 467 observations have been deleted from the data. If $P_m = Q_m = 6$, 7, or 8, are chosen, 582, 749, or 941 observations respectively would be discarded. This would clearly waste the efforts of collaborators and programme organizers alike, and has forced the development of ideas that result in these data being used in some way.

The data's sparsity required feasibility of methods to be checked at every turn. Contingency plans needed to be considered for many circumstances, and incorporated where necessary. The methodology presented in subsequent chapters relies on the G×E matrix having complete connectivity. Connectivity is used to mean the ability to link performances of every pair of genotypes, using existing direct links. Genotypes i and j would have connectivity if:

1. $P_{ij} \geqslant P_m$, or

2. Each of these genotypes has sufficient linkage with at least one other genotype h, i.e. $P_{hj}, P_{hi} \geqslant P_m$.

The last two rows of Table 3.5 describe the details of the two data sets that will be used throughout this work. In Onion Data I and II, as they will be called, the number of times a genotype (environment) was used P_{ii} (Q_{kk}) is at least 7 and 8 respectively. These G×E matrices have complete connectivity; setting $P_m = Q_m \geqslant 9$ results in disconnected G×E matrices. A disconnected matrix would need to be split into a set of connected submatrices, as against a connected set of submatrices. These separate analyses cannot be linked together. Increasing the amount of linkage between all pairs of genotypes, and therefore environments, will clearly improve the quality of the analysis of the Onion Trials Programme.

In Section 3.3 one reason for the inclusion of known failures (where a genotype produces very little or no marketable yield in a certain environment) in the data was given. Another reason for the creation of data to reflect current knowledge became apparent. If it could be said, with confidence, that some varieties would not yield any marketable product if grown in a certain environment, their entries in the G×E matrix could be created (entered as zero). This would then add more entries to some rows and columns of the G×E matrix and could possibly ensure that some genotypes and environments did not get discarded unnecessarily. This scenario actually occurred in the Onion Trials Programme with the trial (coded as X04601) from Cameroon in 1994. This trial grew only six varieties, and three G×E combinations were recorded as zero due to known crop failure. Of these nine entries in the total data, two were for genotypes that were under-represented in the trial programme, and were discarded in the creation of Onion Data I. Without the creation of the zero yields for three G×E combinations this trial's data would have been discarded. Use of the discarded genotypes that were grown in environment X04601 in the future will increase the chance of this environment being part of Onion Data II.

3.7 Adapting current methodology to allow for incomplete data

This section reports on attempts to apply current G×E methodology to sparse data from the Onion Trials Programme. Some stability measures discussed in Section 2.7 were modified and applied. Problems encountered when working with sparse data are highlighted.

Consistency of results is demonstrated to be difficult to measure with sparse data. Parameters based on fewer data have less accuracy; identifying the level of accuracy poses a problem in itself. Use of the methods presented in this section on data from both Onion Data I and II allowed some comparison.

The other main concern of the methods presented in this section was the need to be cognizant of the source of data used to calculate parameters. Some allowance for this was made by modifying three parameters from the G×E literature:

1. The coefficient of variation, as applied by Francis and Kannenberg (1978).

2. Wricke's ecovalence, which, while not in common usage, is cited by Lin *et al.* (1986) and Kang *et al.* (1987) as one of the first stability parameters of G×E research.

3. The superiority measure of Lin and Binns (1988a).

The Francis and Kannenberg (1978) coefficient of variation

The coefficient of variation for the yields of a given genotype was employed by Francis and Kannenberg (1978) as a stability measure, and expressed the ratio of a genotype's

Data set	Genotypes
Onion Data I	Granex Yellow TK, HA-230 HZ, Jaguar PS, Kano Red NI, *Mercedes PS, Red Comet PS, Rojo SS,* Texas Grano 438 AS.
Both Onion Data I and Onion Data II	Belem IPA-9 IP, Colossal PVP SS, Composto IPA-6 IP, Early Red HZ, Galil HZ, Granoble PS, HA-226 HZ, HA-817 HZ, HA-950 HZ, Houston AS, IRAT-69 MA, Niv HZ, Red Burgundy Imp NE, Redbone AS, Regal PVP SS, Regia AS, Ringer Grano SS, Rio Blanco Grande RC, Texas Grano 502 PRR AS,Texas Grano LO, Tropic Ace TK, Tropic Gold NW, Yellow Granex Imp PRR SS.
Onion Data II	Rio Hondo RC.

Table 3.6: Genotypes with an adjusted coefficient of variation, calculated using (3.4), considered high (in excess of 200%). Data from Onion Data I and II were used. Note that genotypes in Onion Data II are also in Onion Data I, and Onion Data I genotypes that are also part of Onion Data II are italicized.

standard deviation to its mean performance (as a percentage). The sparsity of the G×E matrices of Onion Data I and II made such a calculation misleading; the stability measure for all genotypes should be based on the same set of environments for it to be appropriate.

To counter the heterogeneity of environment variance, seen in Figure 3.5, responses were standardized using

$$z_{ik} = \frac{y_{ik} - \bar{y}_{\cdot k}}{s_k} \tag{3.3}$$

where $\bar{y}_{\cdot k}$ and s_k are the kth environment mean and standard deviation, respectively for the data being used; in this instance the square roots of yield were chosen as the response variable. This transformation does not alter the qualitative structure of the G×E interaction that exists in the data. It, therefore, ensures that each environment contributes on an equal footing to the stability measure. This transformation was used extensively throughout this work as a means of countering sparsity and the impact of heterogeneity of environment variance.

The adjusted coefficient of variation $CV_i^\star$ is found using

$$CV_i^\star = \frac{s_i}{\bar{z}_{i\cdot}} \times 100 \tag{3.4}$$

where $\bar{z}_{i\cdot}$ and s_i are the mean and standard deviation of the within-environment standardized yields z_{ik} for the ith genotype. The adjusted coefficient of variation will be expressed as a percentage as in Francis and Kannenberg (1978).

Table 3.6 shows the genotypes of Onion Data I and II that have an adjusted coefficient of variation greater than 200%. These genotypes are contributing significantly towards the G×E interaction of the data. They are varieties that are therefore likely to have good specific adaptation potential.

Using more data can be seen to affect results. When using Onion Data I, instead of

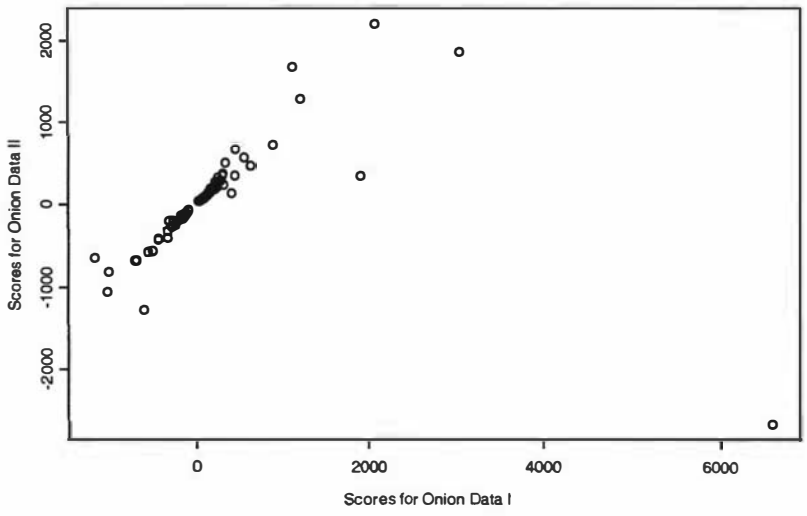


Figure 3.15: Adjusted coefficients of variation calculated using yields in Onion Data I and II, for the 87 genotypes of Onion Data II.

Onion Data II, a further 17 genotypes and 11 environments were added to the analysis. Adjusted coefficients of variation for the 87 genotypes of Onion Data II are plotted against their corresponding scores calculated using Onion Data I in Figure 3.15. A correlation coefficient of 0.128 was found for the scores presented in this figure, and showed that the scores were inconsistent. The adjusted coefficient of variation increased for three genotypes listed in Table 3.6 as a result of using data from more environments (Mercedes PS, Red Comet PS, and Rojo SS), while one genotype's coefficient of variation decreased as a result of using data from more environments (Rio Hondo RC). Adjusted coefficients of variation can be negative when using standardized data, seen in Figure 3.15, as a result of a negative mean yield. A small change in the mean yield, from a positive to a negative yield, could severely alter the adjusted coefficient of variation. The opposite of this problem occurred for the variety 'Red Comet PS' whose mean yield changed from −0.035 to 0.014, with the addition of data from another three environments. When using data with both positive and negative data, the absolute value of the adjusted coefficients should be used instead of the raw scores to avoid this problem. The results presented in Table 3.6 should not, therefore, be considered a complete list of varieties for investigation under this criterion.

The adjusted coefficient of variation should not be relied upon as a criterion for judging a genotype's potential. If used carefully, however, it may be useful for highlighting varieties that have performances worth closer investigation. The next subsection shows another way of looking for this specific adaptability.

Wricke's ecovalance

Wricke's ecovalence was developed to indicate which genotypes contributed significantly to G×E interaction. Wricke's ecovalence, given in (2.18) as

$$\hat{W}_i = \sum_{k=1}^{K} (y_{ik} - \bar{y}_{i\cdot} - \bar{y}_{\cdot k} + \bar{y}_{\cdot\cdot})^2$$

is equal to the sum of squared residuals for each genotype after the additive two-way model from (2.1)

$$Y_{ik} = \mu + G_i + E_k + e_{ik}$$

has been fitted to complete data. It can, therefore, be expressed as

$$\hat{W}_i = \sum_{k=1}^{K} e_{ik}^2 \tag{3.5}$$

This form of Wricke's ecovalence sums all squared residuals for a genotype, and would, if applied to incomplete data, treat genotypes differently depending on the number of entries each has in the G×E matrix. Wricke's ecovalence was adjusted to overcome this problem, so that it could be applied to the data arising from the Onion Trials Programme. The adjusted ecovalence considers the contribution of each genotype to the residual MS rather than its contribution to the residual SS in the model. In effect this means that the denominator is based on degrees of freedom in the model specific to the genotype, and is the number of entries for the genotype in the G×E matrix minus one. The adjusted form of (2.18) is

$$\hat{W}_i^{\star} = \frac{\sum_{\text{available } k} (y_{ik} - \bar{y}_{i\cdot}^{\star} - \bar{y}_{\cdot k}^{\star} + \bar{y}_{\cdot\cdot}^{\star})^2}{P_{ii} - 1} \tag{3.6}$$

where the $\star$ symbol represents the adjusted form of the relevant estimate as determined by the incomplete data, to minimize $\sum_{i=1}^{I} \sum_{\text{available } k} e_{ik}^2$ when the model in (2.1) is fitted. When this two-way model is fitted using incomplete data, the estimates that are produced for the row and column means are not the observed row and column means; that is, $y_{i\cdot} \neq y_{i\cdot}^{\star}$ and $y_{\cdot k} \neq y_{\cdot k}^{\star}$ unless data is complete. The functionality of S-PLUS to fit this two-way model to incomplete data was used to find the residuals based on the estimates of the row and column effects. Adjusted ecovalence can also be expressed as

$$\hat{W}_i^{\star} = \frac{\sum_{\text{available } k} e_{ik}^2}{P_{ii} - 1} \tag{3.7}$$

The application of adjusted ecovalence to the data from the Onion Trials Programme highlighted some problems. Genotypes grown in only one environment could only be used to estimate the mean of that genotype in the data, and were discarded without altering the estimates of environment means. Inclusion of genotypes grown in two to six environments influenced results; this was discovered when the method was applied to Onion Data I

and II, after its application to the entire data set. Results from Onion Data I and II were comparable, but differed markedly from those found using all of the Onion Trials Programme data.

Table 3.7 shows the genotypes that contributed the most to the G×E interaction of the Onion Trials Programme when using only the data from Onion Data I and II. While the results for Onion Data I and II differed little, the values for the adjusted ecovalence reflected the poor fitting of the model used to find them. Results in Table 3.7 indicate potential for specific adaptation, implying that these varieties should be further investigated. One variety that was in both Onion Data I and II (Mercedes PS) was given adjusted ecovalence scores that differed according to the set of environments over which it was calculated. The limited similarity between results presented in Tables 3.6 and 3.7 was another indicator of poor consistency.

Varieties with low adjusted ecovalence would theoretically be those that had good wide adaptation. The adjusted ecovalence scores for Onion Data I and II are unreliable, highlighted by the fact that no varieties were given adjusted ecovalence scores of sufficiently low magnitude. The sparsity of the G×E matrices for Onion Data I and II is assumed to lead to these poor estimates. These estimates are therefore considered indicative rather than instructive.

The Lin and Binns (1988) superiority measure

The superiority measure of Lin and Binns (1988a), given in (2.20), was easily adjusted to take account of the differing number of environments in which genotypes were tested. By taking an average, over environments, of the squared difference between a genotype's performance y_{ik} and the maximum m_k for each environment, a score $p_i^\star$ can be found that indicates the best performing genotypes in terms of wide adaptability. As above, the square roots of yields were used in this investigation. The denominator of the Lin and Binns (1988a) superiority measure was replaced by the number of environments P_{ii}

Data set	Genotypes
Onion Data I and II	Creamgold YA, Jenin HZ, Marix ZU, RS 209 RS, Tropic Ace TK, Utopia AS, Violet de Galmi TS, Yellow Granex Imp PRR SS.
Onion Data I	Cadix ZU, Eytan HZ, HA-891 HZ, Kano Red NI, *Mercedes PS*, Texas Grano 438 AS.

Table 3.7: Genotypes that contribute highly to the G×E interaction in Onion Data I or II, in terms of adjusted ecovalence given in (3.6). The results were comparable between Onion Data I and II, but some additional varieties in Onion Data I met the high standard set for the adjusted ecovalence criterion. One Genotype from Onion Data I that is also part of Onion Data II has been italicized.

Onion Data I		Onion Data II	
Genotype	$p_i^\star$	Genotype	$p_i^\star$
Hurricane RS	0.07	Hurricane RS	0.07
Granex 429 AS	0.09	Granex 429 AS	0.09
Marathon HZ	0.10	Marathon HZ	0.10
Linda Vista PS	0.11	Superex TK	0.11
Superex TK	0.11	Linda Vista PS	0.12
PS 8392 PS	0.12	Houston AS	0.13
Houston AS	0.14	Rio Bravo RC	0.13
Nikita RC	0.15	Equanex PS	0.16
Rio Bravo RC	0.15	Gladalan Brown YA	0.17
Equanex PS	0.15	Savannah Sweet PS	0.18
Gladalan Brown YA	0.17	RAM 710 HZ	0.18
Jaguar PS	0.17	Rio Raji Red RC	0.18
Dessex SS	0.17	Dessex SS	0.18
Mr Max RC	0.18	Granex 33 AS	0.19
Rio Raji Red RC	0.18	Composto IPA-6 IP	0.20

Table 3.8: Names of the best fifteen genotypes selected using the adjusted superiority score $p_i^\star$ for Onion Data I and II. Figure 3.16 shows the distribution of the adjusted superiority scores.

in which the ith genotype was tested, and is given by

$$p_i^\star = \frac{\sum_{\text{available } k} (y_{ik} - m_k)^2}{P_{ii}} \tag{3.8}$$

The idea of this measure is to identify genotypes that perform as consistently close to the maximum yield of environments as possible. When the measure was applied to a reduced set of data, environmental maxima were able to be found in several ways, including:

1. The maximum of the available yields within the reduced set of data.

2. The maximum of the yields from all genotypes tested in an environment.

3. The maximum potential for the environment using extreme value theory.

This last option would require extra consideration of the distribution of yields to be able to establish the correct extreme theory to apply; a step that was considered to be unnecessarily complicated for this investigation. The second option did however provide some recognition of the potential of an environment by use of the maximum yield of the tested genotypes. It was closer to the actual environment potential, and was therefore preferred in this preliminary investigation.

Table 3.8 presents the names of the fifteen genotypes for each of Onion Data I and II which had the lowest adjusted superiority score. The lists are similar, as were the overall

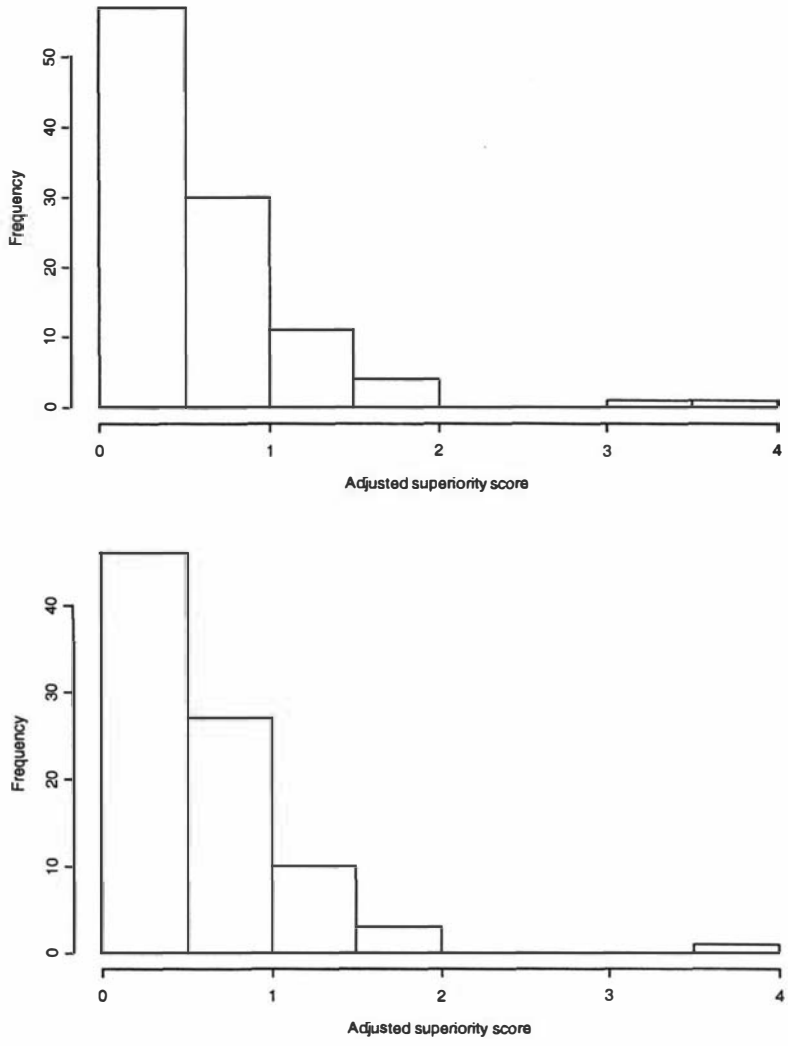


Figure 3.16: Histograms of adjusted superiority scores for the 104 genotypes of Onion Data I (top), and the 87 genotypes of Onion Data II (bottom).

results for the adjusted superiority scores for Onion Data I and II, which had a correlation coefficient of 0.992 for the genotypes common to both data sets.

Histograms presented in Figure 3.16 show that there are a considerable number of genotypes that have adjusted superiority scores less than 0.5. Given the median adjusted superiority scores for Onion Data I and II were 0.458 and 0.486 respectively, it is difficult to recommend use of this measure to select genotypes for future testing. On the other hand, several genotypes have poor adjusted superiority scores and should be removed from future testing. The variety that should not be considered for future testing, according to results from both Onion Data I and II, is 'Marix ZU'. Another variety, 'Cadix ZU', also has a poor adjusted superiority score but was only included in the Onion Data I analysis. Cadix ZU and Marix ZU were tested in seven and eight environments respectively. Before they are totally ignored for future testing an investigation should be made to ensure that these varieties were not grown in a particular type of environment to which they may be unsuited.

The graphical representation of results in Lin and Binns (1988a) was uninformative

Cultivar	Country	Site	Year
Angaco INTA LO	Argentina	San Juan	1997
CV 19 FA	Senegal	St Louis	1993
Early Supreme SS	Benin	Kargui	1998
Jubiley 50 LO	Bulgaria	Plovdiv	1995
Phulkara VR	Sri Lanka	FCRDI	1994
Red Pinoy PH	Mauritius	Richelieu	1998

Table 3.9: The six onion varieties that were used only once in the trials programme and showed the best performance of all varieties in the environments where they were tested. Relevant environment details are also provided.

for sparse data covering a large number of environments such as that of the Onion Trials Programme. An alternative needed to be found which would allow comparison of the relative merits of genotypes using the adjusted superiority scores. It was determined that the number of environments each genotype was tested in had an effect on the outcome of the adjusted superiority scores; this can be seen in Figure 3.17, where the range of superiority scores tends to be greater for genotypes tested in fewer environments. Figure 3.17 plots the reciprocals of adjusted superiority scores against the number of times a genotype was tested in the Onion Trials Programme. Scores for genotypes tested less than five times have been omitted as they showed extraordinarily wide-ranging results. The numerical values of adjusted superiority scores have been ignored in the graphic as it was deemed to be more informative to use a qualitative indicator of the measure. As discussed previously, in Section 2.7, many stability measures have little comparability between data sets because they are dependent on the data directly under examination. Adjusted superiority measures were calculated using the maximum values of environments within the data set concerned.

Reciprocals of adjusted superiority scores were used in order to place 'winners' at the top of the graph, rather than at the bottom. A more useful graphic might have the genotype names on the scatter plot to assist identification of these winners, but fewer points would appear on each scatter plot. This has not been presented because the interest was towards the pattern of performance, rather than particular performances.

Applying adjusted superiority scores to the entire Onion Data resulted in a large range of $p_i^\star$ scores for values of $P_{ii} \leqslant 4$. Of particular note were those high performing varieties that were tested in a single environment, including for example, a local Argentinean variety known as 'Angaco INTA' which was the top performing genotype in San Juan in 1997. This phenomenon occurred at five other sites in the Trials Programme; Table 3.9 lists this group of six one-off successes. This suggested that local varieties should be included in testing as they would potentially exhibit the best specific adaptation to the local environmental conditions.

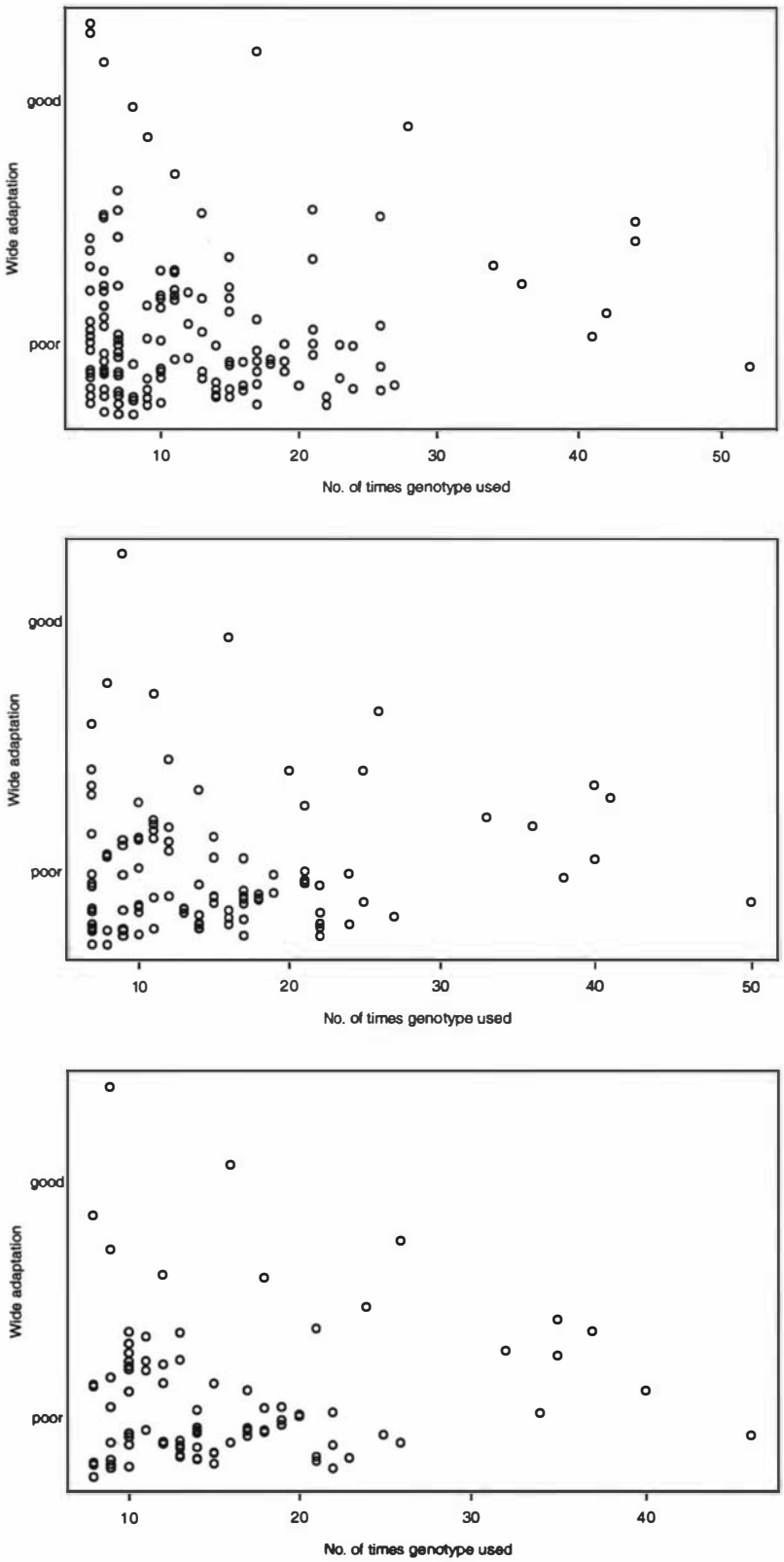


Figure 3.17: Reciprocals of adjusted Lin and Binns (1988a) superiority measures plotted against the number of times genotypes were used in each of three data sets; the entire Trials Programme data (top), Onion Data I (middle), and Onion Data II (bottom). Adjusted values were calculated using (3.8), which reflects the number of environments in which each genotype was tested.

3.8 Analysis with EM-AMMI

The EM-AMMI model of Gauch and Zobel (1990) was implemented using S-PLUS on both Onion Data I and II. The additive main effects and multiplicative interaction (AMMI) model

$$Y_{ik} = \mu + G_i + E_k + \sum_{n=1}^{N} \lambda_n u_{in} v_{kn} + \epsilon_{ik}$$

given in (2.12), and the process of using the Healy-Westmacott algorithm were described in Section 2.5 for fitting a model to incomplete data. EM-AMMI is a particular case of this algorithm which uses imputed values to fit a linear model. Table 3.10 shows the correlation between the fitted values from the models AMMI(0) to AMMI(6) for Onion Data I and II. These correlations are significant and positive between all pairs of models for both data sets; the same findings exist for all models when comparing results from one data set to the other, as presented in Table 3.11. This table also shows the number of

Correlation	AMMI(0)	AMMI(1)	AMMI(2)	AMMI(3)	AMMI(4)	AMMI(5)
AMMI(1)	0.702					
AMMI(2)	0.509	0.392		Onion Data I		
AMMI(3)	0.425	0.313	0.500			
AMMI(4)	0.405	0.297	0.496	0.533		
AMMI(5)	0.284	0.179	0.340	0.409	0.610	
AMMI(6)	0.240	0.145	0.257	0.351	0.528	0.644
AMMI(1)	0.739					
AMMI(2)	0.541	0.433		Onion Data II		
AMMI(3)	0.455	0.346	0.503			
AMMI(4)	0.446	0.350	0.414	0.575		
AMMI(5)	0.337	0.286	0.340	0.438	0.564	
AMMI(6)	0.293	0.248	0.285	0.390	0.493	0.603

Table 3.10: Correlation coefficients for fitted values found by applying various EM-AMMI models applied to Onion Data I and II.

	Number of iterations		Correlation
Model	Onion Data I	Onion Data II	of results
AMMI(0)	43	32	0.999
AMMI(1)	1729	1014	0.989
AMMI(2)	2499	1815	0.515
AMMI(3)	2054	1282	0.615
AMMI(4)	855	1297	0.722
AMMI(5)	911	1092	0.685
AMMI(6)	687	492	0.681

Table 3.11: Results of fitting various EM-AMMI models to Onion Data I and II. The number of iterations is recorded for each model applied to each data set, and the correlation of results between the two data sets for G×E combinations common to both.

	Onion Data I		Onion Data II	
Model	Negative	Excessive	Negative	Excessive
AMMI(0)	31	0	15	0
AMMI(1)	206	54	150	38
AMMI(2)	415	47	293	43
AMMI(3)	728	41	540	39
AMMI(4)	945	31	606	28
AMMI(5)	1266	11	856	24
AMMI(6)	1750	7	1095	5

Table 3.12: Unrealistic yields determined by fitting various EM-AMMI models to Onion Data I and II. Yields that are negative or in excess of 25 kilograms per square metre are considered unrealistic in this instance.

	Onion Data I		Onion Data II	
Model	No. of iterations	Correlation	No. of iterations	Correlation
AMMI(0)	43	1.000	32	1.000
AMMI(1)	4335	0.468	1795	0.591
AMMI(2)	2383	0.507	1432	0.449
AMMI(3)	913	0.338	1475	0.401
AMMI(4)	1139	0.356	715	0.439
AMMI(5)	449	0.402	439	0.375
AMMI(6)	386	0.337	305	0.346

Table 3.13: Effects of using the grand mean of the data set in place of a zero yield for initial imputed values. The number of iterations taken for the EM-AMMI model to converge is given along with the correlation coefficient of results found using the two different sets of initial imputed values.

iterations needed to reach the point where the largest change in an imputed yield was less than 0.005.

Table 3.12 shows the number of yields which were unrealistic because the imputed yields were either negative or well outside the observed range of yields in the Onion Trials Programme.

When applying the EM algorithm in conjunction with a model to impute missing values, it is necessary to choose starting values for the missing G×E combinations. In the results presented thus far, all missing G×E yields were first imputed using a zero yield. This may seem a poor choice when compared to using another means of getting initial estimates based on a more rigorous strategy. To gauge the performance of methods used to determine initial estimates the EM-AMMI models were re-fitted using the mean of observed data in each data set as the starting value. Correlation coefficients between results found using the two methods, and the number of iterations required for these models to converge to the same accuracy as above, are presented in Table 3.13. It was somewhat surprising to note that this second strategy which appears theoretically stronger

took longer to reach convergence. The two sets of imputed values are highly correlated for all EM-AMMI models applied to Onion Data I and II, but using different initial imputed values clearly resulted in a different final set of imputed values. The only exception is the AMMI(0) model which converged to the same set of values irrespective of the initial values chosen.

On the whole the set of EM-AMMI models were very consistent, but they provided answers that have little bearing on reality. The EM-AMMI modelling approach was therefore of little use in the analysis of the Onion Trials Programme. It is likely that this was a direct result of the inability of sparse data to find adequate parameters for fitting a specific model like AMMI.

3.9 Summary

This chapter outlined the history of the Onion Trials Programme, presented the preliminary data analysis, and discussed problems caused by the sparsity of the data. The chief aim of these first three chapters was to show that many methods for analysing G×E data (presented in Chapter 2) were inappropriate for use with incomplete data, that arose from the Onion Trials Programme. After adjustments were made to their formulation, several standard approaches were examined. As discussed in this chapter, the results from these attempts were useful for preliminary investigation of the data, rather than for rigorous parameter estimation which could be used for decision-making. The limitations of the most widely accepted method of fitting a model to incomplete data were exposed. Of the choices initially offered, only one option appears to remain, namely to:

> Devise a suitable method for imputing missing data, which would allow the use of standard G×E methodology to answer the principal research question.

Part II

Development of a Solution

Chapter 4

Distance measures

4.1 Introduction

Finding meaningful results from the Onion Trials Programme was difficult due to the incompleteness of the data. This research project was dominated by the inability of standard G×E analyses to accommodate incomplete data. The solution process presented over the next few chapters arose in a somewhat serendipitous manner. The first major development, a clustering method capable of handling incomplete data, is presented in the next chapter. Its development inspired the imputation process presented in Chapter 6, and the work that follows in subsequent chapters. The new clustering method requires suitable distance measures to function correctly. In this chapter, distance measures are developed, and their application to G×E analyses is considered.

Statistical packages generally offer the user few options for distance measures. The most widely available options are versions of the Minkowski metric, formally defined as

$$D_{ij}^{(Min)} = \left[\sum_{k=1}^{K} |y_{ik} - y_{jk}|^p\right]^{1/p} \tag{4.1}$$

where $p \geqslant 1$. Particular values of p will give common distance measures, such as Manhattan distance and Euclidean distance. Manhattan distance is the simplest form of the Minkowski metric obtained by setting $p = 1$, giving

$$D_{ij}^{(Man)} = \sum_{k=1}^{K} |y_{ik} - y_{jk}| \tag{4.2}$$

Euclidean distance, denoted E_{ij}, is found by setting $p = 2$ in (4.1), giving

$$E_{ij} = \sqrt{\sum_{k=1}^{K} (y_{ik} - y_{jk})^2} \tag{4.3}$$

Squared Euclidean distance E_{ij}^2 is also offered by many statistical software packages. Var-

ious forms of Euclidean distance were used throughout this investigation as it is the most commonly used distance measure for cluster analyses.

One of the specific aims of investigations covered in this chapter was to ensure that the new distance measures introduced were capable of handling varying amounts of data for comparisons among genotypes. Missing values in the G×E matrix would make the calculation of squared Euclidean distance possible only over environments in which both genotypes were grown. Ouyang *et al.* (1995) proposed dividing the sum of squared differences by the number P_{ij} of differences available, so that distances would not be adversely affected by the number of common environments in which the pair of genotypes was grown. This gives the expression

$$\frac{\sum_{\text{common } k} (y_{ik} - y_{jk})^2}{P_{ij}} \tag{4.4}$$

for a mean squared Euclidean distance.

The next section shows how two distance measures can be related to distinct aspects of a genotype's performance across environments. Euclidean distance is successfully partitioned into these two distance measures in Section 4.3, and shows how they can be applied to incomplete data. A method for estimating unmeasurable distances, which arise when working with sparse data, is presented in Section 4.6. Section 4.8 presents an argument for using one of the new distance measures instead of Euclidean distance when there is no significant interaction between observations (genotypes) and variables (environments). Other sections of this chapter cover material that was considered as the new distance measures were developed, including a survey of distance measures in Section 4.7.

4.2 Main effect and interaction distance

Chatfield and Collins (1980) summarized the manner in which this investigation has been conducted, stating "There is no such thing as the 'best' measure of dissimilarity. Rather, the researcher must choose the one which seems most appropriate for his particular problem. Indeed, the researcher must be prepared for situations in which there is no textbook solution...". In G×E analyses where a distance measure was required, usually cluster analyses (Lin and Butler, 1990) or pattern analyses (Mungomery *et al.*, 1974), a different distance measure has effectively been created by altering the data used in some way, including for example Ivory *et al.* (1991). Seldom has the distance measure been explicitly defined in terms of the raw data, and when this has been done, it has usually been to illustrate some relationship of interest (Abou-El-Fittouh *et al.*, 1969; Lin, 1982).

Distances, such as the Minkowski metric presented in (4.1), calculated between rows of raw data in a two-way table, often confound the row main effect and the two-way interaction. In G×E analyses, it has been commonplace to concentrate on the similarity of genotype performances in terms of G×E interaction alone. This has generally been done

by subtracting the row mean from each entry in the G×E matrix, so that the distance is formed by summing squared differences in the centred rows of the G×E matrix. That is, using

$$\left(D_{ij}^{(Int)}\right)^2 = \sum_{k=1}^{K}((y_{ik} - \bar{y}_{i\cdot}) - (y_{jk} - \bar{y}_{j\cdot}))^2 \tag{4.5}$$

to give a squared distance based on G×E interaction. Variants of this measure have been proposed. For example, the correlation between pairs of locations has been used (Abou-El-Fittouh, 1969). Lin (1982) showed that scaling all distances found using (4.5) by $2/(K-1)$ gives a measure that can be related to the interaction mean square in a two-way ANOVA table.

Lin (1982) stated that the adjusted rows of the G×E matrix are indicative of the 'shape' of a genotype's performance across environments, and termed the mean yield of each genotype its 'level'. Differences in these shapes or profiles indicate the existence of G×E interaction. Lin and Binns (1985) noted that if genotypes can be shown to have similar performance profiles across environments to varieties that exhibit well understood performances, the difference in mean performance can then be directly compared to provide a simple method of establishing variety recommendations. The terminology used by Lin (1982) is simple and is now explained further because it underpins many of the theoretical developments that follow.

A genotype performance can be compared to the performance of another genotype in two ways. First, by directly comparing the mean performances across all environments, and second, by the similarity of the environments to which the genotypes are suited. The second notion is related to the G×E interaction, or specific adaptation of the genotypes. It is known that the existence of interaction invalidates comparisons in terms of main effects alone. The interaction plot presented in Figure 4.1 shows two pairs of genotypes. One pair of genotypes, marked with solid lines, have similar mean performance across the environments, while the second pair, marked with dashed lines, are specifically adapted to the same environments. Mean performance across environments will be referred to throughout this investigation as 'level', while 'shape' will be used to indicate the pattern of genotype performances across environments. This pattern is also referred to as the 'interaction profile'.

The principal data of this investigation included a very large number of genotypes that needed to be compared. Cluster analysis was chosen as a tool that could be employed as it was descriptive and was understood by most researchers. Given the ability to use distance measures in cluster analyses that distinguish genotypes with similar interaction profiles, from those with dissimilar interaction profiles, there was then a need to develop a distance measure that would distinguish genotypes in terms of their level similarity.

The cluster analyses employed in this investigation also needed to be capable of handling missing data. Everitt (1993) noted some ways of dealing with missing values in

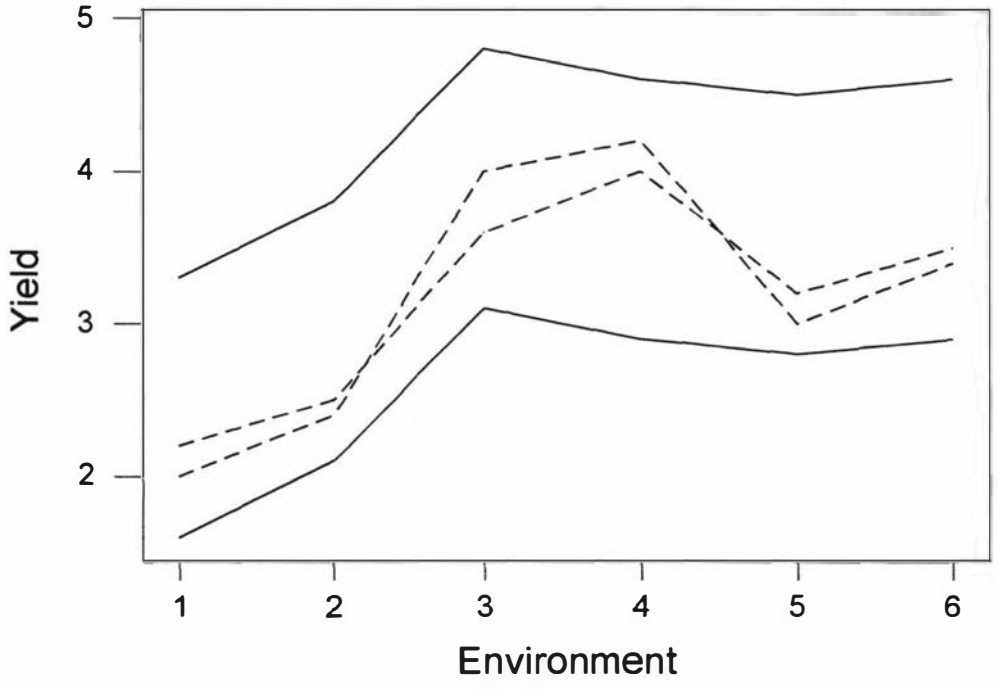


Figure 4.1: An illustration of the difference between the two ideas of 'shape' and 'level' using artificial data for two pairs of genotypes across six environments. Solid lines show a pair of genotypes with similar shape not level, while dashed lines show a pair of genotypes with similar level not shape.

cluster analyses:

1. Deleting observations with missing values.

2. Replacing missing values with estimated values. He noted that using the variable means was not advisable as the estimated values should be found using the data from similar observations, which have not yet been found.

It is worth noting that this second option has been used by Drake (1981), and in the new imputation methodology proposed in Chapter 6.

In Chapter 5 cluster analyses which use distance measures capable of handling missing data will be employed as an alternative to the options given by Everitt (1993). Simply ignoring the missing data is not a solution, as distance measures (in general) sum results over the range of variables for pairs of observations. A pair of observations that are measured over fewer variables will have a lower total than a pair measured over all variables. A measure that is akin to comparison of means is therefore required. Godfrey *et al.* (1999) presented a distance measure based on the difference in levels of genotypes in a G×E

matrix, referred to as the 'main effect distance' M_{ij}, where

$$M_{ij} = \frac{\left| \sum_{\text{common } k} (y_{ik} - y_{jk}) \right|}{P_{ij}}, \tag{4.6}$$

equivalent to

$$M_{ij} = \left| \bar{y}_{i\cdot}^{(j)} - \bar{y}_{j\cdot}^{(i)} \right| \tag{4.7}$$

with $\bar{y}_{i\cdot}^{(j)}$ defined as the mean of the yields y_{ik}, using only the P_{ij} environments common to both genotypes i and j. Note that this distance measure is different to the Manhattan distance measure given in (4.2), which accumulates positive and negative differences, whereas main effect distance allows them to cancel one another.

Use of main effect distance recognizes that the comparison of means as a measure of the difference in level of a pair of genotypes is not valid when some data are missing. It therefore considers only the data from the P_{ij} environments common to both genotypes in a similar fashion to the distance measure of Ouyang *et al.* (1995) given in (4.4).

What attributes should a distance measure have? The answer to this question will guide the use of distance measures, and provide methods to gauge the success of a new distance measure. These desired attributes need to be related to the mathematical properties of the proposed distance measure. In general, the aim of a distance measure is to provide a one-dimensional summary of the differences that exist between pairs of observations in a data set. This summary measure should:

1. Be near zero when there is no actual difference between the pair of observations.

2. Be near the actual difference when the pair of observations are clearly different.

It must be remembered that observations are actually measured with error, so cannot be expected to strictly meet these requirements. Instead of meeting them, a distance measure should be sought that is a function of the actual (error free) difference being measured. As formal properties of distance measures do not explicitly deal with the random component, later sections of this chapter deal with issues such as expected value and bias. The expression for main effect distance given in (4.6) can be rearranged to give

$$M_{ij} = \left| \bar{y}_{i\cdot}^{(j)} - \bar{y}_{j\cdot}^{(i)} \right| \tag{4.8}$$

and is therefore related to the difference of two means, and has the properties of means in general. Notably the expected value of the main effect distance $E(M_{ij})$ is equal to the difference between the pair of genotypes being compared if data are observed without error.

4.3 A partition of Euclidean distance

The familiar relationship

$$\sum_{k=1}^{K} y_k^2 = K\bar{y}^2 + (K-1)\frac{\sum_{k=1}^{K}(y_k - \bar{y})^2}{K-1} \tag{4.9}$$

partitions the sum of squares of observations $y_1, \ldots, y_K$ into orthogonal and independent components, the first related to the level or the sample mean and the second to the variability or the sample variance.

Let y_{ik} be the yield of the ith genotype in the kth environment. Substituting the difference $y_{ik} - y_{jk}$ of two genotypes in the kth environment for y_k gives

$$\sum_{k=1}^{K}(y_{ik} - y_{jk})^2 = K(\bar{y}_{i\cdot} - \bar{y}_{j\cdot})^2 + (K-1)\frac{\sum_{k=1}^{K}((y_{ik} - y_{jk}) - (\bar{y}_{i\cdot} - \bar{y}_{j\cdot}))^2}{K-1} \tag{4.10}$$

where $\bar{y}_{i\cdot}$ and $\bar{y}_{j\cdot}$ are the means of y_{ik} and y_{jk} respectively across all K environments. The left-hand-side of the last equation is the well known expression for squared Euclidean distance

$$E_{ij}^2 = \sum_{k=1}^{K}(y_{ik} - y_{jk})^2, \tag{4.11}$$

as used in G×E analyses (Mungomery *et al.* (1974) for example), while the right-hand-side exposes two components: a measure of the difference in genotype means and a measure of the G×E interaction. As has been discussed, the G×E interaction term has been used as a distance measure in clustering genotypes in the past (Lin, 1982).

Letting $I_{ij}^2 = \left(D_{ij}^{(Int)}\right)^2/(K-1)$ denote the squared interaction distance, the partition of squared Euclidean distance can be expressed as

$$E_{ij}^2 = K(\bar{y}_{i\cdot} - \bar{y}_{j\cdot})^2 + (K-1)I_{ij}^2 \tag{4.12}$$

This expression is now modified to allow for missing values in a G×E matrix.

Care is needed when centring the rows of G×E matrices with missing values, as now genotype means will generally be based on distinct sets of environments, rendering the orthogonal decomposition invalid. To remedy this, genotype means for each pair of genotypes i and j are calculated using only values from the P_{ij} environments for which both y_{ik} and y_{jk} have been recorded. The orthogonal partition of the squared Euclidean distance

for rows i and j is then

$$E_{ij}^2 = P_{ij}(\bar{y}_{i\cdot}^{(j)} - \bar{y}_{j\cdot}^{(i)})^2 + (P_{ij}-1)\frac{\sum_{\text{common } k}\left((y_{ik} - \bar{y}_{i\cdot}^{(j)}) - (y_{jk} - \bar{y}_{j\cdot}^{(i)})\right)^2}{P_{ij}-1} \tag{4.13}$$

The means used for row-centring genotypes and the partitioning of squared Euclidean distance are therefore dependent on the particular pair of genotypes that are being compared. The partition of squared Euclidean distance, when data is incomplete, uses P_{ij} in place of the total number of environments K in (4.12).

The partition of squared Euclidean distance can now be expressed completely in terms of main effect and interaction distances as

$$E_{ij}^2 = P_{ij}M_{ij}^2 + (P_{ij}-1)I_{ij}^2 \tag{4.14}$$

where the interaction distance I_{ij} is now given by

$$I_{ij} = \sqrt{\frac{\sum_{\text{common } k}\left((y_{ik} - \bar{y}_{i\cdot}^{(j)}) - (y_{jk} - \bar{y}_{j\cdot}^{(i)})\right)^2}{P_{ij}-1}} \tag{4.15}$$

This distance expression, which measures G×E interaction differences among genotypes, is appropriate when there are missing entries in the G×E matrix. Its construction takes two ideas into account: the value used for row-centring a given genotype is tailored to the other genotype in the pair, and as shown below, appropriate averaging is used.

Figure 4.2 shows the partition of Euclidean distance for a pair of genotypes measured in two environments. Note the Pythagorean triangle shows the orthogonality of main effect M_{ij} and interaction distance I_{ij} for genotypes i and j. In this complete data illustration, the main effect distance is equal to the length of the line segment that would join the projections of the points representing the yields of genotypes i and j onto the equiangular line where $y_{i1} = y_{i2}$. The plane that is normal to this line can be used to measure interaction distance in a similar way. Distances I_{ij} and M_{ij} will be used in Chapter 5 to cluster genotypes efficiently.

Relating the partition of Euclidean distance to a two-way model

The distance measures I_{ij} and M_{ij} are now related to a standard two-way model for G×E data. Such a model, assuming no replication in a cell, is

$$Y_{ik} = \mu + G_i + E_k + (GE_{ik} + \epsilon_{ik}) \tag{4.16}$$

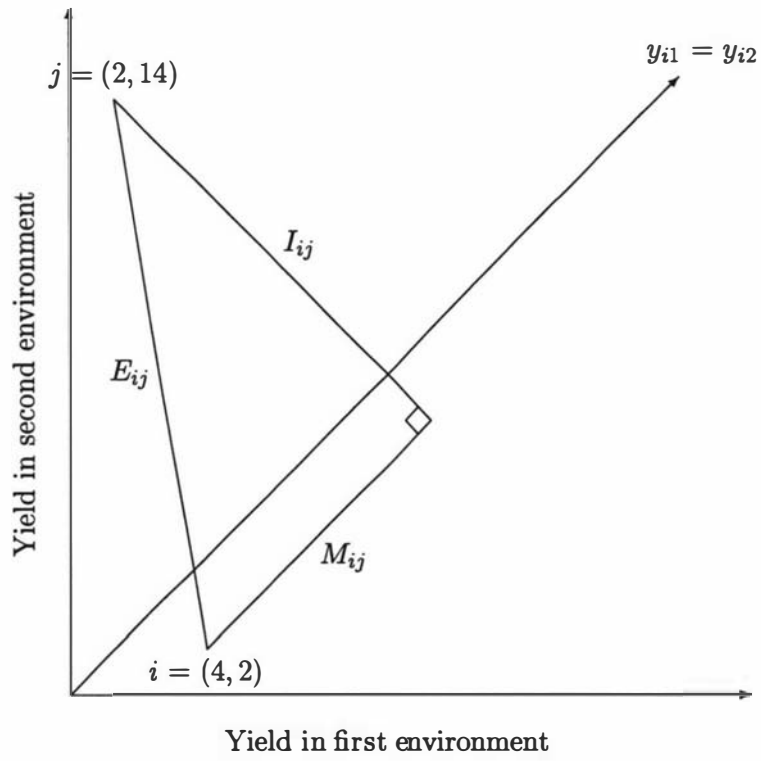


Figure 4.2: Graphical representation of the partition of Euclidean distance E_{ij} for two genotypes i and j in two environments. Note that the line segment representing main effect distance is parallel to the equiangular line for equal yields in all environments. The line segment for interaction distance I_{ij} is perpendicular to this line.

where the ϵ_{ik} are independent and normally distributed, with mean zero and variance σ^2. Note that the G×E interaction GE_{ik} and error ϵ_{ik} are confounded when there is no replication. Assuming that there is no missing data, the squared Euclidean distance between genotypes i and j is

$$E_{ij}^2 = \sum_{k=1}^{K} (\hat{G}_i - \hat{G}_j + \widehat{GE}_{ik} - \widehat{GE}_{jk} + e_{ik} - e_{jk})^2, \tag{4.17}$$

and therefore combines the genotype main effect and the difference in G×E interaction. Here a 'hat' denotes an estimator; note that $\bar{y}_{i\cdot} = \hat{\mu} + \hat{G}_i$. Also,

$$\left(D_{ij}^{(Int)}\right)^2 = \sum_{k=1}^{K} \left((\widehat{GE}_{ik} - \widehat{GE}_{jk}) + (e_{ik} - e_{jk}) \right)^2, \tag{4.18}$$

and,

$$M_{ij}^2 = \left(\sum_{k=1}^{K} \left((\hat{G}_i - \hat{G}_j) + (\widehat{GE}_{ik} - \widehat{GE}_{jk}) + (e_{ik} - e_{jk}) \right) \right)^2 / K^2 \tag{4.19}$$

Note that when genotypes i and j have the same interaction pattern (so $GE_{ik} = GE_{jk}$, for all k), $\left(D_{ij}^{(Int)}\right)^2 / 2\sigma^2$ will follow a χ^2_{K-1} distribution and hence $\left(D_{ij}^{(Int)}\right)^2$ will have expected value $2\sigma^2(K-1)$. Thus the expected value of $I_{ij}^2 = \left(D_{ij}^{(Int)}\right)^2 / (K-1)$ is $2\sigma^2$ and so does not depend on the number of environments. This ensures comparability of the I_{ij}^2 interaction distance measures, from one pair of interaction similar genotypes to another, when missing values are encountered.

When genotypes i and j have the same interaction profile, $KM_{ij}^2/2\sigma^2$ follows a non-central χ_1^2 distribution, with non-centrality parameter $\sqrt{K/2\sigma^2}$. It follows that M_{ij}^2 has expected value

$$2\sigma^2/K \; + (\hat{G}_i - \hat{G}_j)^2 \tag{4.20}$$

The quantity $2\sigma^2/K$ found in this expectation is generally small when compared to $(\hat{G}_i - \hat{G}_j)^2$. Thus M_{ij}^2 serves as a satisfactory measure of difference in genotype level.

4.4 Computation of main effect and interaction distances

Discussion of the formulation of interaction and main effect distances would be incomplete without consideration of the practical computation of the distance measures. S-PLUS routines have been written for the two distance measures and can be found on the CD-ROM accompanying this volume.

The functionality of S-PLUS for handling missing values can be used to advantage when calculating main effect and interaction distances. As a particular example, the main effect distance in its original form suggests taking the mean of the values of each genotype after discarding the environments they do not have in common with each other; then finding the difference in these means. The fact that the mean of differences is equal to the difference of means is used in the following way. A missing value in either part of a difference causes S-PLUS to return a missing value. When the mean of this vector of differences is found, S-PLUS discards all missing values, giving a mean of the available differences that come from environments in which both genotypes were grown. Interaction distance can be found in similar fashion, using the unbiased sample standard deviation of the vector of differences.

The matrix of interaction distances can be shown to be the square roots of the entries of the sample covariance matrix, if the missing data is handled appropriately. Unfortunately, existing routines attempt to find the sample standard deviation even when there is only one available observation to use, returning values of positive or negative infinity. In this instance the value returned should be 'undefined', expressed by S-PLUS as 'NA'. Correc-

tion of this anomaly uses computation time and leaves this method of finding interaction distance less efficient.

Use of the partition of Euclidean distance has proved advantageous in the calculation of interaction distance. Interaction distance is found using

$$I_{ij} = \sqrt{\frac{E_{ij}^2 - P_{ij}M_{ij}^2}{P_{ij} - 1}} \tag{4.21}$$

and is more efficient when the matrices of distances are used, rather than looping through pairwise combinations of observations.

4.5 The metric nature of main effect and interaction distances

Mathematicians have used the term 'metric' to define distance measures D_{ij} that meet certain axioms. These formal properties of a metric are:

1. $D_{ij} \geqslant 0$, for all i, j.
2. $D_{ij} = 0, \iff i = j$.
3. $D_{ij} = D_{ji}$, for all i, j.
4. $D_{hj} \leqslant D_{hi} + D_{ij}$, for all h, i, j.

This fourth axiom is known as the triangle inequality, and is used in the following section as a mechanism for estimating unobserved distances. This section shows that main effect and interaction distances are metrics when working with complete data, but are not necessarily so when working with incomplete data.

The theory of metrics shows:

1. The sum of two metrics is also metric.
2. If D_{ij} is metric, and w a positive constant, then $D'_{ij} = \frac{D_{ij}}{w + D_{ij}}$ is also a metric.
3. The square of a metric is not guaranteed to be metric as the triangle inequality may not hold (Chatfield and Collins, 1980).

In their discussion of metrics and the use of dissimilarity measures, Chatfield and Collins (1980) point out that the triangle inequality has its uses, but that a metric is not always the best measure to use.

Main effect distance and complete data

Main effect distance can be expressed as

$$\begin{aligned} M_{ij} &= \left| \frac{1}{K} \sum_{k=1}^{K} x_{ik} - x_{jk} \right| \\ &= |\bar{x}_{i\cdot} - \bar{x}_{j\cdot}| \end{aligned} \tag{4.22}$$

when there is complete information. Using three observations h, i, and j and without loss of generality letting $\bar{x}_{h\cdot} \geqslant \bar{x}_{i\cdot} \geqslant \bar{x}_{j\cdot}$, the main effect distance is metric when working with complete data if $M_{hj} \leqslant M_{hi} + M_{ij}$.

$$\begin{aligned} \bar{x}_{h\cdot} - \bar{x}_{j\cdot} &\leqslant \bar{x}_{h\cdot} - \bar{x}_{i\cdot} + \bar{x}_{i\cdot} - \bar{x}_{j\cdot} \\ |\bar{x}_{h\cdot} - \bar{x}_{j\cdot}| &\leqslant |\bar{x}_{h\cdot} - \bar{x}_{i\cdot}| + |\bar{x}_{i\cdot} - \bar{x}_{j\cdot}| \\ M_{hj} &\leqslant M_{hi} + M_{ij} \end{aligned} \tag{4.23}$$

So the triangle inequality holds for main effect distance, thus proving that main effect distance is a metric when working with complete data.

Interaction distance and complete data

Euclidean distance is known to be a metric (Chatfield and Collins, 1980). Interaction distance can be found in complete data circumstances by calculating Euclidean distance on the row centred G×E matrix.

Figure 4.2 showed the relationship between main effect, interaction, and Euclidean distances. Interaction distance can be viewed as the length of the line joining the projections of the points representing the two observations onto the hyperplane that is orthogonal to the equiangular line that represents the mean of all variables. The lines joining any three points on this hyperplane form a triangle, over which Euclidean distance can be calculated. The triangle inequality must hold, and therefore, interaction distance is a metric when working with complete data.

Main effect and interaction distance with incomplete data

In incomplete data circumstances the metric nature of main effect, interaction and Euclidean distance does not hold. The simplest way to illustrate this fact is by using a proof by contradiction.

Take for example, the incomplete G×E matrix $\begin{bmatrix} 2 & 2 & 2 & - \\ - & 3 & 3 & 3 \\ 6 & - & 4 & 4 \end{bmatrix}$. The distance measures between rows of this matrix would be

Distance Measure	Equation No.	dhi	dij	dhj
Euclidean	4.4	1	1	$\sqrt{10}$
Interaction	4.15	$\sqrt{2}$	$\sqrt{2}$	$\sqrt{22}$
Main effect	4.6	1	1	3

The distance between genotypes (rows) h and j is greater than the distance between genotypes h and i added to the distance between genotypes i and j for the three distance measures. The triangle inequality does not necessarily hold for main effect, interaction, and Euclidean distances, so they are not metric when working with incomplete data.

Consequences of non-metric nature of distances with incomplete data

Chatfield and Collins (1980) noted that it is not necessary to use a metric distance measure in cluster analyses, but the next section uses the triangle inequality to estimate unobserved distances. If the data are assumed to be 'missing completely at random' (MCAR), as defined on page 58, the fact that main effect and interaction distances are not metric when working with incomplete data can be disregarded.

Empirical distance measures are based on data that is observed with error, and are therefore subject to some sampling error. The range of variables over which distance measures are calculated is also a sample of the population of variables. When working with a subset of the set of variables observed, as is the case when working with incomplete data, the sampling nature is even more obvious.

When data are MCAR, observed distances are calculated over a random subset of variables and can be assumed to have the same expected value as distances calculated over complete data. In some situations where data are not MCAR, but are 'missing at random' (MAR) the same conclusion may be drawn. If genotypes were grown in a subset of environments covering all conditions, but in no more than two environments of the same type, data would be MAR. In this situation the incompleteness may be sufficiently random to allow its analysis as if it were MCAR. Problems may be encountered when pairs of genotypes are not grown in any environments that would have highlighted their differences. In such situations the data is censored to the fact that this pair of genotypes are not as similar as the missing at random data would suggest.

If genotypes that were known to fail at certain locations had not been included in the data set, the distances between these genotypes and those that succeed in those environments would be less than the current estimates. These data would then not be classified as MAR, let alone MCAR, and the estimates of genotype differences would be flawed.

Observed distances will be good estimates of the 'true' differences between pairs of genotypes if the data are MCAR, but if the genotypes have been grown in enough dissimilar environments to highlight their differences, having data that are MAR should be sufficient. All observed distances can then be assumed to be estimates of the 'true' distance, irrespective of the number of environments over which they were calculated. They

can therefore, be used in the following section to estimate unobserved distances using the triangle inequality, and in cluster analyses in subsequent chapters.

4.6 Estimating unobserved distances

As discussed in Section 2.6, cluster analyses have been used frequently in G×E analyses to discover relationships among genotypes or environments. The sparse data arising from the Onion Trials Programme could not be used in a standard cluster analysis because some inter-genotype distances were unobserved, and others were based on a small number of environments. A strategy for estimating unobserved distances was required, irrespective of the distance measure to be employed. This section shows how an acceptable work around for this problem was found.

Existing methods for finding distances with incomplete data

Lawrence and DeLacy (1993) had twelve G×E matrices, one for each year that they used to find inter-location distance matrices. They used squared Euclidean distance and location centred data, which had the effect of using G×E interaction differences between locations. They broke their data into two six year periods as the genotypes and locations changed over this time. This separation was appropriate as a means of reducing the sparsity of their data set — a luxury that was not an option for the data arising from the Onion Trials Programme. For each of these two time periods, six inter-location distance matrices were averaged to give one inter-location distance matrix. They also suggested that the approach should be repeated when the set of tested genotypes has significantly altered, to re-evaluate the relationships among test sites.

Lawrence and DeLacy (1993) averaged over years without weighting distance matrices to account for the amount of information that was obtained in each year. If this had been done, the distance matrix would be heavily weighted to the years when more lines were grown. This non-weighting however does not weight each line equally and therefore the distance obtained is not equivalent to the interaction distance presented in this chapter.

All pairs of locations were able to have an inter-location distance calculated in the study of Lawrence and DeLacy (1993), due to the averaging over six years of data and the fact that every pair of locations was used in the same season at least once. Ouyang *et al.* (1995) used the maximum value in the observed distance matrix to estimate unobserved inter-location distances. Their use of the maximum value in the distance matrix was justified as the selection of genotypes on trial was based on geographic location. Distances between locations that had little commonality of genotype test sets were likely to be large.

Ouyang *et al.* (1995) discounted the distances measured over a low number of common genotypes. They chose to combine the observed distance with the maximum of all observed distances $D_{ij}^{(Max)}$ in the following way.

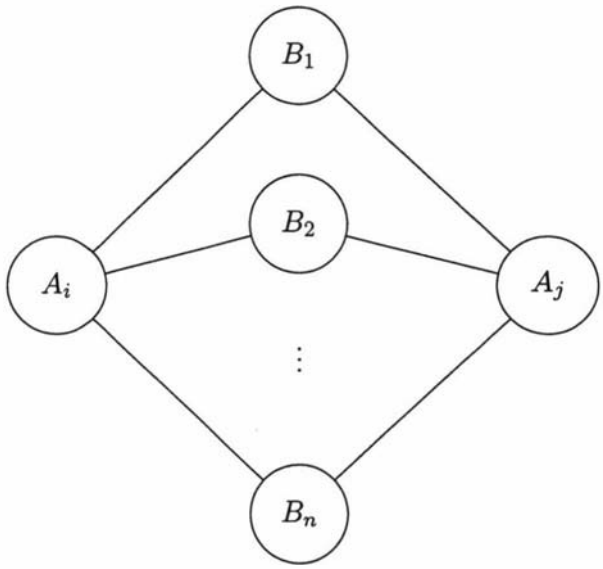


Figure 4.3: Two genotypes A_i and A_j have no direct link. They both, however, have a link to n other genotypes, thus forming indirect links between A_i and A_j .

For any pair of genotypes the number of common environments P_{ij} in which both were grown is examined. If $P_{ij} \geqslant q$, where q is a pre-determined number, the distance was deemed to have been calculated over a sufficient number of environments, and the observed distance was used in clustering. If $1 \leqslant P_{ij} < q$, the distance was estimated using

$$\left(P_{ij}D_{ij} + (q - P_{ij})\, D_{ij}^{(Max)}\right)/q \tag{4.24}$$

where D_{ij} is the observed distance between observations i and j. An adaptation of the strategy of Ouyang *et al.* (1995) for estimating unavailable distances between pairs of genotypes and adjusting distances based on few comparisons is now described.

The new proposed method

As with Ouyang *et al.* (1995), a method for estimating unobserved distances, where $P_{ij} = 0$, is given and subsequently used to adjust distances based on lower numbers of common environments, that is where $0 < P_{ij} < q$. The strategy of Ouyang *et al.* (1995) for estimating unobserved distances is not satisfactory, so use of the following 'shortest path' estimate is proposed.

Suppose that the two genotypes A_i and A_j in Figure 4.3 do not have any environments in common, but n other genotypes $B_1, \ldots, B_n$ share at least q environments with both A_i and A_j. An upper bound $D_{ij}^{(Upp)}$ for the distance between A_i and A_j can be estimated as the minimum of the summed pairs of distances,

$$D_{ij}^{(Upp)} = \min\left\{d(A_iB_1) + d(B_1A_j), \ldots, d(A_iB_n) + d(B_nA_j)\right\} \tag{4.25}$$

The d used in this expression denotes a distance function, whether it be squared Euclidean, interaction, or main effect distance. This upper bound can then be used as a conservative estimate of the unobserved distance.

A general rule for all distances measured over a small number of environments, $0 \leqslant P_{ij} < q$, can now be found by extending the Ouyang *et al.* (1995) strategy. The observed distance D_{ij} and $D_{ij}^{(Upp)}$ are merged using

$$\left(P_{ij}D_{ij} + (q - P_{ij})\, D_{ij}^{(Upp)}\right)/q \tag{4.26}$$

to give an estimated distance which combines direct and indirect information. This provides a complete set of distances, allowing clustering (Chapter 5), and then imputation (Chapter 6) to be performed.

4.7 A survey of distance measures

In previous sections of this chapter some distance measures used in G×E analyses have been introduced. During the preliminary investigations in relation to distance measures it was necessary to review the literature to ensure that the developments of this chapter did not repeat the work of others.

The distance measures covered in this section are restricted to those applicable to continuous measurements of observations over a range of variables, and therefore can be applied to genotypes or environments in G×E analyses. Measures that were designed for binary data will not be covered, although some of the measures given could be applied to binary data. Asymmetric distance measures and those created for a specific graphic such as those presented in Chapter 2 are not reviewed in this section either.

This collection of distance measures was accumulated from a number of references, including Anderberg (1973), Chatfield and Collins (1980), Gower (1985), Krzanowski (1988), Everitt (1993), and Manly (1994). Formulae given have been presented in their simplest form, using the standard notation of this investigation for ease of comparison.

As noted in Section 4.1, some of the most commonly used distance measures are forms of the Minkowski metric, given in (4.1), including:

1. The city-block, Manhattan, or taxi-cab metric, where $p = 1$. It usually appears in the form presented in (4.2).

2. Euclidean distance, where $p = 2$, given in (4.3).

3. The Chebychev metric, found as the limit of $D_{ij}^{(Min)}$ as $p \to \infty$, or more simply $\max_k \{(y_{ik} - y_{jk})\}$.

The Minkowski metric is presented differently according to the preference of the source. For example, Gower (1985) expressed the metric in such a way that included a 'normalizing'

factor r_k, and scaled the result by a factor related to the number of variables K. Gower's (1985) representation is

$$D_{ij}^{(G)} = \left[\frac{1}{K} \sum_{k=1}^{K} \frac{|y_{ik} - y_{jk}|^p}{r_k^p} \right]^{1/p} \tag{4.27}$$

Gower (1985) then defined the 'taxonomic distance' as the particular case of (4.27) where $p = 2$. Gower (1985) presented some options for the values r_k, including the standard deviation and sample range of $y_{\cdot k}$. The Penrose distance is a specific case of (4.27) and can be found using

$$D_{ij}^{(P)} = \sum_{k=1}^{K} \frac{(y_{ik} - y_{jk})^2}{K\sigma_k^2} \tag{4.28}$$

where σ_k^2 is the variance of the kth variable. Mahalanobis distance is one of the most well-known distance measures used in multivariate statistics. Its particular importance is that it overcomes the tendency of correlated variables to provide the same information on the observations being compared; having too many variables measuring the same underlying attribute would disproportionately weight the distance measure towards that attribute. Mahalanobis distance is found using

$$D_{ij}^{(Mah)} = \sum_{k=1}^{K} \sum_{m=1}^{M} (y_{ik} - y_{jk})\, v_{km}\, (y_{im} - y_{jm}) \tag{4.29}$$

where i and j refer to observations (genotypes), k and m are variables (environments), and v_{km} refers to the element in the kth row and mth column of the inverse covariance matrix of variables. Manly (1994) observed that when there are few observations to calculate the covariance matrix, Penrose distance should be used instead of Mahalanobis distance. Unfortunately, he did not offer a threshold to surpass, but did suggest that more than 100 degrees of freedom when calculating the covariance matrix should be ample for use of Mahalanobis distance. This will prove difficult for the majority of G×E data to surpass; it is certainly the case for the Onion Trials Data.

Mahalanobis and Penrose distances have mainly been used for calculating the distance between two or more groups in a population, where the difference of the group means for each variable is used in the numerator of (4.28) or (4.29). The ability to use group means in this way needs to be carefully considered when data is incomplete.

The following list of distance measures is given with additional comments where appropriate. Some distance measures are closely related and these relationships have been noted.

1. The Lance and Williams non-metric distance measure

$$D_{ij}^{(L\&W)} = \frac{\sum_{k=1}^{K} |y_{ik} - y_{jk}|}{\sum_{k=1}^{K} (y_{ik} + y_{jk})}. \tag{4.30}$$

was presented by Anderberg (1973). This distance measure is unlikely to be of use in G×E analyses due to the way the denominator impacts on results. A difference between a pair of genotypes will be weighted by the sum of their performances across environments. The higher this sum, the lower the effect of the difference in the numerator. It does, however, have the advantage that it counters the effect of increasing variance with increasing mean, common to agricultural data. Using this measure on data that has negative values would lead to problems as there is a risk that a negative distance will result. Gower (1985) attributed this distance to a different source (Bray and Curtis), but does not provide a complete reference.

2. The Canberra metric as described by Gower (1985), is found using

$$D_{ij}^{(Ca)} = \sum_{k=1}^{K} \frac{|y_{ik} - y_{jk}|}{|y_{ik}| + |y_{jk}|}. \tag{4.31}$$

Its limitation is that differences of the same magnitude will be treated differently depending on the values of the observations being compared. A pair of higher performing genotypes would then be deemed more similar than a pair of lower performing genotypes even though the differences between the pairs was exactly the same. This concern is common to the Lance and Williams distance, but is specific to the performances within a variable rather than across all variables. The absolute value signs are often removed from (4.31) when $y_{ik} \geqslant 0, \forall\, i, k$.

3. The Soergel distance measure replaces the denominator of (4.30) with the maximum of the two observations for each variable. It was given by Gower (1985) as

$$D_{ij}^{(S)} = \frac{\sum_{k=1}^{K} |y_{ik} - y_{jk}|}{\sum_{k=1}^{K} \max\{y_{ik}, y_{jk}\}}. \tag{4.32}$$

4. The Czekanowski coefficient

$$D_{ij}^{(Cz)} = 1 - \frac{2\sum_{k=1}^{K} \min\{y_{ik}, y_{jk}\}}{\sum_{k=1}^{K} (y_{ik} + y_{jk})} \tag{4.33}$$

also appears similar to the Lance and Williams distance of (4.30), and is non-metric.

5. The divergence distance is given as

$$D_{ij}^{(D)} = \sqrt{\frac{1}{K}\sum_{k=1}^{K} \frac{(y_{ik} - y_{jk})^2}{(y_{ik} + y_{jk})^2}}. \tag{4.34}$$

6. The angular separation distance can be found using

$$D_{ij}^{(A)} = \frac{\sum_{k=1}^{K} y_{ik} y_{jk}}{\sqrt{\sum_{k=1}^{K} y_{ik}^2 \sum_{k=1}^{K} y_{jk}^2}}. \tag{4.35}$$

7. The Ware and Hedges distance measure was given by Gower (1985) as

$$D_{ij}^{(W\&H)} = \frac{1}{K}\sum_{k=1}^{K}\left(1 - \frac{\min\{y_{ik}, y_{jk}\}}{\max\{y_{ik}, y_{jk}\}}\right) \tag{4.36}$$

but could be represented in another way using absolute values rather than minima, as

$$\begin{aligned} D_{ij}^{(W\&H)} &= \frac{1}{K}\sum_{k=1}^{K}\left(1 - \frac{\min\{y_{ik}, y_{jk}\}}{\max\{y_{ik}, y_{jk}\}}\right) \\ &= \frac{1}{K}\sum_{k=1}^{K}\left(\frac{\max\{y_{ik}, y_{jk}\}}{\max\{y_{ik}, y_{jk}\}} - \frac{\min\{y_{ik}, y_{jk}\}}{\max\{y_{ik}, y_{jk}\}}\right) \\ &= \frac{1}{K}\sum_{k=1}^{K}\frac{|y_{ik} - y_{jk}|}{\max\{y_{ik}, y_{jk}\}}. \end{aligned} \tag{4.37}$$

The Ware and Hedges distance is therefore a mean of ratios, whereas the Soergel distance is the ratio of sums using the same quantities. Many cases of the rearrangement in the order of operations exist in distance measure formulation, as illustrated by this relationship and that of the Manhattan and main effect distance measures.

Other distance measures have been proposed that have algorithmic rather than functional methods of determination. For example, the Calhoun distance is based on the

number of 'third-party' observations that have a response between a pair of observations in at least one variable. The algorithm counts the observations that are within three categories:

1. N_i, the number of points between the two points of interest on at least one variable.

2. N_b, the number of points that do not fall between the two points of interest for any variable, but are equal to one or more of them.

3. N_z, the number of points which tie with both points of interest; that is, all three points share the same value.

These three quantities are used to calculate the normalized Calhoun distance using

$$D_{ij}^{(NC)} = \frac{6N_i + 3N_b + 2N_z}{6\,(N-2)} \tag{4.38}$$

where N is the total number of observations. It is not a metric for three reasons:

1. Two points need not be identical to have a Calhoun distance of 0.

2. Two points that are identical need not have a zero distance.

3. The triangle inequality does not necessarily hold.

This distance measure was not applied in this investigation, mainly for the reason that it is algorithmic rather than formulaic, making its calculation in a large data set time-consuming.

Distance and dissimilarity measures can be constructed from similarity measures. For example, correlation coefficients have been used in G×E analysis (Abou-el-Fittouh *et al.*, 1969). This is possible because all variables are in the same units in a data set and no transformation is required. The correlation coefficient ρ_{ij} between a pair of genotypes i and j can be converted to a distance measure using

$$D_{ij}^{(Cor)} = 2\,(1 - \hat{\rho}_{ij})\,, \tag{4.39}$$

which results in a metric distance measure.

The majority of distance measures presented in this section cannot be applied to incomplete data without having their formulation altered to reflect the differing numbers of variables over which calculations are made. Squared Euclidean distance has already been modified to account for incomplete data by Ouyang *et al.* (1995), and in this chapter, two distance measures were described that measure two aspects of genotype performances while allowing for the incompleteness of data.

4.8 Comparison of main effect and Euclidean distances

It was necessary to prove that main effect distance is a better method of distinguishing a difference than Euclidean distance when there is a constant difference between a pair of observations across a set of variables. In a G×E context, this means that there is no significant G×E interaction in the data. Establishing the advantage of one distance over the other when there is no significant interaction was done by finding the distribution of the main effect and Euclidean distances under two scenarios. First, when there is a known difference in the pair, and second when no difference exists. The ratio of these two values will be used to distinguish these scenarios; the higher ratio will indicate the better performance.

All distance measures over-estimate the difference between a pair of genotypes because observed distances are generally positive, even when no actual difference exists. A ratio has been used to show how well an actual difference is reflected by the observed distance compared to the observed distance based on error alone. The additive two-way model

$$Y_{ik} = \mu + G_I + E_k + \epsilon_{ik}$$

of (2.1) is assumed to hold for this analysis. There is no interaction and the ϵ_{ik} are assumed to be independent and normally distributed with mean 0 and variance σ^2.

Main effect distance for a known underlying difference

Main effect distance can be expressed as the absolute value of the mean difference between two genotypes across all environments. The actual expression used in complete G×E matrices is

$$\begin{aligned} M_{ij} &= \frac{1}{K}\left|\sum_{k=1}^{K} y_{ik} - y_{jk}\right| \\ &= \frac{1}{K}\left|\sum_{k=1}^{K}[(G_i + \epsilon_{ik}) - (G_j + \epsilon_{jk})]\right| \qquad (4.40) \\ & \qquad (4.41) \end{aligned}$$

When $G_i - G_j = d$ across all K environments the main effect distance M_d can be expressed as

$$M_d = \frac{1}{K}\left|Kd + \left(\sum_{k=1}^{K}\epsilon_{ik} - \epsilon_{jk}\right)\right| \qquad (4.42)$$

which expressed in terms of normally distributed random variables is

$$\begin{aligned} M_d &\sim \frac{1}{K}\left|\mathcal{N}\left\{Kd, 2K\sigma^2\right\}\right| \qquad (4.43) \\ &\sim \frac{1}{K}\left[\mathcal{N}\left\{Kd, 2K\sigma^2\right\}^2\right]^{1/2} \qquad (4.44) \end{aligned}$$

Standardizing the normal variate to get unitary variance gives

$$\begin{aligned} M_d &\sim \frac{1}{K}\left[\left(\sigma\sqrt{2K}\,\mathcal{N}\left\{\frac{d}{\sigma}\sqrt{\frac{K}{2}},1\right\}\right)^2\right]^{1/2} \\ M_d^2 &\sim \frac{2\sigma^2}{K}\left[\mathcal{N}\left\{\frac{d}{\sigma}\sqrt{\frac{K}{2}},1\right\}\right]^2 \end{aligned} \tag{4.45}$$

The special case when $d = 0$ (no underlying difference) reduces (4.45) to be

$$M_0^2 \sim \frac{2\sigma^2}{K}\left[\mathcal{N}\{0,1\}\right]^2 \tag{4.46}$$

Expressing the square of the main effect distance as a χ^2-distributed random variable, rather than a normally distributed random variable gives

$$M_d^2 \sim \frac{2\sigma^2}{K}\,\chi_1'^{\,2}\left\{\frac{d}{\sigma}\sqrt{\frac{K}{2}}\right\} \tag{4.47}$$

and has expected value

$$\mathcal{E}\left[M_d^2\right] = \frac{2\sigma^2}{K}\left[1+\frac{Kd^2}{2\sigma^2}\right] \tag{4.48}$$

$$= \frac{2\sigma^2}{K} + d^2 \tag{4.49}$$

Euclidean distance for a known underlying difference

When working with complete data the Euclidean distance is the square root of the sum of K squared normally distributed differences. The Ouyang adjusted distance has been used so that changes in the number of environments K does not affect the magnitude of the observed distance

$$E_d \sim \left[\frac{1}{K}\left(\sum_{k=1}^{K}\mathcal{N}\left\{d,2\sigma^2\right\}^2\right)\right]^{1/2} \tag{4.50}$$

standardizing the normal variates gives

$$E_d \sim \left[\frac{1}{K}\left(\sum_{k=1}^{K}\sigma\sqrt{2}\,\mathcal{N}\left\{\frac{d}{\sigma\sqrt{2}},1\right\}^2\right)\right]^{1/2} \tag{4.51}$$

and rearranging gives

$$E_d^2 \sim \frac{2\sigma^2}{K}\sum_{k=1}^{K}\left[\mathcal{N}\left\{\frac{d}{\sigma\sqrt{2}},1\right\}\right]^2 \tag{4.52}$$

The special case when there is no underlying difference reduces this last expression to

be

$$E_0^2 \sim \frac{2\sigma^2}{K} \sum_{k=1}^{K} [\mathcal{N}\{0,1\}]^2 \tag{4.53}$$

Expressing the distribution of E_d^2 as a χ^2-distributed random variable rather than the sum of squared normally distributed random variables gives

$$E_d^2 \sim \frac{2\sigma^2}{K} \chi'^2_K \left\{ \frac{d}{\sigma\sqrt{2}} \right\} \tag{4.54}$$

which has expected value

$$\mathcal{E}\left[E_d^2\right] = \frac{2\sigma^2}{K} \left[K + \frac{d^2}{2\sigma^2} \right] \tag{4.55}$$

Comparing the ratios for main effect and Euclidean distances

The expected value of a non-central χ^2-distributed random variable is

$$\mathcal{E}\left[\chi'^2_v\{\lambda\}\right] = v + \lambda^2 \tag{4.56}$$

Letting $\mathcal{R}[M]$ denote the ratio of the expected value of the main effect distance when there is some underlying difference d, to the expected value of the main effect distance when there is no underlying difference $\mathcal{E}[M_0]$, and similarly for the Euclidean distance gives the expressions

$$\mathcal{R}[M] = 1 + \frac{Kd^2}{2\sigma^2} \tag{4.57}$$

and

$$\mathcal{R}[E] = 1 + \frac{d^2}{2K\sigma^2} \tag{4.58}$$

Therefore the ratio of squared main effect distance is always greater than the same ratio for squared Euclidean distance. The squared distance measures should not be used when working with expected values, however, because the square root of the expected value of the squared distances is not equal to the expected value of the distances.

A comparison using expected values and mean squared error

It proved easier to compare the expected distances by simulating distances given a certain underlying difference between genotypes, and subsequently obtaining a plot of the expected value and MSE of distances against the underlying difference. Adjusting the underlying difference d and the variance of the randomly generated data for both main effect and Euclidean distance gave the plots found in Figures 4.4 and 4.5. In these figures the expected value and MSE were found for a series of underlying differences in the range zero to six. The standard error of the additive two-way model, given in (2.1), was chosen at six different levels (0.25, 0.50, 0.75, 1, 2, and 3) and provides the six lines in these figures. The number of environments was held constant at five, but altering this would result in plots

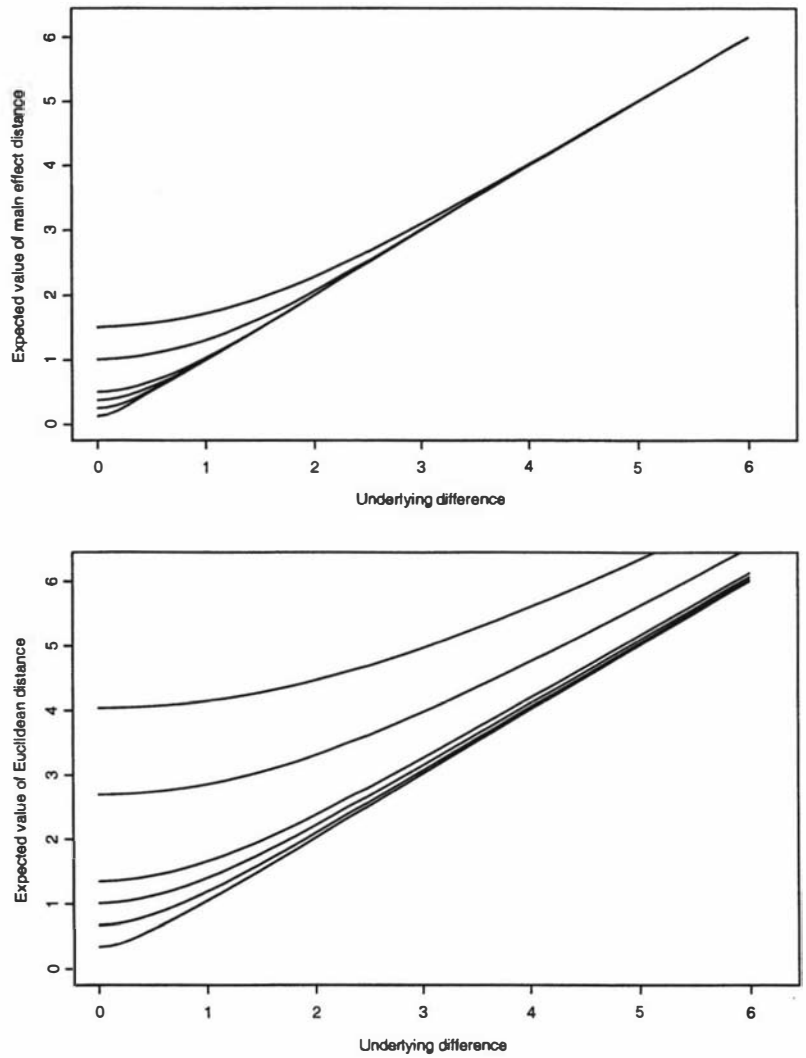


Figure 4.4: The expected value of main effect (upper) and Euclidean (lower) distances plotted against underlying differences between a pair of observations. Six different levels of error variance (3, 2, 1, 0.75, 0.5, and 0.25) in the observations were used giving the six lines from top to bottom respectively in each panel.

similar to those found for the varying standard deviations used in Figures 4.4 and 4.5. As the number of environments increases, the distance measured is closer to the underlying difference.

The MSE of the distance measures is a combination of the bias of the estimator and the variance of the estimate. From Figure 4.4 it can be seen that both the main effect and Euclidean distance measures are biased when the underlying difference is small compared to the error variance of observations. The graphics in Figure 4.5 show that when the underlying difference is low, the MSE of main effect distance is lower than the MSE of Euclidean distance. The amount of variation in the estimated distance should be reduced due to the restricted (non-negative) range of values the estimate can take, when the underlying difference is low.

There appears to be a tradeoff between bias and variance for the main effect distance when the underlying difference is low, but for underlying differences greater than the error

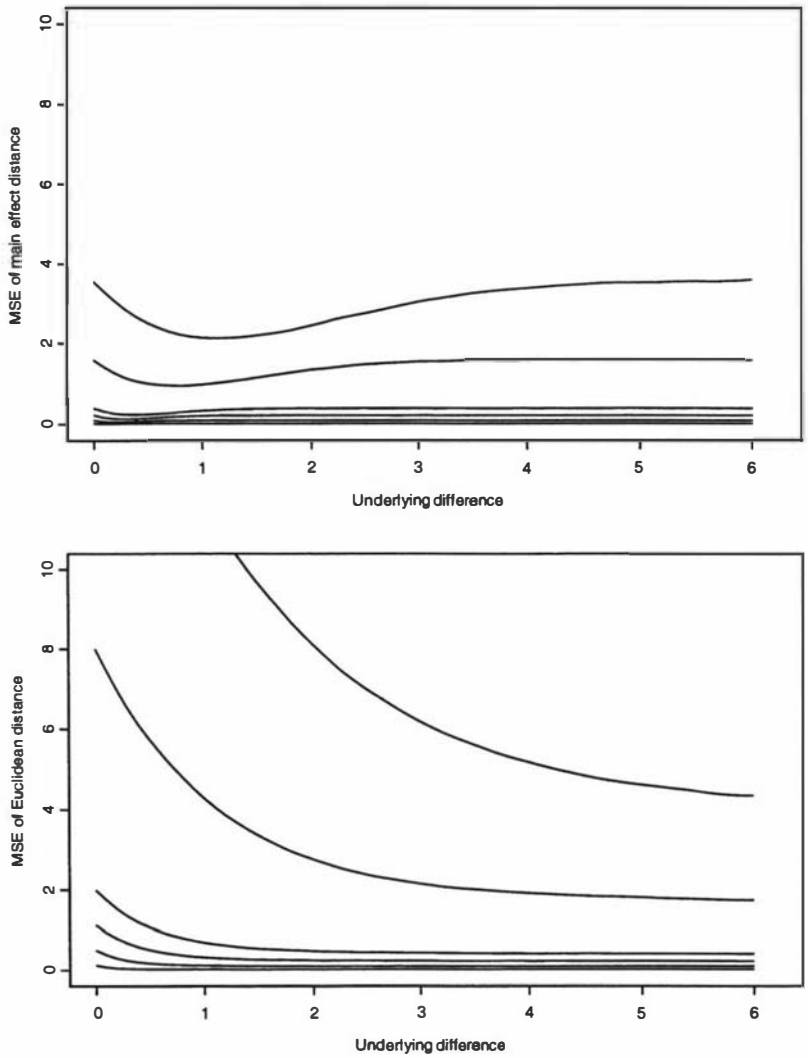


Figure 4.5: The mean squared error (MSE) of main effect (upper) and Euclidean (lower) distances plotted against underlying differences between a pair of observations. Six different levels of error variance (3, 2, 1, 0.75, 0.5, and 0.25) in the observations were used giving the six lines from top to bottom respectively in each panel.

variance of observations the MSE is relatively constant. On the other hand, the Euclidean distance has much greater bias across the entire range of underlying differences presented in Figure 4.5. Although this bias is decreasing over the range of underlying differences, the amount of bias is dependent on the error variance from the two-way model. As similar graphics would have been created by altering the number of environments and holding the error variance constant, the Euclidean distance has differing bias as the number of environments changes. It will therefore be less valid than main effect distance when working with incomplete data.

The expected value formula for squared main effect distance given in (4.20) includes a relatively small constant term based on the error variance of observations and the number of observations over which the distance was measured. It would seem reasonable to suggest that the bias in the main effect distance measure is only important when the underlying difference is less than one third of the error variance of observations. This judgement

is based on using five observations for estimation of main effect distance. The point at which the bias becomes ignorable is a function of both the number, and error variance, of observations, and if greater accuracy is required for main effect distance this quantity could be estimated.

Criteria for a good distance measure were introduced on page 99. It is clear from Figure 4.4 that the main effect distance meets these criteria better than Euclidean distance, and will therefore better distinguish pairs of genotypes with similar mean performance from those with dissimilar mean performance. Main effect distance has been shown in this section to be more accurate and more precise than Euclidean distance in scenarios where there is no significant interaction between observations (genotypes) and variables (environments).

4.9 Summary

The clustering methodology to be presented in the next chapter is dependent on distance measures that satisfy two conditions:

1. The measures need to be capable of handling incomplete data.

2. The distance measures need to reflect different aspects of genotype performance across environments.

The main effect and interaction distance measures presented in (4.6) and (4.15) respectively, satisfy both of these conditions. Both distance measures use appropriate averaging to ensure that distances measured over varying numbers of environments have the same expected value.

Interaction distance for a pair of genotypes is found using genotype means calculated using only the environments in which both genotypes were grown. Main effect distance is appropriate for use when there is no significant G×E interaction between the genotypes being compared.

Euclidean distance was shown able to be partitioned into the main effect and interaction distances in Section 4.3, even when the G×E matrix is incomplete. In Section 4.5 the three distances were shown to be non-metric when data were incomplete, although they are metric when working with complete data.

Because the data arising from the Onion Trials Programme was so sparse, distances between some pairs of genotypes were unobservable. A strategy of Ouyang *et al.* (1995) was enhanced in Section 4.6. Instead of using the maximum observed distance of the entire data set, the new strategy uses a shortest path approach to estimate an upper bound for the unobserved distance. Where possible, this estimate was combined with observed distances that were measured over few environments.

A survey of existing symmetric distance measures for continuous data was provided in Section 4.7 as further background material, while Section 4.8 showed that main effect distance was a better distance measure than Euclidean distance when there is no significant G×E interaction in the data.

Chapter 5

Two-stage clustering

5.1 Introduction

This chapter describes a clustering method that was developed to handle the incomplete G×E data which arose from the Onion Trials Programme. The new method uses the distance measures developed in Chapter 4 which discriminate between genotype main effect and G×E interaction.

The main purpose of cluster analysis is to group similar observations together while leaving dissimilar observations separate. Cluster analysis is driven by the data rather than by selection of a particular model. Two components combine to create this data description tool; a distance measure and the method chosen to form new clusters at each step of the process (discussed later in this section).

Exploring the long history of cluster analysis and G×E data reveals its use in many varying ways (Abou-El-Fittouh *et al.*, 1969; Mungomery *et al.*, 1974; Lin and Thompson, 1975; Byth *et al.*, 1976; Shorter *et al.*, 1977; Seif *et al.*, 1979; Lin, 1982; Lefkovitch, 1985; Lin and Binns, 1985; DeLacy and Lawrence, 1988; Lin and Butler, 1990; Ivory *et al.*, 1991; Lin and Morrison, 1992; Baril *et al.*, 1994; Cooper and DeLacy, 1994; Ouyang *et al.*, 1995). Many of these references were reviewed in Section 2.6. They and others, referred to later in this chapter, led to development of two-stage clustering. The remainder of this section deals with background material that puts two-stage clustering into context.

Simplification of G×E analyses

The presence of G×E interaction makes comparisons among genotypes a difficult task. Lin (1982), in a complete data set, removed the differences in the levels of genotypes by centering the rows of the G×E matrix, so that clusters of genotypes that performed similarly across environments could be found. Mean performances were then compared to establish which genotypes in each cluster performed best. This approach enables a researcher to reduce the number of genotypes that need to be compared, in future testing, to the number of clusters found. On the other hand, Ivory *et al.* (1991) used column

(environment) centred yields to cluster environments, and then compared genotypes by their mean performance in each of these clusters of 'similar' environments.

In both cases, the aim was to find a set of genotypes which performed similarly in a set of environments, that is, to find subsets of the original data matrix where there is no significant G×E interaction.

Ivory *et al.* (1991) used a G×E matrix that was rectangular in shape, having 6 rows (genotypes) and 19 columns (environments). The choice for which dimension to reduce to make comparisons was simple. The effect of clustering environments was to reduce the analysis to a more manageable 6×7 G×E matrix. Byth *et al.* (1976) clustered genotypes and environments independently to reduce their 49×63 G×E matrix to one with ten genotype groups and ten environment groups. Each of these approaches makes analysing G×E data a simpler exercise; none, however, are immediately capable of handling data sets which have missing entries in the G×E matrix.

Corsten and Denis (1990) developed a method for simultaneously clustering rows and columns in a G×E matrix, which was subsequently implemented by Baril *et al.* (1994) on a larger series of G×E matrices. This method chooses whether to form a new cluster of either genotypes or environments at each step of the process, based on the current clustering of both genotypes and environments. Byth *et al.* (1976), Corsten and Denis (1990), and Baril *et al.* (1994) have the same aim, but the mechanisms they proposed differ, and therefore would provide different results.

Cluster formation

Clusters of homogeneous observations are created in a number of ways. By far the most common method of forming clusters in G×E analyses is 'hierarchical agglomerative' clustering, which starts with single observations and forms clusters until all observations are combined into one cluster. The selection of which clusters to merge at a given step is dependent on the distance measure employed at the outset, and the criterion used to gauge the value of forming a new cluster. Known as the 'linkage method', this criterion describes how to measure the distance between a single observation and a previously formed cluster, or between two previously formed clusters. The dendrogram created is therefore dependent on the options chosen for distance measure and linkage method. Many references exist which provide greater detail on linkage methods, and little will be gained by presenting an in-depth summary of these; see for example, Chatfield and Collins (1980), Everitt (1993), and Manly (1994).

The opposite strategy to hierachical agglomerative clustering starts with all observations in a single cluster; clusters are then split to create the greatest distinction between the two new clusters formed at each stage. 'Divisive' clustering methods are also hierarchical, and dependent on the distance measure and division criterion chosen.

Non-hierarchical clustering methods include the mixture model approach used by

Ganesalingam and McLachlan (1978), followed by McLachlan and Basford (1988). In general, this approach determines the likelihood that each observation belongs to each of a pre-determined number of clusters. The clusters themselves are in turn determined by their members.

A mixture model clustering is usually determined through use of the EM algorithm because there are two distinct sets of parameters that must be used to optimize the likelihood of the mixture model. The following steps describe this implementation of the EM algorithm:

1. Determine the number of clusters, and their initial 'locations'.
2. Determine the probability that each observation belongs to each cluster. This is the 'E' step in the EM algorithm.
3. Re-evaluate the location of each cluster, given the current membership of clusters. This is the 'M' step in the EM algorithm.
4. Iterate through the 'E' and 'M' steps until some convergence criterion is satisfied.

This iterative approach can be quite slow, and there is a need to pre-determine the number of groups that exist in the data. McLachlan and Basford (1988) provide much greater detail with regard to the implementation of mixture model clustering.

The outcome of mixture model clustering is a set of probabilities that observations belong to each cluster. If all of these probabilities are either zero or one, the clustering is said to be 'crisp' or 'hard'. In general, however, the outcome of mixture model clustering will give what has been called a 'fuzzy clustering', as the probabilities lie somewhere in the range zero to one. Mixture model clustering is a powerful tool but the added complexity that would be introduced through its use is, for the time being, considered unnecessary in this work.

Cluster analyses with missing data

There is a dearth of papers dealing with clustering of incomplete two-way data in general, let alone in G×E data. Ouyang *et al.* (1995) clustered ninety environments using incomplete genotype data for each year from 1985 to 1990. The distance measure applied was capable of handling incomplete data. When locations were clustered using data from all years, these authors used a method described in DeLacy *et al.* (1996), which averages the within-year distance matrices to give a single set of inter-location distances. This method for combining cluster analyses weights entries from within-year inter-location distance matrices by the number of genotypes used to form each inter-location distance. Baril *et al.* (1994) could have used this strategy to combine all of their data, but presented eight yearly cluster analyses instead.

The use of distance measures that take account of the incomplete nature of data

being clustered is therefore not new to G×E researchers. Distance measures that reflect differences in main effect and interaction, developed in Chapter 4, are now incorporated into a method for clustering a large number of genotypes. The next section describes the two-stage clustering strategy; subsequent sections show how this method was applied to a set of G×E data from the literature and then the data sets arising from the Onion Trials Programme. The final section of this chapter presents a model for yield which utilizes the results from two-stage clustering.

5.2 Description of two-stage clustering

G×E researchers are often interested in the specific adaptability of genotypes to environments. The Onion Trials Programme is a particular instance where this is true. Once groups of genotypes that have similar specific adaptation potential are determined, it would also be useful to know which of these varieties perform similarly across all environments.

As identified in the previous section, currently clustering methodology does not appear to appropriately handle incomplete data. A notable exception is the method of Ouyang *et al.* (1995) which confounds main effects with G×E interaction. The two-stage clustering strategy, presented in this section, differentiates between G×E interaction and mean performance in spite of the data's incompleteness. The development of distance measures capable of handling incomplete G×E data in Chapter 4 allowed the following two-stage clustering method to be proposed.

First stage

(i) Calculate all interaction distances I_{ij} using (4.15) and the G×E matrix.

(ii) Cluster genotypes using these interaction distances. This produces clusters of shape-similar genotypes.

Second stage

(i) Calculate main effect distances M_{ij} within each first stage cluster using (4.6) and the G×E matrix.

(ii) Cluster genotypes within the first stage clusters using these main effect distances. This produces final clusters of level-and-shape similar genotypes.

Given dissimilarities between pairs of genotypes (in both first and second stages) a decision must be made on how clusters are to be formed at each step of the clustering process. The incremental sum of squares method for forming clusters (Ward, 1963) was chosen for use in this research. It successively merges two current clusters so as to minimize the increase in within-clusters sum of squared distances. This method had the advantage of leading to a simple stopping criterion. Cluster formation is stopped when the next cluster to be formed would have an average within-cluster sum of squared distances that

is greater than the average sum of squared distances in the entire set of observations. This method was used by Baril *et al.* (1994) when they simultaneously clustered genotypes and environments using their G×E interaction. More complex stopping criteria have been developed and are reviewed in Everitt (1993), while Ghaderi *et al.* (1980) and Lin (1982) provide examples in the G×E literature.

Stabilization of variance across environments

Changes in variance across environments will limit the success of this approach; In agreement with the findings of Fox and Rosielle (1982) and Yau (1991), transformation within environments is advocated to equalize variance, ensuring that environments contribute similarly to distance measures. This was of particular importance in the case of missing G×E data as the genotypes needed to be given an opportunity to contribute equally to results, irrespective of the subset of environments in which they were grown. If the raw data were used, performances from environments with greater variation would have greater influence on results than those from environments with lower variation.

Heterogeneity of variance from environment to environment needed to be investigated. This could be performed using Levene's test, however, problems had been found implementing this test on unbalanced data. O'Neill and Mathews (2000) argued that weightings should be used when calculating Levene's test for homogeneity of variance when the treatments are not equally replicated. They also noted that different statistical software applications use different versions of the Levene's test, and that some are inappropriate in their current form for unequal treatment replication.

5.3 An example of two-stage clustering

In this section, two-stage clustering is illustrated by applying it to a data set, well known in the G×E literature. Mungomery *et al.* (1974) first reported the experiment from which the data originated. It has been analysed in many different and sometimes innovative ways, including Basford (1982) and McLachlan and Basford (1988). While Basford and Tukey (1998) published the data set in full, the data used in this and subsequent sections is available on the CD-ROM accompanying this volume.

The experiment measured six response variables for 58 soybean genotypes from four locations over two years. This data has been commonly analysed using the location-year combination as the environment, which gives a G×E matrix that is 58×8 in size with entries being mean yields of two replicates from a randomized block design. As variation within environments might have affected the analysis in this illustrative example, the yields were first standardized within each environment, using the transformation previously presented in (3.3)

$$z_{ik} = \frac{y_{ik} - \bar{y}_{\cdot k}}{s_k}$$

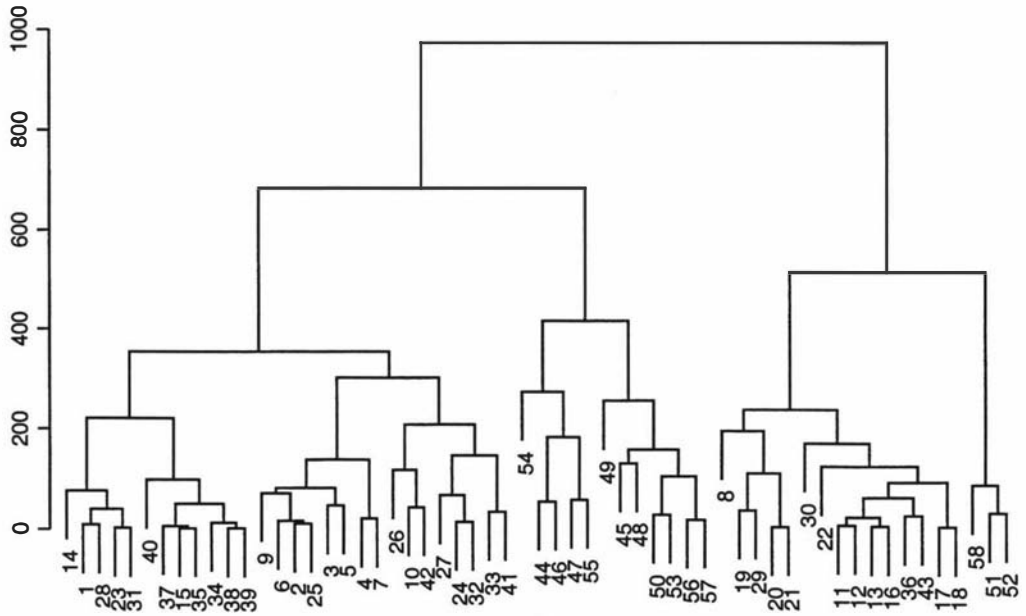


Figure 5.1: First stage clustering of 58 genotypes using complete data. Interaction distance I_{ij} and the incremental sum of squares method were applied.

where $\bar{y}_{\cdot k}$ and s_k are the kth environment mean and standard deviation, respectively. This transformation did not alter the qualitative structure of the G×E interaction that existed in the data, but ensured that each environment contributed on an equal footing to the distance measures calculated.

Before investigating the use of two-stage clustering on incomplete data, the desired outcome needed to be identified. Ideally, the clustering of genotypes using incomplete data should approximate the results that would have been gained if the complete data were available. Figure 5.1 shows first stage clustering of the complete data. The dendrogram was formed using interaction distance I_{ij} and the incremental sum of squares method for forming clusters. The stopping criterion, described in Section 5.2, determined that the dendrogram should be truncated at the level 238, and that there were nine clusters of genotypes.

Fifteen of the 464 G×E combinations were then deleted as an example of two-stage clustering on incomplete data. These G×E combinations were chosen at random, and are shown in Table 5.1. Figure 5.2 shows the first stage dendrogram for the 58 genotypes clustered using interaction distance I_{ij}, appropriate when data is incomplete, and the incremental sum of squares method.

First stage clustering was truncated at level 196 with ten clusters remaining, as determined by the stopping criterion. In contrast, the complete data had only 9 clusters when the dendrogram was truncated. Table 5.2 shows the number of genotypes in each cluster of Figure 5.1 cross-tabulated against the number of genotypes in each cluster of Figure 5.2. It is evident from this table that the number of genotypes that stay together from one clustering to another is high. In fact only two distortions took place when the

Genotype	Environment	Genotype	Environment	Genotype	Environment
2	R70	10	L71	26	B71
5	L71	14	N70	30	N70
5	N71	19	B70	37	N71
6	N71	19	L71	52	B70
7	R71	24	B70	53	R70

Table 5.1: Fifteen G×E combinations randomly deleted to create an incomplete set of data.

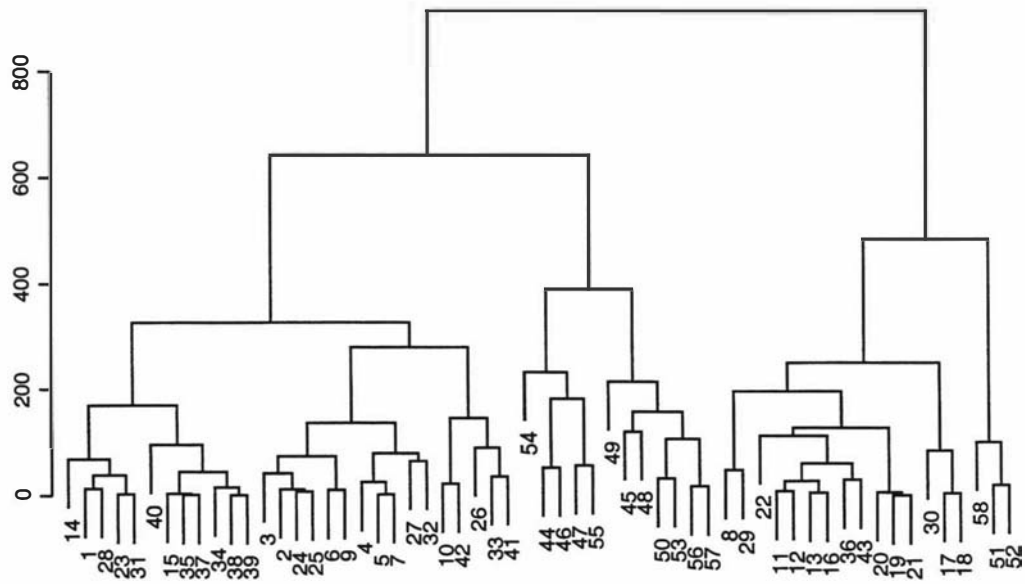


Figure 5.2: First stage clustering of 58 genotypes. Interaction distance I_{ij} and the incremental sum of squares method were applied to the incomplete data created by the deletion of fifteen values from the G×E matrix.

Cluster number	Cluster number from Figure 5.1								
from Figure 5.2	1	2	3	4	5	6	7	8	9
1	12								
2				4					
3		12							
4					6				
5			5						
6			3		8				
7							3		
8	3								
9								1	
10									1

Table 5.2: Cluster memberships of 58 genotypes when clustered using complete (columns) and incomplete data (rows) in Figures 5.1 and 5.2 respectively. The numbers indicate the cardinality of each cluster.

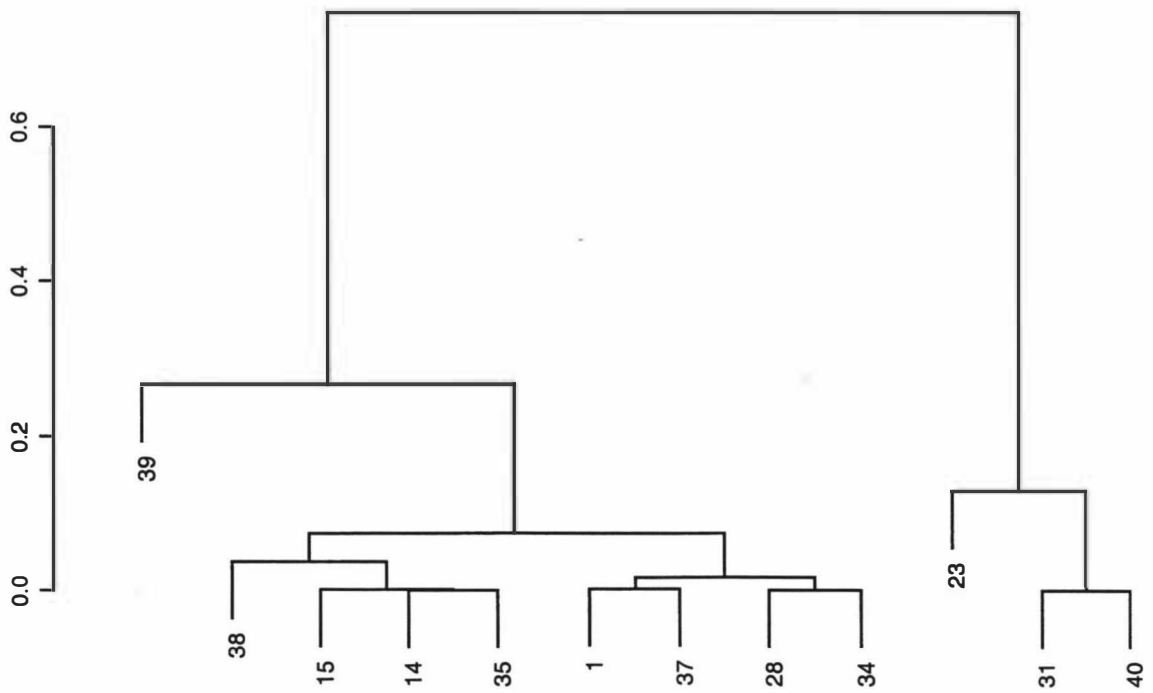


Figure 5.3: Second stage clustering of twelve similar genotypes. A first stage cluster (seen at the left of Figure 5.2), was re-clustered using main effect distance M_{ij} and the incremental sum of squares method.

incomplete data is used. First, the last cluster (column 1) formed using complete data was left as two clusters (rows 1 and 8) when the incomplete data was used. This was not caused by premature truncation of Figure 5.2, as the next step would have instead merged clusters 4 and 9.

The second distortion was the misplacement of three genotypes which should be in cluster 5, but were instead put into cluster 6. It should be remembered that a cluster analysis uses empirical yield results, which include error in their outcome. If exact results were available, the degree to which the incomplete data reflected the same information as complete data could be measured. It cannot be stated with complete surety that the incomplete data was giving an inaccurate depiction of the real situation. What can be said is that the dendrograms constructed using complete and incomplete data are very similar.

To illustrate second stage clustering, one cluster of interaction similar genotypes, seen at the left of Figure 5.2 is examined.. This is the third to last cluster formed and contains genotypes 14, 1, 28, 23, 31, 40, 15, 35, 37, 34, 38, and 39. Figure 5.3 shows the second stage dendrogram for this first stage cluster, using main effect distance M_{ij} appropriate for incomplete data, and the incremental sum of squares linkage method.

Sets of genotypes, similar in mean performance over environments (from second stage clustering), and that are specifically adapted to the same kinds of environments because they have similar G×E interaction profiles (as determined by first stage clustering) were

then determined. For example, genotypes 14, 15, and 35 are now deemed by two-stage clustering to have similar performances across environments in terms of both G×E interaction (shape), and the average yield (level). Having demonstrated two-stage clustering on the Mungomery *et al.* (1974) data, the next section describes its application to the data arising from the Onion Trials Programme.

5.4 Application of two-stage clustering to the trials programme data

Before applying two-stage clustering to Onion Data I and II, the nature of the data within each set needed to be investigated. In particular, it was necessary to deal with any heterogeneity in the data, across either genotypes or environments.

Clustering of genotypes should depend on results within environments, rather than on where each genotype was grown. The decision was therefore taken to standardize within environments, to ensure that each environment contributed equally to the outcome of clustering. Figures 5.4 and 5.5 show the effects of this transformation on the raw data of Onion Data I and II; Figures 5.6 and 5.7 repeat this exercise for the square roots of yields from Onion Data I and II respectively. Normality plots in these figures showed no advantage in using either raw or square roots of yields, but the histograms highlighted some differences between the untransformed and transformed data. Some negative skewness in the data was created by within-environment standardization of the square roots of yields. In all four of the normal probability plots presented, there are some points at the bottom left that do not fit the trend. This was caused by the data created to reflect the failure of crops, as discussed in Sections 3.3 and 3.6.

The scatter plots show that there was no advantage in using within-environment standardization of the square roots of yields, instead of the raw yields, to gain a more consistent range of genotype standard deviations. Correlations between standard deviations and means of genotypes were investigated and showed that the downward trend apparent for the within-environment standardized square roots of yield were significant. Correlations between standard deviations and means of genotypes based on within-environment standardized yields were also significant, and it was determined that in keeping with the initial data analysis of Section 3.4, the square roots of yields should be used in favour of the raw yield data in this analysis.

Figure 5.8 shows the first stage clustering of Onion Data I genotypes using interaction distance I_{ij} and the incremental sum of square method of forming clusters as described in Section 5.2. The stopping criterion used throughout this chapter determined that there are 24 first stage clusters, whose members are listed in Table 5.3. Using the same methodology for the first stage clustering of Onion Data II, resulted in 22 clusters of interaction similar genotypes. Figure 5.9 shows the dendrogram for this clustering, while the memberships

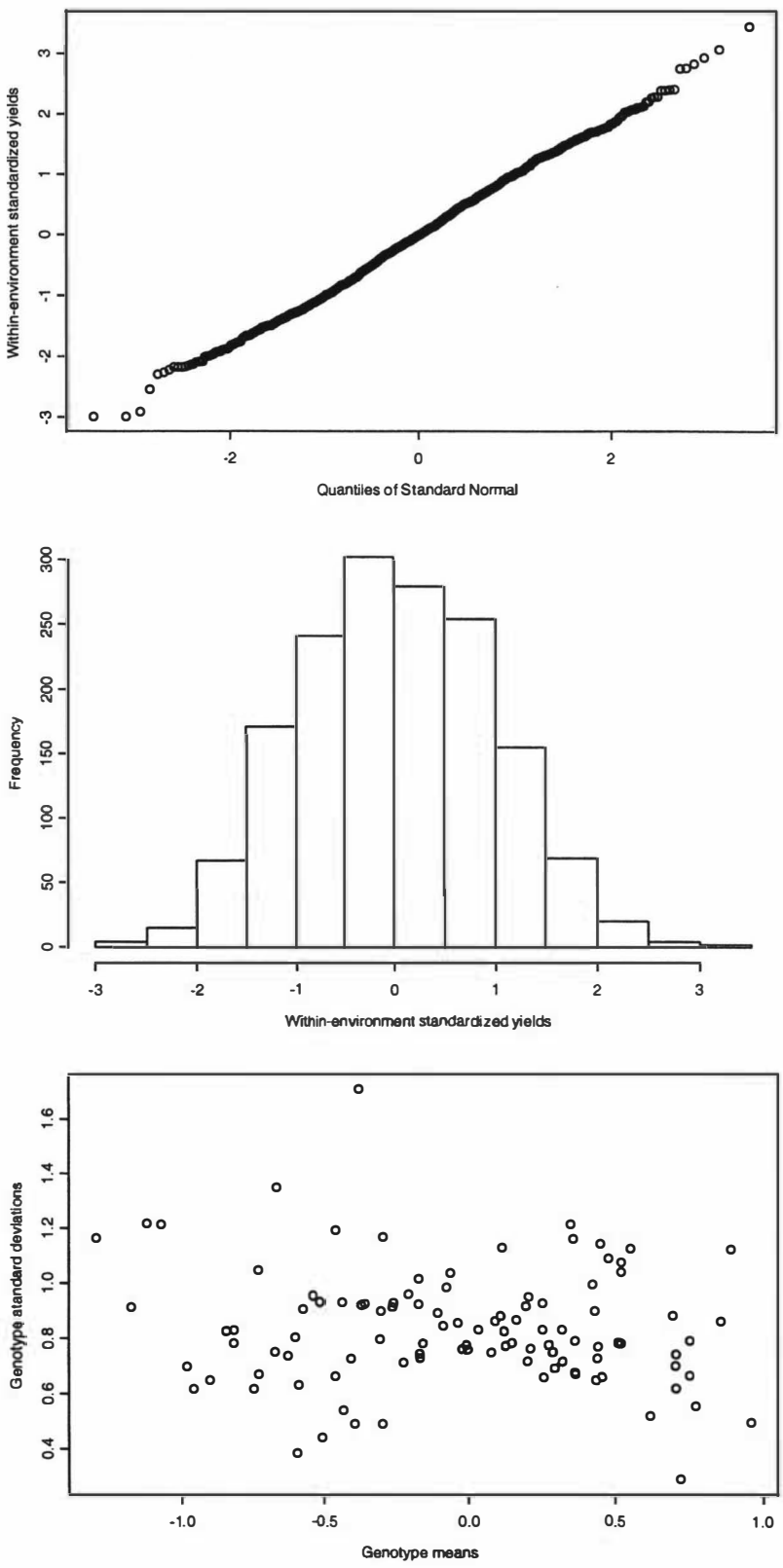


Figure 5.4: Effects of using within-environment standardization on Onion Data I. The normal probability plot (top), and histogram (centre) show the distribution of the transformed data, while the scatter plot (bottom) shows the effect of this transformation on the relationship between genotype means and standard deviations. Within-environment standardized yields were originally measured in kilograms per square metre.

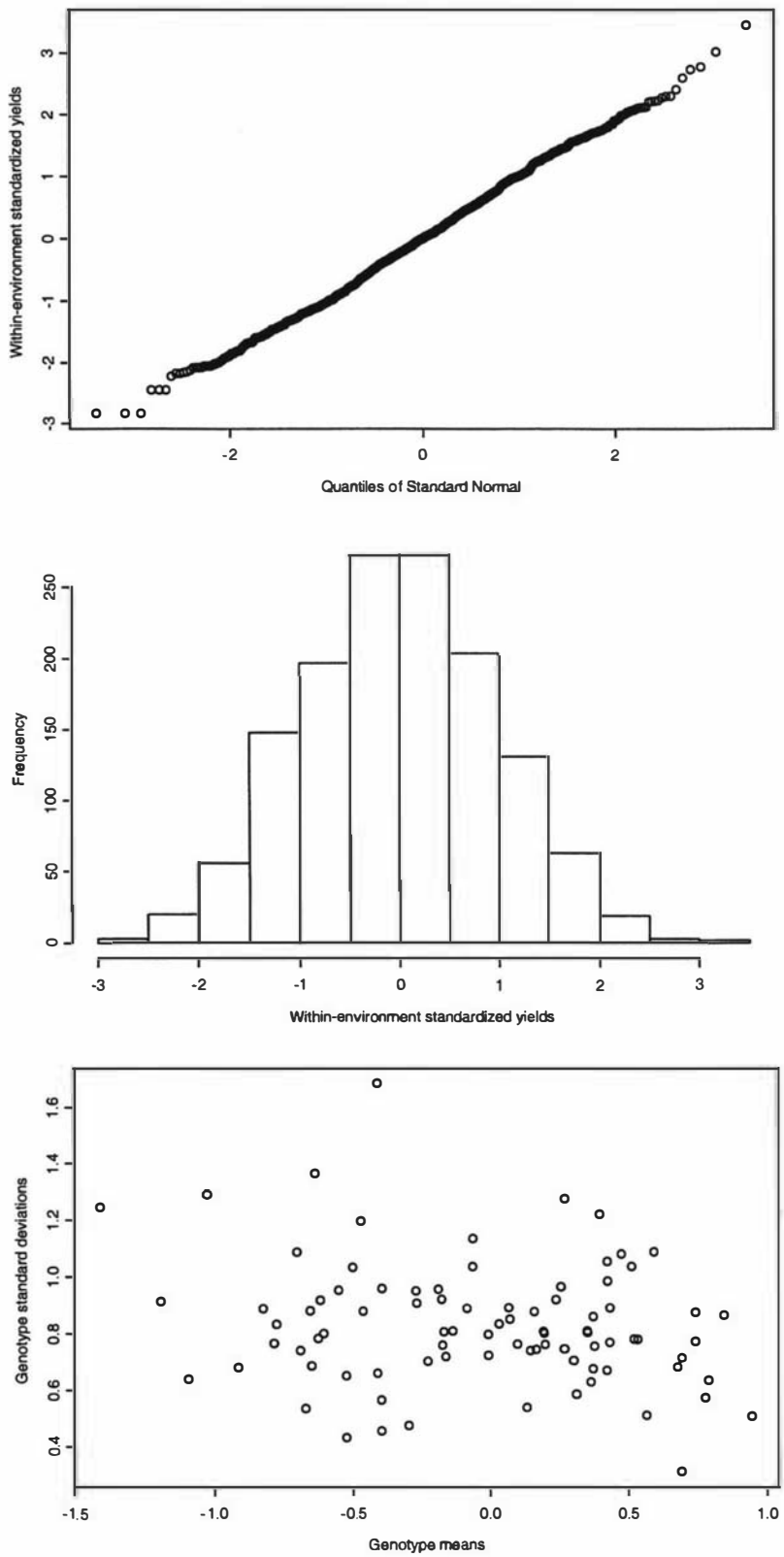


Figure 5.5: Effects of using within-environment standardization on Onion Data II. The normal probability plot (top), and histogram (centre) show the distribution of the transformed data, while the scatter plot (bottom) shows the effect of this transformation on the relationship between genotype means and standard deviations. Within-environment standardized yields were originally measured in kilograms per square metre.

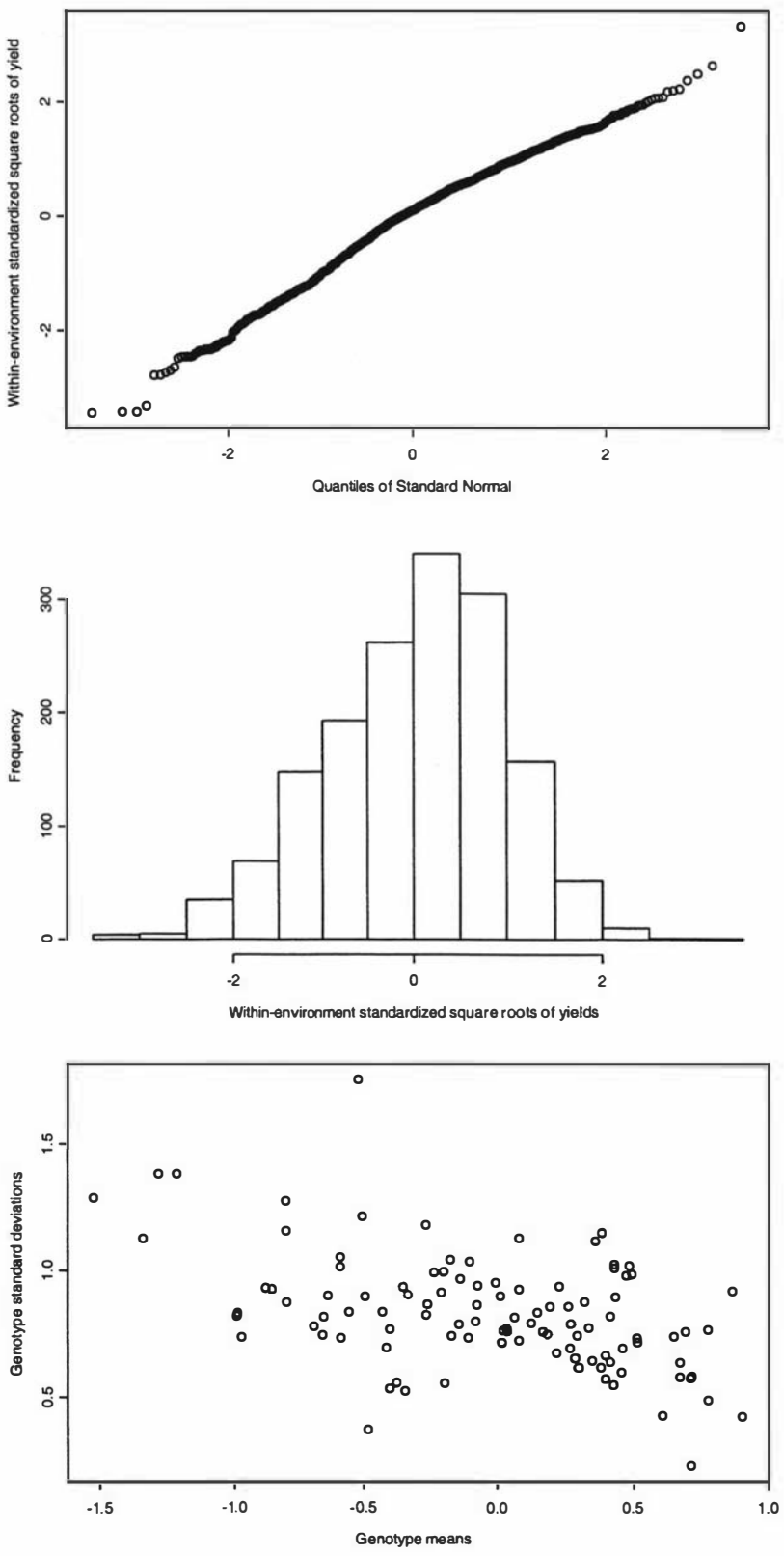


Figure 5.6: Effects of using within-environment standardization on the square roots of yields from Onion Data I. The normal probability plot (top), and histogram (centre) show the distribution of the transformed data, while the scatter plot (bottom) shows the effect of this transformation on the relationship between genotype means and standard deviations.

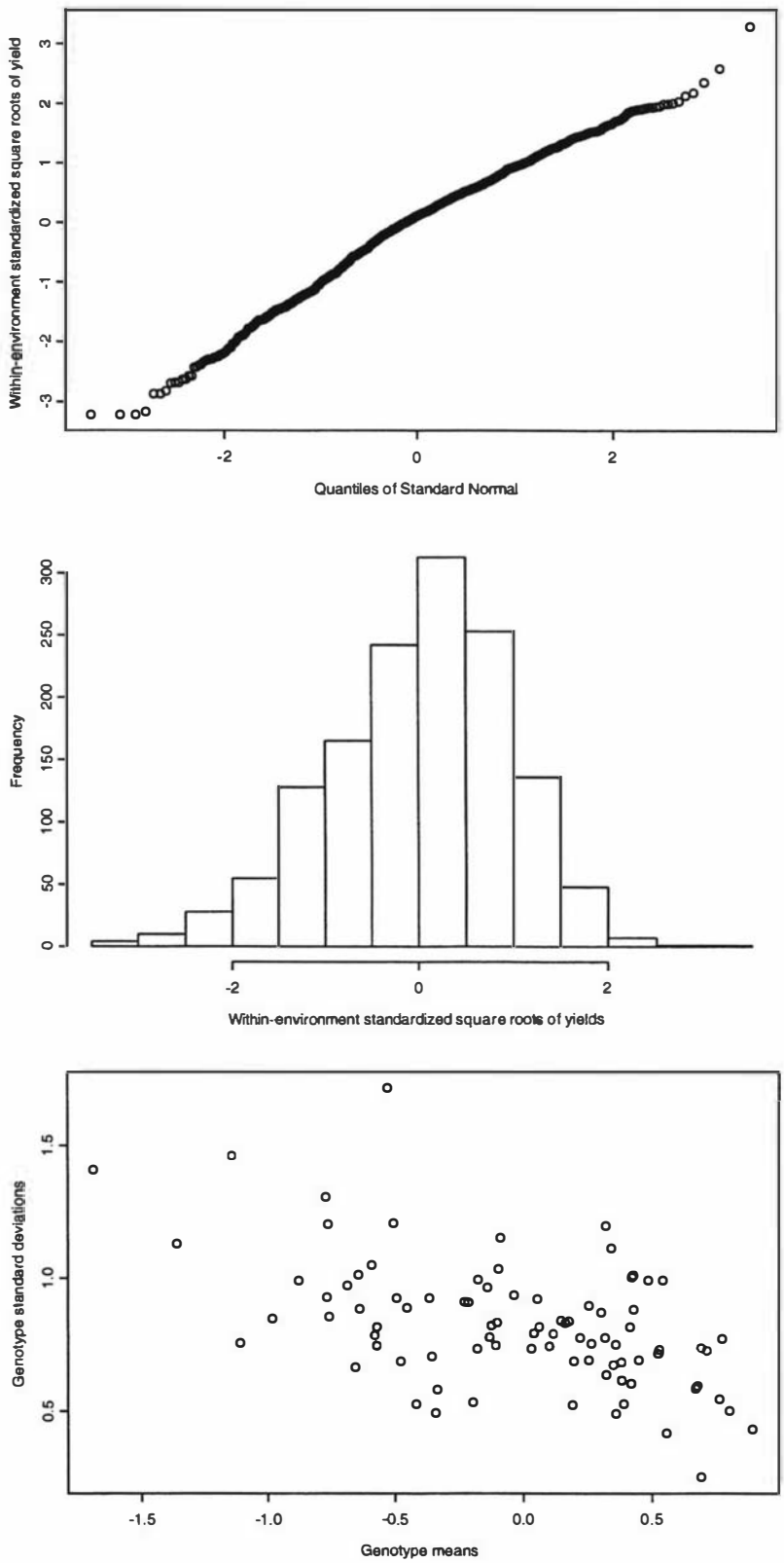


Figure 5.7: Effects of using within-environment standardization on the square roots of yields from Onion Data II. The normal probability plot (top), and histogram (centre) show the distribution of the transformed data, while the scatter plot (bottom) shows the effect of this transformation on the relationship between genotype means and standard deviations.

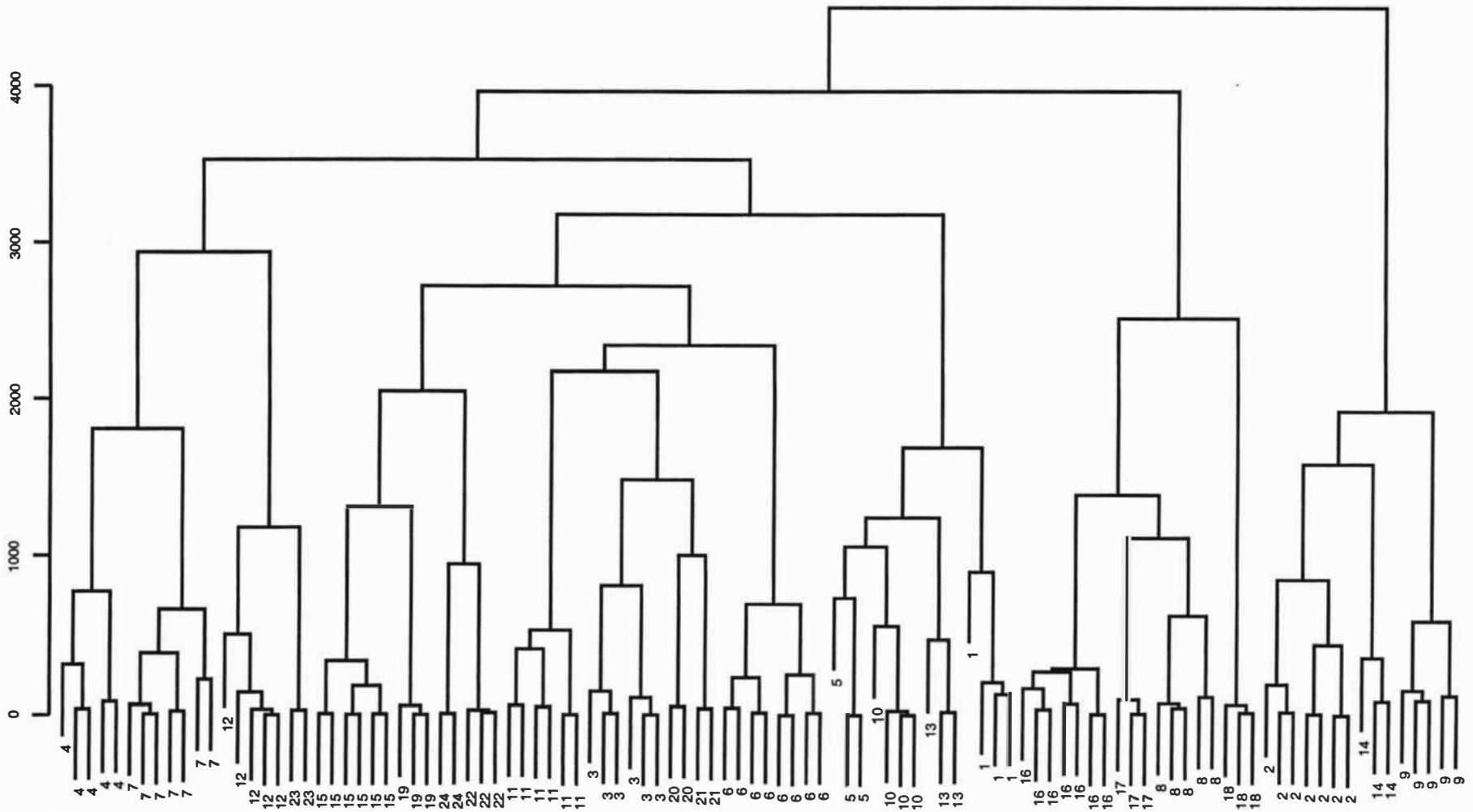


Figure 5.8: First stage clustering of 104 genotypes from Onion Data I. The incremental sum of squares method and interaction distance I_{ij} were applied. The stopping criterion truncated the process at the level 898, forming 24 first stage genotype clusters. Labels refer to the cluster, which can be linked to particular genotypes using Table 5.3.

Cluster	Members
1	Cadix ZU, Eytan HZ, Jenin HZ, Marix ZU.
2	HA-891 HZ, Hurricane RS, Jaguar PS, Mercedes PS, PS 8392 PS, Regia AS, Superex TK.
3	Ben Shemen HZ, Bon Accord HT, Galil HZ, NuMex BR-1 RC, Pyramid SA, Tropic Gold NW.
4	Agrifound Dark Red AF, Agrifound Light Red AF, Australian Brown ST, Red Creole LO, Rojo SS.
5	Brownsville AS, Gladalan White YA, Savages Flat White YA.
6	Colossal PVP SS, Dessex SS, Equanex PS, Granex Yellow TK, Ringer Grano SS, Rio Bravo RC, Rio Raji Red RC, Texas Grano 502 PRR AS.
7	Creole Red PRR PS, Dehydrator No 3 SS, Extra Early Creamgold NW, HA-222 HZ, HA-226 HZ, IRAT-69 MA, Rouge de Tana TS.
8	Red Creole Credo RS, Rio Hondo RC, Tropic Ace TK, Tropicana RS, Yellow Creole SS.
9	HA-950 HZ, Linda Vista PS, Nikita RC, RS 209 RS, Violet de Galmi TS.
10	Creamgold YA, Early Lockyer Brown YA, Supply YA, Yodalef HZ.
11	Belem IPA-9 IP, Composto IPA-6 IP, Granex 429 AS, Redbone AS, Savannah Sweet PS, Serrana AS.
12	Agrifound Rose AF, Red Burgundy Imp NE, Red Creole PRR PVP SS, Red Creole SA, Red Creole Select NE.
13	Kano Red NI, Lockyer Gold NW, Red Synthetic HZ.
14	Pera IPA-4 IP, Riviera AS, Utopia AS.
15	Arad HZ, El Ad HZ, Gladalan Brown YA, HA-489 HZ, Marathon HZ, RAM 710 HZ.
16	Deko HZ, Early Red HZ, Niv HZ, Regal PVP SS, Rio Blanco Grande RC, Rio Ringo RC, Texas Grano LO.
17	Granoble PS, Pusa Red AF, Ringo RS.
18	N-53 LO, Nasik Red LO, Red Bombay RS.
19	Granex 33 AS, Mr Max RC, Sivan HZ.
20	Houston AS, Texas Grano 438 AS.
21	Primero SS, Red Creole AS.
22	HA-675 HZ, Ori HZ, Red Star PS.
23	Red Comet PS, Yellow Granex Imp PRR SS.
24	HA-230 HZ, HA-817 HZ.

Table 5.3: The members of the 24 clusters from the first stage clustering of Onion Data I, given in Figure 5.8.

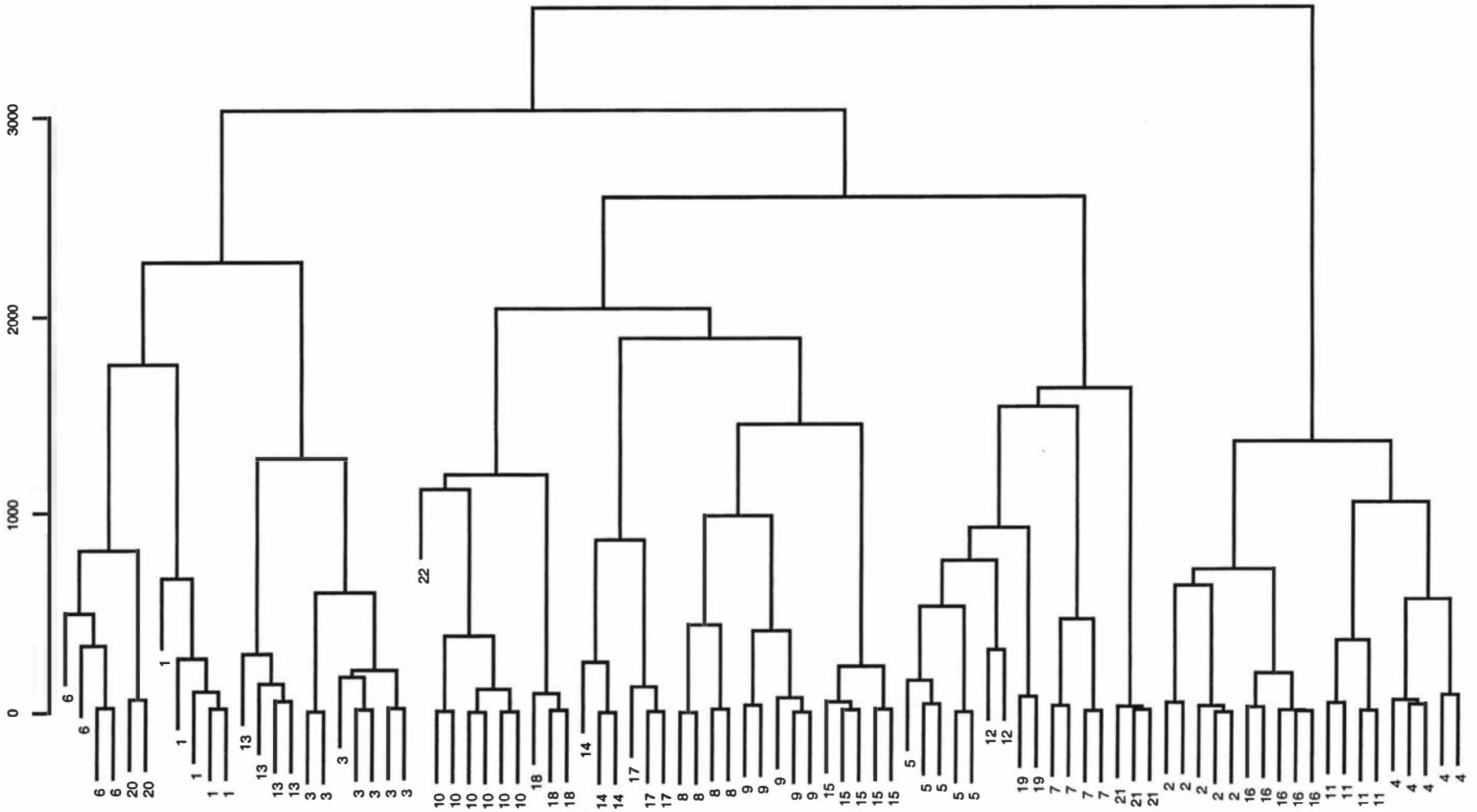


Figure 5.9: First stage clustering of 87 genotypes from Onion Data II. The incremental sum of squares method and interaction distance I_{ij} were applied. The stopping criterion truncated the process at the level 667, forming 22 first stage genotype clusters. Labels refer to the cluster, which can be linked to particular genotypes using Table 5.4.

Cluster	Members
1	Red Burgundy Imp NE, Red Comet PS, Red Creole PRR PVP SS, Red Creole SA, Yellow Granex Imp PRR SS.
2	Deko HZ, Early Lockyer Brown YA, Red Bombay RS, Supply YA, Yodalef HZ.
3	Creole Red PRR PS, Dehydrator No 3 SS, HA-222 HZ, HA-226 HZ, IRAT-69 MA, Redbone AS, Rouge de Tana TS.
4	Red Creole Credo RS, Rio Hondo RC, Tropic Ace TK, Tropicana RS, Yellow Creole SS.
5	Granex 429 AS, HA-950 HZ, Mercedes PS, RS 209 RS, Regia AS.
6	Agrifound Dark Red AF, Agrifound Rose AF, Australian Brown ST, Rojo SS.
7	HA-817 HZ, Hurricane RS, Serrana AS, Superex TK.
8	Brownsville AS, Gladalan White YA, Lockyer Gold NW, Red Synthetic HZ.
9	Colossal PVP SS, Galil HZ, Pyramid SA, Ringer Grano SS, Tropic Gold NW.
10	Arad HZ, El Ad HZ, Gladalan Brown YA, HA-489 HZ, Marathon HZ, RAM 710 HZ.
11	Granoble PS, Pusa Red AF, Regal PVP SS, Ringo RS.
12	Pera IPA-4 IP, Violet de Galmi TS.
13	Creamgold YA, Extra Early Creamgold NW, Jenin HZ, Red Creole AS.
14	Belem IPA-9 IP, Composto IPA-6 IP, Savannah Sweet PS.
15	Dessex SS, Equanex PS, Rio Bravo RC, Rio Ringo RC, Texas Grano 502 PRR AS.
16	Early Red HZ, Niv HZ, Rio Blanco Grande RC, Rio Raji Red RC, Texas Grano LO.
17	Granex 33 AS, Houston AS, Sivan HZ.
18	Ben Shemen HZ, Bon Accord HT, NuMex BR-1 RC.
19	Linda Vista PS, Utopia AS.
20	Agrifound Light Red AF, Red Creole LO.
21	HA-675 HZ, Ori HZ, Red Star PS.
22	Marix ZU.

Table 5.4: The members of the 24 clusters from the first stage clustering of Onion Data II, given in Figure 5.9.

are listed in Table 5.4.

The truncation levels for these dendrograms are 898 and 667, which is based on the average distance between observations in the data being clustered. Onion Data I genotypes therefore covered a broader range of G×E interaction patterns than did the genotypes in Onion Data II. If clustering of Onion Data II genotypes had been truncated at the level 898, there would have been only 18 clusters of genotypes; or if clustering Onion Data I genotypes had been truncated at the level 667, there would have been 30 clusters. This indicated some sensitivity to the level chosen to truncate clustering. Choosing the number of genotype clusters to be the same for both cluster analyses (23) would not result in perfect matching of cluster memberships from Table 5.3 to Table 5.4. This would indicate that the seventeen additional genotypes in Onion Data I had an impact on the way that the 87 Onion Data II genotypes clustered together. An automated (formula driven) stopping rule was preferred to avoid the need for subjective evaluation of large and complex dendrograms. There was no reasonable way of linking the number of genotype clusters for Onion Data I and II in conjunction. The number of clusters was determined

by the data.

Second stage clustering was applied and the outcome presented for one first stage cluster from each of Onion Data I and II. The first stage cluster with the most genotypes in each case was chosen to show the alteration between first and second stage clustering. Figure 5.10 presents dendrograms for first and second stage clustering of first stage cluster 6 from Figure 5.8 which has eight members. Corresponding dendrograms for first stage cluster 3 from Figure 5.9 which has seven members are shown in Figure 5.11. In both of these figures, the upper dendrogram re-presents the clustering of the first stage cluster using interaction distance. It is therefore a portion of the full first stage dendrogram presented previously. There was very little difference between the members of these clusters in terms of G×E interaction pattern compared to the other genotypes in the data. The lower dendrograms present second stage clustering and used main effect distance in their construction, appropriate because there was little or no G×E interaction to consider when comparing these genotypes. Comparing the genotype means presented in the scatter plots of Figures 5.6 and 5.7, to the scale on the second stage dendrograms, indicated substantial differences between genotypes based on mean performances across environments. These results indicated that while these genotypes are suited to the same kinds of environments, there are potential gains to be made in terms of yield by selecting the varieties with the higher mean performance. It is also evident that there was very little difference between such pairs of genotypes as 'Colossal PVP SS' and 'Equanex PS' in Figure 5.10, or between 'Dehydrator No. 3 SS' and 'HA-222 HZ' in Figure 5.11.

5.5 Expressing clustering in a parametric model framework

In this section, various clustering strategies are linked to linear models by providing functions that explain a genotype performance y_{ik} using variables to represent genotype or environment cluster memberships. Fitting these models will allow parameter estimation and could be used to gauge the significance of the models. Once clustering has been performed, indicator variables showing membership of clusters can be created. A linear model could then be fitted to the yield data using these factors as explanatory variables. The model thus created would clearly be driven by the data itself.

Recall the standard model for yield in situations where there is no replication within G×E combinations,

$$Y_{ik} = \mu + G_i + E_k + (GE_{ik} + \epsilon_{ik})$$

from (4.16).

Let the subscript notation $g(i)$ indicate that genotype i is a member of genotype cluster g, and similarly let $h(k)$ indicate that environment k is a member of environment cluster h. The clustering given by various authors can now be described parametrically.

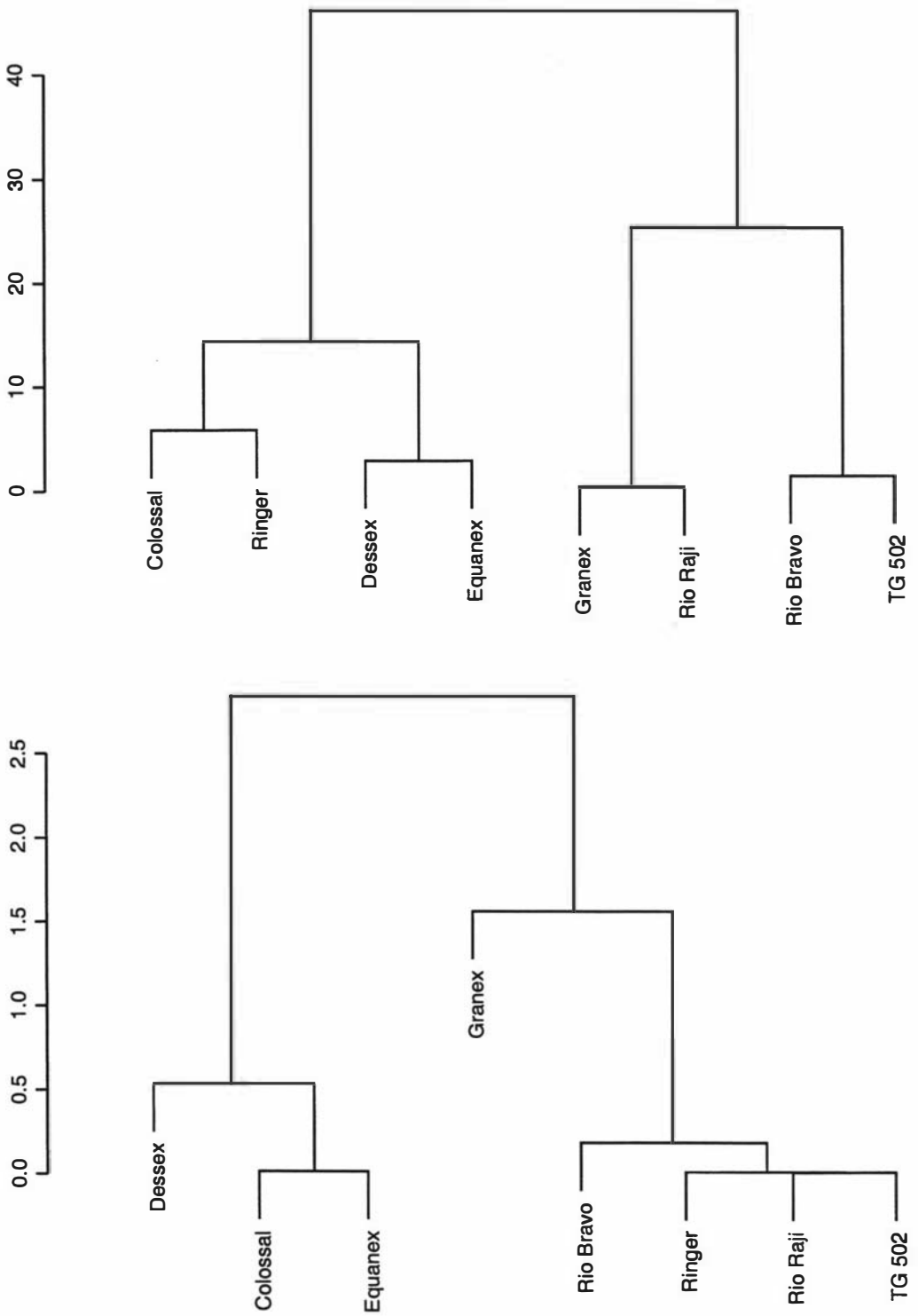


Figure 5.10: Two-stage clustering of first stage cluster 6 from Figure 5.8. The first stage clustering (top) of Onion Data I genotypes has been re-presented to allow comparison with its second stage clustering (bottom). Genotype names have been abbreviated for clarity.

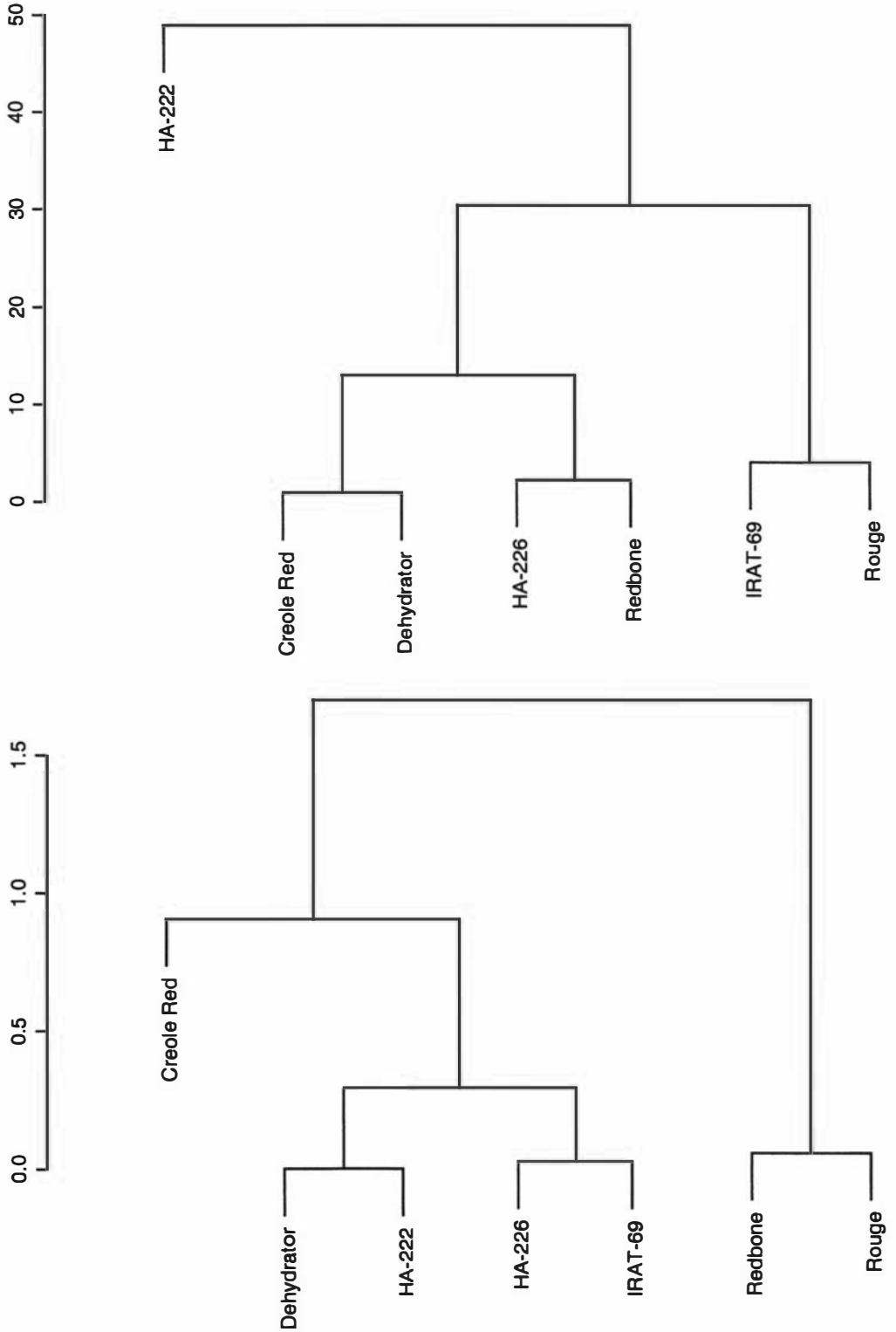


Figure 5.11: Two-stage clustering of first stage cluster 3 from Figure 5.9. The first stage clustering (top) of Onion Data II genotypes has been re-presented to allow comparison with its second stage clustering (bottom). Genotype names have been abbreviated for clarity.

For example, the clustering of genotypes using Euclidean distance could be expressed as

$$Y_{ik} = \mu + \left(G_{g(i)} + GE_{g(i)k}\right) + G'_i + E_k + \left(GE'_{ik} + \epsilon_{ik}\right) \tag{5.1}$$

where the bracketing of the terms $(G_{g(i)} + GE_{g(i)k})$ represents the merging of the main effect and interaction used to form cluster g and would be fitted in a linear model using an explanatory variable for $G_{g(i)}$, along with the interaction of this term and the variable for E_k. The terms G'_i and GE'_{ik} reflect differences between the mean of genotypes in cluster g and the individual performance of genotype i in environment k. It should be noted that the G×E interaction not accounted for by the model is still confounded with the error term, and therefore remains bracketed. The term G'_i is nested within the term $G_{g(i)}$ for the cluster g to which genotype i belongs; it could therefore be fitted in a linear model accordingly. Use of a 'prime' in the models of this section indicates that the term is actually part of the unexplained variation and as such would generally not be used when fitting the linear model. These unexplained sources of variation would form the residual term of the model, when added to ϵ_{ik}.

More complicated clustering strategies lead to more complicated models, but are no more difficult to implement in a linear model framework. When genotypes are clustered together based on the similarity of G×E interaction, as in Lin (1982), the model can be expressed as

$$Y_{ik} - \bar{Y}_{i\cdot} = E_k + GE_{g(i)k} + \left(GE'_{ik} + \epsilon_{ik}\right) \tag{5.2}$$

Note that Lin (1982) centred the rows of the G×E matrix which is reflected by the alteration to the left-hand-side of the equation. The column centring used by Ivory *et al.* (1991) can be expressed as

$$Y_{ik} - \bar{Y}_{\cdot k} = G_i + GE_{ih(k)} + \left(GE'_{ik} + \epsilon_{ik}\right) \tag{5.3}$$

Byth *et al.* (1976) presented a model for two-way clustering, which using the current notation would appear as

$$\begin{aligned} Y_{ik} \;=\; & \mu + \left(G_{g(i)} + E_{h(k)} + GE_{g(i)h(k)}\right) + G'_i + E'_k \\ & + \left(GE'_{g(i)k} + GE'_{ih(k)} + GE'_{ik} + \epsilon_{ik}\right) \end{aligned} \tag{5.4}$$

The bracketing of terms representing genotype and environment main effects and the G×E interaction reflects the fact that genotype (environment) clusters were formed using a distance measure that confounded the genotype (environment) main effect with the G×E interaction.

Alteration of the bracketing used in this model highlights the difference between the Byth *et al.* (1976) and Corsten and Denis (1990) clusterings. Corsten and Denis (1990) clustered genotypes and environments simultaneously on the basis of similar G×E inter-

action. Their model could be expressed as

$$\begin{aligned} Y_{ik} &= \mu + G_i + E_k + GE_{g(i)h(k)} \\ &\quad + \left(GE'_{g(i)k} + GE'_{ih(k)} + GE'_{ik} + \epsilon_{ik}\right) \end{aligned} \tag{5.5}$$

For two-way clustering to be represented in a linear model framework, a single indicator variable would need to be created to reflect the term $GE_{g(i)h(k)}$ in (5.5), but would also incorporate the genotype and environmental main effects bracketed in (5.4).

Two-stage clustering, as proposed in this investigation, shows greater definition, as shown by the model

$$Y_{ik} = \mu + GE_{f(i)k} + G_{g(i)} + G'_i + E_k + \left(GE'_{ik} + \epsilon_{ik}\right) \tag{5.6}$$

In a two-stage clustering framework, a genotype will have some difference in its interaction profile GE'_{ik}, as well as a difference in level, to the first stage cluster to which it has been classified. The difference in level is broken into two components — one for the second stage cluster g and the other for any remaining difference it has in level from its second stage cluster G'_i. If genotype i is in first stage cluster f, and second stage cluster g, the two terms GE'_{ik} and G'_i in the above expression are confounded with the model's error component ϵ_{ik} and will be realized as residuals by fitting the model.

5.6 Summary

Whether it be by use of dendrograms or the model proposed in (5.6), two-stage clustering has been shown in this chapter to have the power to describe the performance of a large number of genotypes over a large number of environments. It seeks to group genotypes by similarity of performance in two ways; first by clustering on the similarity of G×E interaction profiles, and then using the similarity of mean performance over the set of environments. In the next chapter, information gained from two-stage clustering is used to impute missing values in a G×E matrix, thus allowing standard G×E analyses to occur.

Chapter 6

Two-stage imputation

6.1 Introduction to imputation of missing data

Often there are insufficient resources to test all genotypes in all environments, yet researchers would still like to estimate the yield from an untested combination. The knowledge gained from two-stage clustering of genotypes, discussed in the previous chapter, indicates how a genotype would perform in an environment in which it has not been tested. In this chapter, this knowledge, combined with observed data, is used to estimate unknown performances.

Before moving into the discussion of what will be known as 'two-stage imputation' some background is provided. Little and Rubin (1987) is well recognized as a definitive reference for data imputation methods. While much of the methodology presented therein was not relevant to the principal data of this work, the questions asked provided the benchmark against which all imputation methods should be evaluated. These authors are responsible for the language of missing data, and an understanding of the key definitions of their work is essential to this study.

The method developed in this chapter assumes data to be 'Missing at Random', defined by Little and Rubin (1987) to mean the cause of missing data does not depend on the missing values themselves. This does not suggest that missing values are a simple random sample from the entire data, which would be defined as 'Missing Completely at Random'. 'Missing Completely at Random' is therefore a restricted case of 'Missing at Random' (Little and Rubin, 1987).

To put this into context, data would not be 'Missing at Random' if onion varieties were tested in only those environments where it was known they would succeed. The scenario that arose in the Onion Trials Programme was that any G×E combinations known to yield poorly were not tested. In this instance there is a possibility that imputation results would give erroneous predictions for these missing yields. When performing imputation, all causes of missing data should be quantified, and then removed in order to improve validity and accuracy of the imputed values. If this cannot be achieved, the scope of results may

need to be reconsidered. Creating data that reflected crop failure was therefore necessary to ensure data were 'Missing at Random'. This gave even greater incentive to include 'zero yields' in the data sets than was already discussed in Section 3.6.

A summary of some imputation methods from Little and Rubin (1987) follows to provide a background to this chapter. These methods were considered for use with the data arising from the Onion Trials Programme; their failure highlighted the need to develop new methodology, and exposed undesirable attributes which needed to be avoided by a new method. The first method considered is due to Hartley in 1956, and can be used to find an imputed value for a single missing observation. This method:

1. Creates three augmented data sets by taking different estimates for the missing value.
2. Fits the model for each of these three augmented data sets.
3. Fits a quadratic to the residual sum of squares resulting from these three analyses.
4. Minimizes this quadratic to find an estimate for the imputed value that would have the lowest residual sum of squares.

The method is computationally inexpensive as, unlike many imputation methods, it is not iterative. Joint consideration of combinations of missing values would result in exponentially increasing complexity. If there are two missing values in a data set, nine combinations of the missing values would need to be put into analyses and a bivariate quadratic function minimized. This method was rejected as it would be difficult to implement for sparse data, such as the data from the Onion Trials Programme.

The Healy-Westmacott method, also proposed in 1956, is an iterative method that in general:

1. Selects a set of initial starting values to replace missing values in an analysis.
2. Fits a model to the augmented data.
3. Finds a new set of predicted values for the missing values using this model.
4. Updates the augmented data by replacing missing values with the current set of predicted values.
5. Iterates until the updated values do not change between iterations, or some other stopping criterion is satisfied.

The Healy-Westmacott method was shown to be an application of the EM algorithm by McLachlan and Krishnan (1997), if the errors from the model are normally distributed. The application of the EM-AMMI model of Gauch and Zobel (1990) in Section 3.8 is an example of this algorithm. Its use is discussed further towards the end of Section 6.2.

Little and Rubin (1987) presented what they called the 'unconditional means' imputation method. In general this method uses the observed mean of a variable to impute all missing values in that variable, and is therefore sometimes known as 'mean substitution. In a G×E context, this would be equivalent to using the means of each environment in place of any missing values. This method was clearly undesirable when working with sparse data as the environment means were dependent on the varieties that were tested. This approach would also not help determine which varieties were 'winners' in environments. Little and Rubin (1987) showed how this method could be used when data are 'Missing Completely at Random', but noted that estimates of variance provided are unsatisfactory. As the data arising from the Onion Trials Programme are known not to be 'Missing Completely at Random', the unconditional means method was seen to have little potential in this work. One point in its favour is that the use of variable means to substitute for missing values results in parameter estimates that are usually more accurate than simple deletion of any observation with missing data (Malhotra, 1987).

Buck's method was also presented by Little and Rubin (1987) and was referred to as the 'conditional means' method. This method uses only the observations with complete data in a multiple regression. The variable having a missing value would be the response in this multiple regression, and all other variables used as the explanatory variables. The available data for the observation with the missing value are inserted into the generated multiple regression model to predict the missing value. This imputed value is therefore conditional on the complete data used for its prediction. If this method is applied directly to G×E data, the data from other environments would be used as one option for predicting performances in the environment where data are missing. In most G×E analyses, the environments are independent of one another making this method inappropriate. If available data is multi-attribute, however, there may be a relationship between attributes that can be used to advantage. The multi-attribute data used extensively by Basford and Tukey (1998) has six responses for each G×E combination. In this case a missing yield could be imputed using a multiple regression of yield on the other five response variables. A second option would be to use the two factors for genotypes and environments as explanatory variables in the linear model used to predict missing yields. There is no scope for prediction of yields that includes G×E interaction effects in this case, so the method is limited to additive G×E data.

In their summary of these imputation methods, Little and Rubin (1987) noted the following reasons why it is hard to recommend any of the above methods:

1. Their performance is unreliable.

2. *Ad hoc* adjustments are often required to yield satisfactory estimates.

3. It is not easy to distinguish situations when the methods work from situations when they do not.

4. They are unable to provide simple correct answers when interval estimates are required.

Another method of imputing missing multivariate data is to use a pre-determined number of randomly selected sets of values; each of these imputed data sets is then analysed separately, and results are then averaged. 'Multiple imputation', as this method is known, is reliant on choosing the correct distribution from which to take random observations that will substitute missing entries in the incomplete data set. The two-way model for unreplicated G×E data

$$Y_{ik} = \mu + G_i + E_k + (GE_{ik} + \epsilon_{ik})$$

given in (4.16) can be used to find the expected value of missing values. The expected value of the bracketed terms for G ×E interaction and error is zero, as

$$\sum_{i=1}^{I}\sum_{k=1}^{K} GE_{ik} = \sum_{i=1}^{I}\sum_{k=1}^{K} \epsilon_{ik} = 0.$$

Assuming these values are normally distributed, a normal distribution for the imputed values can be established in the following way:

1. Fit the additive model
$$Y_{ik} = \mu + G_i + E_k + \epsilon_{ik}$$
given in (2.1) to the data.

2. Use the parameter values from this model to determine the expected values of the missing entries.

3. Use the error MS from this model as the variance of the distributions used to impute missing entries.

This method for multiple imputation is therefore an extension of the second option for applying Buck's method discussed above.

What benefit will such a proposed method provide? Given that the expected value of the imputed values does not include any G×E interaction, its usefulness is likely to be minimal. Multiple imputation techniques focus on the distribution of parameters, rather than on the actual imputed values themselves. Advantage cannot be taken of beneficial interactions, without establishing some mechanism for determining the nature of the G×E interaction effects on particular G×E combinations.

If the G×E interaction component in the analysis was viewed as minor compared to the genotype main effect, this approach could provide a useful means of determining which genotypes showed the greatest wide adaptation potential. In this case the distribution of

particular parameter estimates would be compared to decide which genotypes had the highest average yield across all environments.

If the number of missing entries in the G×E matrix is small, multiple imputation is likely to provide a useful means of parameter estimation. With data as sparse as that arising from the Onion Trials Programme, initial estimates of genotype and environment main effects would not be as accurate as those that would arise from nearly complete data. The parameter estimates used to create imputed values are then likely to be altered significantly as a result of fitting the same (additive) model to the imputed data set. There seemed little point in creating a large number of data sets that would have little in common to allow parameter estimation.

When working with sparse G×E data, the potential of methods described in this section was limited by one (or both) of two reasons. First, the methods could not allow for the existence of G×E interaction and therefore had limited value in deciding which genotypes were likely to succeed in different types of environment. Second, the amount of missing data was too great for the methods to function with the surety of attaining sufficient accuracy. A method is now proposed that can be applied to sparse G×E data which recognizes the importance of G×E interaction.

6.2 Imputing missing G×E data

This section describes how two-stage clustering was adapted to create two-stage imputation. Models were presented in Section 5.5 that describe yields in terms of the results of clustering observations from both two-stage clustering and clustering based on Euclidean distance. Two-stage imputation is described in terms of the splitting of the two components of genotype performance — shape and level, rather than making reference to the model of (5.6). Approaches developed in Chapters 4 and 5 were used to formulate the following step-by-step process for imputing missing values.

Step 1 Perform first-stage clustering of genotypes in the standard way, or, if necessary, until the cluster of the genotype with the missing value has an observation in the environment of the missing value.

Step 2 For a genotype with a missing G×E yield, identify the shape-similar genotypes from Step 1.

Step 3 Determine the difference in level, between the genotype with the missing yield and a shape-similar genotype; add this difference to the yield of the shape-similar genotype in the environment where the yield is missing. Repeat for all shape-similar genotypes that have a yield in this environment.

Step 4 Calculate the mean of the estimates found in Step 3. This is the imputed value for the missing yield.

Step 5 If an imputed value is greater than the observed maximum (or less than the observed minimum) for a given environment then replace the imputed value by the observed environment maximum (minimum).

Continuation of clustering in Step 1 is necessary if imputed values for all untested combinations are to be found. When working with sparse data it is common for a first stage cluster to have no observed data for some environments; in such cases, the imputation will be based on the yields of the most shape-similar genotypes that do have data in the environment where the yield is missing.

The difference in level mentioned in Step 3, is equal to $\bar{y}_{i\cdot}^{(j)} - \bar{y}_{j\cdot}^{(i)}$. If few or no common environments exist, the intermediate genotypes $B_1, \ldots, B_n$, introduced in Section 4.6, are used to estimate the required difference in level. An overall difference in level can be approximated by the sum, over a path, of the differences in level. These overall difference estimates are then averaged. Using the notation of Section 4.6 this level difference is

$$\sum_{l=1}^{n} \left(\left[\bar{y}_{A_i \cdot}^{(B_l)} - \bar{y}_{B_l \cdot}^{(A_i)} \right] + \left[\bar{y}_{B_l \cdot}^{(A_j)} - \bar{y}_{A_j \cdot}^{(B_l)} \right] \right) / n \tag{6.1}$$

The final step was included to prevent imputation of negative yields. In fact, it produces imputed values that fall inside the observed range of yields in each environment, thus ensuring that there are no unrealistic imputed values. When decisions were made over selection of the best genotypes for an environment, this 'trimming' guaranteed that at least one tested genotype would be selected. The best values therefore came from a group that contained the genotype with the maximum observed yield.

Two-stage imputation is now illustrated using Figure 6.1 which shows the performance pattern of three genotypes across six environments; the data are artificial but allows imputation methods to be contrasted. Genotype A has a missing yield in the sixth environment, while in the other five environments it is most similar in shape to genotype C (a distance measure based on G×E interaction would be near zero for genotypes A and C). Measuring the difference in level between A and C and adjusting the yield of C in the sixth environment by this amount would give an estimate of the missing yield, indicated by the point marked with an open circle in Figure 6.1. The imputation is therefore based on the fact that the genotypes have similar interaction profiles; it does not mix the main effect and interaction as would proximity-based imputation strategies.

Proximity-based imputation strategies

An existing imputation strategy proposed by Drake (1981) uses output from clustering to estimate missing observations. Routines were obtained for the statistical package S-

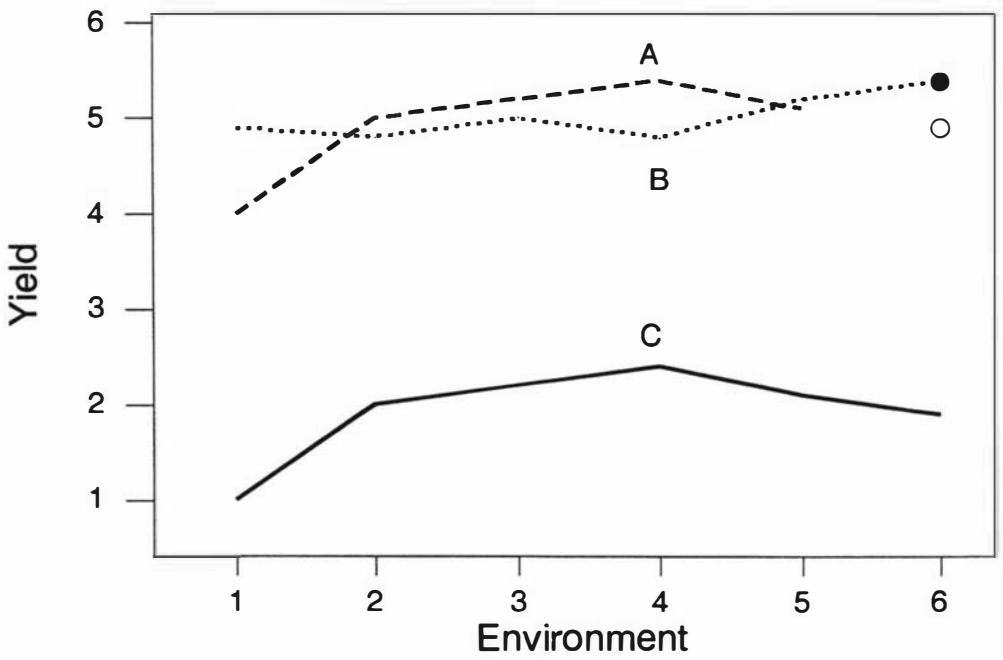


Figure 6.1: Yields of three genotypes plotted against environment. Genotypes A and C are deemed similar by interaction distance, while genotypes A and B are closest using squared Euclidean distance.

Plus that apply this approach, which will be referred to as the 'nearest cluster' method. For a genotype with missing values the method uses the closest cluster of genotypes, under squared Euclidean distance (the numerator of (4.4)), to impute the missing values. Specifically, a genotype with a missing yield that first merges with a cluster of other genotypes will use the mean of that cluster as the estimate for the missing yield.

A pair of genotypes which are nearest under Euclidean distance, such as genotypes A and B in Figure 6.1, do not necessarily provide good substitutes for each other when values are missing. This is seen in Figure 6.1, where the yield of genotype B in the sixth environment, marked with a solid circle, appears less appropriate for genotype A than the two-stage imputed value.

The method of Drake (1981) imputes missing values as the mean yield of all other genotypes with which a genotype first clusters, in the environment where the yield is missing. As a means of gauging the value of this averaging, a second Euclidean distance based imputation method was also applied. This method, called 'closest observation', uses the data from the closest genotype that has data available in the environment where the yield is missing. If the closest genotype also has a missing value, the second closest would be used as the substitute.

Two-stage imputation versus model-based imputation

Imputation methods use a model to determine imputed values (Little, 1988); in many instances these models are implied, while for others such as EM-AMMI (Gauch and Zobel, 1990), the model is explicitly used as part of the procedure. The model presented in (5.6) for two-stage clustering is implied by two-stage imputation. The success of two-stage imputation relies on the ability of remaining data to reconstitute truly similar interaction profiles; it is, therefore, driven by the data itself rather than by some predetermined class of models. The ability of a class of models to adequately describe the data is difficult to establish with incomplete data. As a consequence, there is a risk in choosing an unsuitable class of model for an incomplete data set. Another concern is the inability of certain models, when only partial data are available, to reconstitute information that would be gained from complete data. This difficulty is particularly evident when no replicates for a given genotype-environment combination are available. For these reasons, the chosen model may strongly influence the value that is imputed, as seen when the EM-AMMI model was implemented in Section 3.8. The EM-AMMI method, or any model based method, appears to be more useful when imputing missing data in replicated trials, than when imputing missing data in unreplicated trials. In the first case a satisfactory model is more clearly determined, while in the second case the two-stage imputation method is expected to be more appropriate.

Section 6.4 discusses the results after testing two-stage imputation on simulated incomplete data. Although EM-AMMI modelling offers an alternative to two-stage and other clustering imputation strategies, there are several problems that make its simulation testing impracticable. The particular AMMI model to be employed for each data set is difficult to determine, as shown by the wide ranging results found in Section 3.8. Although this is a concern, two other decisions will impact on the outcome when applying the EM algorithm: first, a method must be chosen to determine starting values for use in place of missing values and second, the point to stop iterating through the algorithm must be determined. In Section 3.8 the criterion was a maximum change of 0.005 in the fitted values, but there was little guarantee that this stopping criterion would stop the algorithm at the optimal solution. If simulation testing was to include EM-AMMI models, this stopping criterion should be set so that the algorithnm is guaranteed to stop at the optimal solution. The EM-AMMI model is therefore impractical for comparative analysis through simulation testing due to its lengthy computation time to reach convergence.

6.3 An example of two-stage imputation

This section presents an example of imputing missing G×E data using both two-stage imputation and nearest cluster imputation. The data set used in this example was introduced in Section 5.3. The fifteen randomly deleted observations and their imputed

Genotype	Environment	Omitted yields	Two-stage imputation	Nearest cluster imputation
2*	R70	−0.562	−0.33	−0.218
5	L71	−0.648	0.868	0.799
5*	N71	0.497	0.996	1.301
6*	N71	1.23	0.551	−0.145
7	R71	1.434	0.612	0.948
10*	L71	0.852	0.923	−0.062
14	N70	0.93	0.141	0.469
19*	B70	−1.308	−0.847	−0.121
19	L71	0.375	−1.282	−1.263
24	B70	0.809	−0.563	−0.557
26	B71	0.738	−0.291	0.039
30*	N70	−0.48	−0.174	−0.082
37*	N71	−0.599	−0.093	−0.007
52*	B70	−1.483	−0.939	−0.841
53*	R70	1.674	1.625	1.469

Table 6.1: The fifteen induced missing yield observations (standardized within environments) and the imputed values found using both the two-stage and nearest cluster approaches. An asterisk marks genotypes for which the closer imputed value was found by the two-stage approach. (Data source: Basford and Tukey, 1998)

values are listed in Table 6.1. Imputation calculations for each method are now provided in detail.

Figure 6.2 presents the clustering of genotypes using the mean squared Euclidean distance from Ouyang *et al.* (1995) in (4.4) and the incremental sum of squares method of forming clusters. Use of this averaged distance measure was an improvement on the original implementation of Drake (1981), being more appropriate for the reasons discussed in Chapter 4. Applying nearest cluster imputation would, for instance, use the yields of genotypes 51 and 52 in place of each other's missing yields as these are deemed closest to each other. Thus the missing yield of genotype 52 in environment B70 would be imputed using the yield of genotype 51 in B70, namely −0.841.

The similarity of the clustering in Figure 6.2 with that of Figure 5.2 (on page 127) reflects the low importance of the differences in level between many of the genotypes under examination. An exception to this is genotype 54, which is clustered differently in the two figures. Its level similarity with genotypes 8 and 29 forces the clustering in Figure 6.2, while its distinct G×E interaction profile forces it to remain outside the clustering until late in Figure 5.2.

The first stage cluster that contains genotypes 51, 52, and 58 (at the right of Figure 5.2) is used to illustrate the imputation of the missing value of genotype 52 in B70. The yields of these three genotypes are plotted against environments in Figure 6.3.

Two-stage imputation of a missing yield does not use the results of second stage clus-

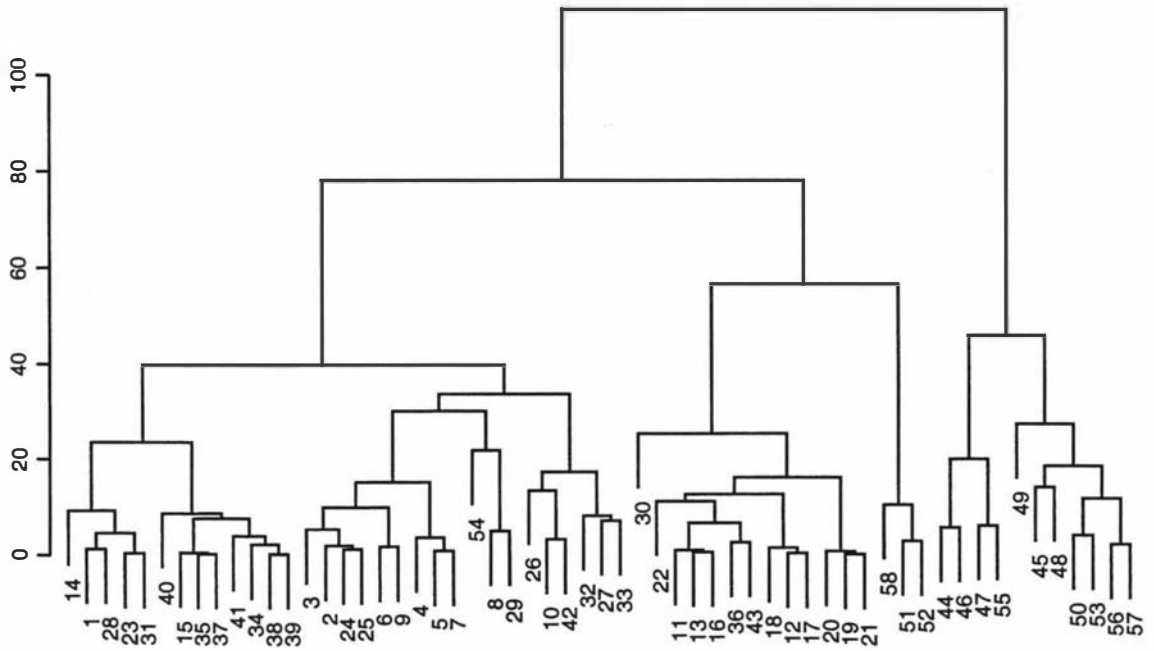


Figure 6.2: Fifty-eight genotypes clustered using the mean squared Euclidean distance of Ouyang *et al.* (1995) and incremental sum of squares.

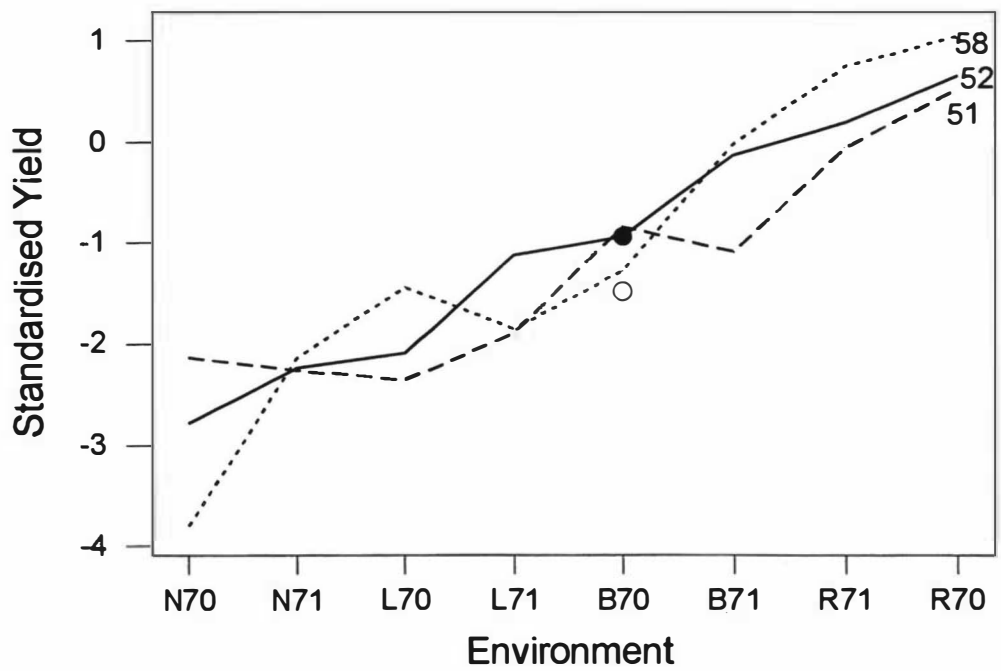


Figure 6.3: Yields of the three similar genotypes (51, 52, and 58) plotted against an ordered environmental index, calculated as the mean of these genotypes using imputed values where necessary. The point marked with a solid circle is the imputed value of genotype 52, and the open circle marks the omitted value.

tering but does use the second stage distance. Recall that first stage clustering produces clusters of genotypes that have similar interaction profiles; differences in level between these genotypes were then used to impute a missing value. Thus, for example, the missing yield of genotype 52 in B70 was imputed using genotypes 51 and 58, as these were the only genotypes deemed similar to genotype 52 when first stage clustering was truncated. In B70, genotype 51 yielded −0.841, while genotype 58 yielded −1.278. Genotype 51 on average yielded 0.250 less than genotype 52, so provided an estimate of the standardized yield for genotype 52 in B70 of −0.591. Similarly, genotype 58 on average yielded 0.009 more than genotype 52, giving an estimate of −1.287. These two values were then averaged to give the imputed value for genotype 52 in B70 of −0.939, found in Table 6.1.

Figure 6.3 reproduces the yield profiles of the three genotypes mentioned in the imputation of genotype 52 and indicates the omitted yield of genotype 52 in B70. Genotypes 51 and 58 are similar in their interaction profiles to genotype 52 as determined by first stage clustering. The imputed value for genotype 52, found using the two-stage approach, is marked with a solid circle in Figure 6.3, while an open circle marks the omitted yield.

It can be seen from Table 6.1 that in nine of the fifteen cases two-stage imputation provided a closer imputed value than the nearest cluster approach. This is typical for this amount of missing data, as will be shown in Section 6.4.

For this example, an overall comparison of the results from the two imputation methods, using mean squared error (MSE) of the estimate, showed that two-stage imputation performed better, with an MSE of 0.727 compared to an MSE of 0.896 for the nearest cluster method. Section 6.4 compares the methods by varying the amount of missing data in the G×E matrix, using a number of G×E data sets available in the literature.

6.4 Comprehensive testing of two-stage imputation

Comprehensive testing of two-stage imputation was required to compare its effectiveness with that of the nearest cluster method. This section also compares two-stage imputation's effectiveness to that of the closest observation method described in Section 6.2, and imputation using values randomly selected from the observed yields of other genotypes within the same environment. Varying amounts of missing data were simulated by randomly deleting values from complete G×E matrices in order to compare the imputation methods. The following procedure was used on a variety of data sets from the G×E literature:

1. Randomly remove the desired number of elements from the complete G×E matrix. Data in these reduced matrices are therefore 'Missing Completely at Random'.

2. Check this new matrix for the representation of each genotype and environment. Every genotype (environment) must be represented in some minimum number of

environments (genotypes); if $P_{ii} < P_m$, or $Q_{kk} < Q_m$ for some genotype i or environemnt K, then start again.

3. If this matrix is partitioned, that is, there is no direct or indirect commonality of environments between every pair of genotypes, then start again.

4. Impute missing values using all four methods.

5. Record results for the following criteria:

   - The mean squared error for each method
   - The proportion of cases imputed more accurately by each method, in comparison to each other method
   - Spearman's rank correlation coefficient between the deleted values and the imputed values for each method.

Greater detail is now added to the above summary.

Data used for simulation testing

The effectiveness of two-stage imputation has been tested using six data sets; the Mungomery *et al.* (1974) data described in Section 5.3, data sets from Gauch (1992), Ramey and Rosielle (1983), two from Flores *et al.* (1998), and another from Fox and Rathjen (1981). These data sets were used as they covered a range of sizes, and more importantly, were complete. These data sets are available on the CD-ROM accompanying this volume, while the full data can be obtained from the original references. Table 6.2 shows the size and shape of the G×E matrices and the G×E interaction sum of squares. Means over G×E

Data set	G×E matrix size	Genotype SS	G×E interaction SS	Interaction SS as % of total SS	Transformed
Flores 1	15×12	17120782	18191558	51.52	no
		21.97	146.03	86.92	yes
Flores 2	11×16	5634380	24798300	81.49	no
		0.69	159.31	99.57	yes
Fox	22×14	69.28	224.72	76.44	yes
Gauch	7×10	7117668	39728718	84.81	no
Mungomery	58×8	238.44	217.56	47.71	yes
Ramey	15×9	9,506,462	16334640	63.21	no

Table 6.2: Summary details of data sets used in testing. Data sets were extracted from Flores *et al.* (1998), Fox and Rathjen (1981), Gauch (1992), Mungomery *et al.* (1974), and Ramey and Rosielle (1983); Shorthand has been used in this and the following tables for brevity. Note that the transformed data sets are standardized within environments, and that this has an effect on the magnitude of the genotype main effect and G×E interaction sums of squares. The exact data used are contained on the CD-ROM in the back of this volume for convenience.

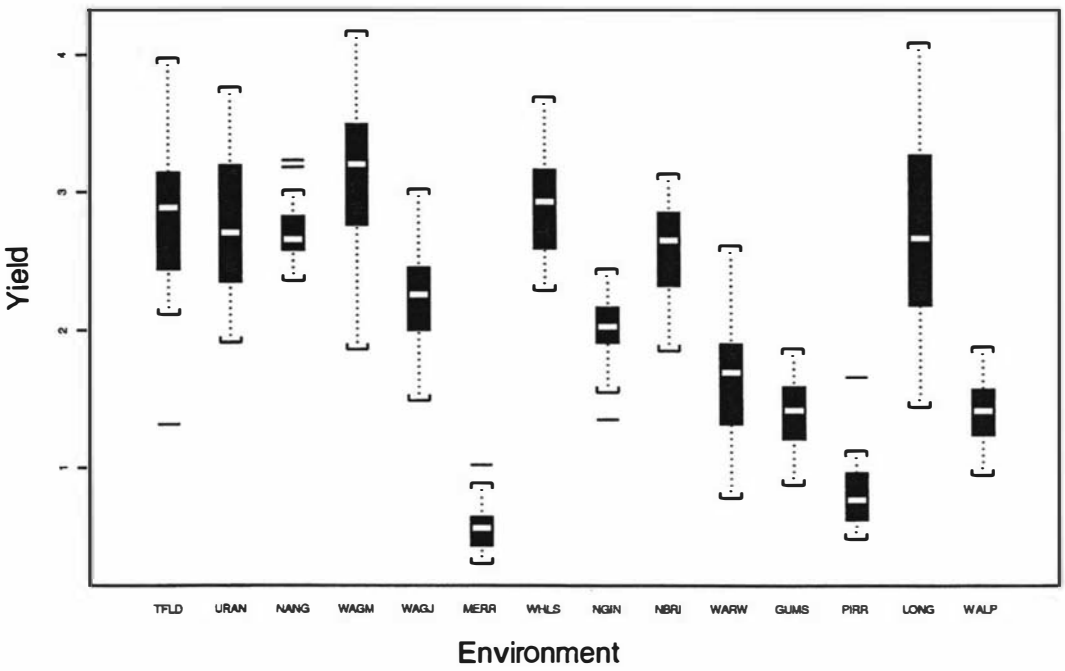


Figure 6.4: Boxplots of yields for each of the 14 environments in the Fox and Rathjen (1981) data.

combinations have been used where replicates were available. As discussed in Section 5.2, appropriate transformation should be considered as a general rule to equalize within-environment variances. For example, the Fox and Rathjen (1981) data was standardized within environments in the same way as the Mungomery *et al.* (1974) data in Section 5.2. This transformation was deemed necessary due to the heteroscedasticity of the environments. Figure 6.4 shows the boxplots of the yields from each of the environments in this data. Minitab software was used to perform Levene's test for homogeneity of variance, and yielded a test statistic of 5.933 which determined that the within-environment variances were significantly different. There was clearly a need to transform this data. The effects of standardizing data were investigated through use of raw and transformed versions of the two Flores *et al.* (1998) data sets.

Two approaches were used to ensure that sufficient information remained in the G×E matrix after data removal. First, each incomplete matrix was checked to ensure that all genotypes were grown in a minimum number of environments, $P_{ii} \geqslant P_m$. In the testing, this minimum level P_m was chosen at four environments as this was at least half of the total number of environments in the smaller data sets used. Second, care was taken that data deletion used to form the incomplete G×E matrices did not lead to two unlinked data sets, so that it remained possible to impute all missing G×E yields.

Imputation via two-stage, nearest cluster, or using the closest observation method would be impossible for some G×E combinations if the G×E matrix became unlinked. A pair of genotypes were, however, allowed to have no common environments. In this

instance the distance between these genotypes would need to be estimated using the method discussed in Section 4.6. The value of P_s from (4.26) was set at four in this work, as this integer must be less than or equal to the minimum representation of environments in a given G×E matrix. If P_s was allowed to exceed the minimum representation P_m, an under-represented genotype would have undefined distances to every other genotype.

Results from simulation testing

Imputations were then compared with the deleted values. Each pair of methods was compared using both the MSE of the imputed values and the proportion of values that were imputed more accurately by one method than the other. For comparative purposes a value randomly selected from those within the environment was also used as a fourth imputation method. One thousand runs were performed for each data set and missing data amount so that a sufficiently large number of randomly imputed values could be compared to imputed values found by the other three methods. These results are shown in the tables on pages 157 through 160. Note that it is possible for two methods to give the same imputed value; this is counted in favour of the second listed method in any comparison and thus the figures provided understate the performance of the first listed method.

The two-stage method consistently outperformed the nearest cluster method; this margin increased as the amount of missing data increased in the sets that had a large number of environments. The reason for this appeared to be the ability of the reduced data set to retain the qualitative G×E interaction structure of the complete matrix when there were more environments over which to compare genotypes. In these situations the genotype clustering was less likely to be altered by the deletion of particular G×E combinations. The improvement of the two-stage method over the nearest cluster method was always greater than in the case when only three points were removed from the Ramey and Rosielle (1983) data. The mean squared error in this worst case was lower in only 59.6% of all runs. When the second criterion was considered, the two-stage imputed values were consistently more likely to be closer to the omitted value than the nearest cluster imputed values. The case where 60 points were removed from the second standardized data set from Flores *et al.* (1998) had the lowest average proportion (0.495) of omitted values imputed better by two-stage imputation. The mean squared error comparison for this worst case, however, showed that while on average just under half of the omitted values were imputed more accurately using two-stage imputation, the MSE was lower for two-stage imputation in 81.5% of runs. The advantage gained by the two-stage imputed values, when they were closer to the omitted yields than nearest cluster imputed values, exceeded that gained by the nearest cluster imputed values when they were closer to the omitted yields.

Neither the nearest cluster, nor the closest observation method, consistently outperformed random imputation in terms of the proportion of imputed values that were closer

Data set	Number of points removed	Two-stage v. nearest cluster	Two-stage v. random	Nearest cluster v. random
Flores 1	3	76.7 (0.627)	79.7 (0.664)	64.0 (0.567)
	5	80.0 (0.610)	84.1 (0.646)	72.2 (0.566)
	10	87.6 (0.606)	92.3 (0.648)	79.6 (0.562)
	15	89.7 (0.596)	95.4 (0.637)	85.0 (0.552)
	20	92.9 (0.590)	97.2 (0.630)	88.9 (0.551)
	25	92.5 (0.583)	98.3 (0.631)	90.4 (0.549)
	30	92.8 (0.580)	98.8 (0.634)	92.8 (0.554)
	35	89.3 (0.569)	99.1 (0.624)	93.9 (0.548)
Flores 2	3	66.2 (0.562)	66.3 (0.606)	60.0 (0.549)
	5	67.6 (0.545)	71.2 (0.598)	62.4 (0.542)
	10	70.1 (0.541)	72.4 (0.583)	64.6 (0.527)
	15	69.3 (0.523)	73.6 (0.576)	65.3 (0.520)
	20	75.4 (0.529)	75.8 (0.576)	64.3 (0.513)
	25	74.4 (0.523)	79.3 (0.576)	68.9 (0.512)
	30	77.2 (0.522)	80.0 (0.573)	67.2 (0.510)
	35	77.8 (0.522)	79.2 (0.564)	65.6 (0.498)
	40	79.3 (0.517)	79.4 (0.561)	64.3 (0.493)
	45	79.5 (0.516)	79.8 (0.557)	64.4 (0.492)
	50	81.3 (0.513)	79.4 (0.552)	65.6 (0.485)
	55	80.0 (0.508)	82.3 (0.551)	63.4 (0.486)
	60	81.0 (0.500)	80.9 (0.544)	64.8 (0.480)
Gauch	3	82.2 (0.667)	91.6 (0.710)	71.2 (0.484)
	5	77.6 (0.601)	95.0 (0.670)	82.7 (0.477)
	10	82.3 (0.636)	92.3 (0.679)	81.8 (0.501)
Ramey	3	59.6 (0.522)	75.6 (0.633)	65.8 (0.574)
	5	63.7 (0.514)	77.8 (0.620)	69.5 (0.573)
	10	71.6 (0.509)	84.4 (0.622)	73.8 (0.567)
	15	72.8 (0.507)	86.5 (0.615)	74.9 (0.567)

Table 6.3: Comparison of all pairs of methods using four untransformed data sets and varying levels of missing data. The two measures are the percentage of 1000 runs for which the mean squared error for the first method is lower, and (in brackets) the average proportion of the first method's imputed values that are closer to the missing values. *Continued on page 158.*

Data set	Number of points removed	Two-stage v. closest observation	Nearest cluster v. closest observation	Closest observation v. random
Flores 1	3	76.8(0.631)	35.5(0.215)	63.0(0.552)
	5	82.2(0.616)	39.7(0.235)	70.7(0.550)
	10	89.5(0.617)	36.9(0.235)	75.8(0.538)
	15	92.1(0.602)	39.4(0.230)	82.0(0.532)
	20	93.8(0.605)	34.4(0.237)	85.9(0.525)
	25	94.5(0.593)	35.8(0.231)	88.1(0.528)
	30	94.6(0.590)	34.2(0.233)	90.1(0.531)
	35	93.0(0.577)	36.9(0.241)	91.8(0.523)
Flores 2	3	63.2(0.575)	44.1(0.243)	59.1(0.498)
	5	63.5(0.561)	46.4(0.252)	62.8(0.501)
	10	62.9(0.545)	48.6(0.250)	64.6(0.497)
	15	67.2(0.537)	46.5(0.251)	63.2(0.483)
	20	70.9(0.529)	47.8(0.251)	63.6(0.484)
	25	67.1(0.525)	48.8(0.245)	69.3(0.485)
	30	70.8(0.526)	49.6(0.253)	66.4(0.478)
	35	70.0(0.522)	52.2(0.252)	66.8(0.468)
	40	71.4(0.518)	53.6(0.252)	66.8(0.464)
	45	71.3(0.515)	53.3(0.250)	68.7(0.464)
	50	71.7(0.515)	53.5(0.249)	67.0(0.455)
	55	69.1(0.507)	54.8(0.248)	68.5(0.457)
	60	72.7(0.504)	53.5(0.256)	66.3(0.446)
Gauch	3	81.4(0.643)	14.0(0.023)	72.5(0.484)
	5	72.5(0.568)	28.1(0.051)	84.4(0.479)
	10	69.6(0.590)	63.5(0.160)	86.7(0.489)
Ramey	3	56.1(0.515)	32.1(0.174)	68.0(0.568)
	5	62.3(0.517)	38.8(0.207)	69.2(0.556)
	10	65.8(0.509)	45.4(0.218)	77.9(0.551)
	15	65.4(0.509)	48.4(0.234)	79.6(0.548)

Table 6.3: *Continued from page 157.* Comparison of all pairs of methods using four untransformed data sets and varying levels of missing data. The two measures are the percentage of 1000 runs for which the mean squared error for the first method is lower, and (in brackets) the average proportion of the first method's imputed values that are closer to the missing values.

Data set	Number of points removed	Two-stage v. nearest cluster	Two-stage v. random	Nearest cluster v. random
Flores 1	3	75.5 (0.610)	77.8 (0.644)	66.0 (0.548)
	5	77.7 (0.601)	82.9 (0.633)	69.2 (0.539)
	10	82.7 (0.585)	93.4 (0.641)	81.6 (0.559)
	15	88.2 (0.589)	95.8 (0.632)	84.4 (0.546)
	20	90.5 (0.579)	97.8 (0.639)	89.7 (0.557)
	25	90.5 (0.579)	99.3 (0.635)	92.5 (0.551)
	30	92.3 (0.573)	99.3 (0.628)	92.5 (0.550)
	35	92.4 (0.573)	99.6 (0.625)	95.0 (0.545)
Flores2	3	57.1 (0.542)	74.2 (0.617)	68.8 (0.560)
	5	63.3 (0.556)	79.3 (0.612)	71.3 (0.553)
	10	70.0 (0.552)	86.9 (0.618)	79.1 (0.548)
	15	71.8 (0.533)	89.4 (0.613)	79.5 (0.541)
	20	75.7 (0.527)	91.5 (0.608)	82.3 (0.540)
	25	73.6 (0.516)	92.4 (0.605)	83.8 (0.543)
	30	76.2 (0.516)	94.0 (0.594)	84.6 (0.526)
	35	79.3 (0.511)	93.7 (0.592)	84.1 (0.523)
	40	78.4 (0.505)	94.0 (0.582)	85.0 (0.514)
	45	80.0 (0.501)	92.8 (0.572)	81.7 (0.510)
	50	80.2 (0.504)	93.3 (0.570)	79.5 (0.501)
	55	81.6 (0.497)	94.9 (0.562)	84.8 (0.496)
	60	81.5 (0.495)	95.1 (0.558)	81.1 (0.489)
Fox	3	61.7 (0.560)	66.5 (0.594)	61.3 (0.555)
	5	64.6 (0.553)	71.0 (0.591)	61.7 (0.539)
	10	73.4 (0.558)	76.1 (0.596)	64.7 (0.537)
	15	75.5 (0.552)	80.3 (0.594)	65.8 (0.538)
	20	80.2 (0.556)	82.6 (0.583)	67.5 (0.532)
	25	81.1 (0.551)	84.7 (0.589)	69.4 (0.529)
	30	81.0 (0.549)	85.2 (0.579)	70.9 (0.520)
	35	83.1 (0.554)	88.2 (0.578)	71.9 (0.520)
	40	83.9 (0.546)	90.3 (0.582)	75.5 (0.522)
	45	85.9 (0.550)	90.5 (0.577)	73.9 (0.520)
	50	86.4 (0.546)	91.8 (0.575)	76.7 (0.519)
	55	87.5 (0.549)	93.3 (0.575)	77.6 (0.517)
	60	87.3 (0.545)	92.8 (0.576)	77.3 (0.519)
	65	87.3 (0.547)	92.8 (0.570)	79.7 (0.513)
	70	88.2 (0.547)	94.9 (0.571)	78.3 (0.514)
	75	89.4 (0.546)	94.7 (0.572)	79.1 (0.513)
Mungomery	3	64.3 (0.586)	85.9 (0.696)	77.5 (0.644)
	5	70.1 (0.597)	90.0 (0.702)	82.0 (0.648)
	10	77.4 (0.598)	97.0 (0.701)	90.3 (0.635)
	15	84.9 (0.598)	98.3 (0.699)	92.4 (0.631)
	20	86.0 (0.588)	99.6 (0.707)	97.3 (0.640)

Table 6.4: Comparison of all pairs of methods using four sets of within-environment standardized data and varying levels of missing data. The two measures are the percentage of 1000 runs for which the mean squared error for the first method is lower, and (in brackets) the average proportion of the first method's imputed values that are closer to the missing values. *Continued on page 160.*

Data set	Number of points removed	Two-stage v. closest observation	Nearest cluster v. closest observation	Closest observation v. random
Flores 1	3	78.5 (0.618)	30.4 (0.202)	64.0 (0.542)
	5	79.7 (0.608)	36.3 (0.212)	67.5 (0.529)
	10	86.1 (0.594)	34.0 (0.223)	77.3 (0.543)
	15	93.1 (0.595)	33.2 (0.226)	78.9 (0.526)
	20	94.9 (0.590)	32.3 (0.224)	85.9 (0.533)
	25	93.5 (0.581)	34.0 (0.228)	89.8 (0.535)
	30	94.8 (0.575)	36.3 (0.227)	90.5 (0.529)
	35	94.6 (0.571)	40.4 (0.231)	93.5 (0.528)
Flores2	3	61.2 (0.547)	23.9 (0.245)	61.9 (0.511)
	5	66.6 (0.551)	31.3 (0.237)	62.9 (0.507)
	10	75.5 (0.549)	32.0 (0.226)	71.3 (0.515)
	15	80.5 (0.540)	32.9 (0.224)	72.5 (0.504)
	20	81.6 (0.541)	32.7 (0.234)	76.5 (0.501)
	25	82.2 (0.528)	32.1 (0.236)	78.6 (0.504)
	30	83.5 (0.530)	33.8 (0.233)	77.9 (0.491)
	35	83.3 (0.523)	36.4 (0.235)	81.0 (0.490)
	40	81.5 (0.516)	38.2 (0.239)	81.6 (0.481)
	45	80.0 (0.507)	41.7 (0.233)	81.1 (0.479)
	50	80.2 (0.509)	44.0 (0.236)	81.0 (0.473)
	55	81.6 (0.502)	42.8 (0.239)	81.8 (0.466)
	60	81.5 (0.498)	47.8 (0.237)	80.3 (0.462)
Fox	3	58.2 (0.534)	45.3 (0.255)	65.0 (0.561)
	5	60.6 (0.526)	52.1 (0.253)	68.3 (0.547)
	10	59.0 (0.537)	59.4 (0.265)	73.6 (0.546)
	15	60.5 (0.523)	65.0 (0.245)	78.1 (0.554)
	20	62.3 (0.527)	67.3 (0.244)	79.6 (0.550)
	25	65.0 (0.532)	64.6 (0.246)	78.8 (0.539)
	30	63.3 (0.525)	69.2 (0.240)	80.8 (0.532)
	35	65.5 (0.527)	68.9 (0.239)	81.4 (0.533)
	40	66.4 (0.530)	69.1 (0.237)	85.8 (0.532)
	45	67.1 (0.529)	71.6 (0.232)	84.4 (0.529)
	50	66.9 (0.528)	70.9 (0.233)	87.0 (0.529)
	55	68.4 (0.531)	74.0 (0.236)	84.4 (0.528)
	60	71.4 (0.533)	68.3 (0.231)	86.4 (0.523)
	65	72.3 (0.531)	71.0 (0.232)	87.5 (0.520)
	70	71.9 (0.532)	70.9 (0.234)	87.4 (0.520)
	75	72.5 (0.529)	73.5 (0.230)	87.9 (0.519)
Mungomery	3	57.1(0.538)	48.9(0.173)	80.8(0.659)
	5	63.2(0.552)	56.3(0.179)	85.0(0.665)
	10	68.3(0.557)	63.9(0.174)	94.1(0.658)
	15	75.6(0.560)	66.2(0.177)	96.1(0.653)
	20	74.4(0.555)	67.7(0.184)	98.3(0.661)

Table 6.4: *Continued from page 159.* Comparison of all pairs of methods using four sets of within-environment standardized data and varying levels of missing data. The two measures are the percentage of 1000 runs for which the mean squared error for the first method is lower, and (in brackets) the average proportion of the first method's imputed values that are closer to the missing values.

to the omitted yield. On the other hand, when the criterion used to gauge effectiveness was the MSE, both these methods consistently outperformed the use of random values. Based on MSE, the two-stage method always improved on random insertion more strongly than either the nearest cluster or the closest observation methods. A comparison of the benefits of two-stage imputation versus each of the methods based on confounding of main effect and interaction, showed that the advantage of two-stage over nearest cluster always exceeded the advantage of two-stage over closest observation. There was one exception to this rule; the first data set from Flores *et al.* (1998), both raw and standardized, consistently showed the opposite result. The MSE results for the comparison of nearest cluster and closest observation imputation were generally low, and this was especially noticeable for both raw and standardized forms of this data set.

Both two-stage and nearest cluster imputation use results from clustering, while closest observation takes the single closest available data point to give the imputed value. The benefit of the clustering step in these imputation methods must be questioned. When comparing the two Euclidean distance based methods of imputing missing values, all data sets were found to initially show superior results from using closest observation imputation rather than nearest cluster imputation; on the other hand, the benefits of using a strategy based on clustering increased as more and more data were randomly removed. Four of the eight data sets did not reach the point where the nearest cluster method became superior in terms of the percentage of runs that had a lower MSE. As previously stated, the comparison favours the second method in the results given. The propensity for the nearest cluster and the closest observation methods to give the same imputed value was higher than that for any other pair of methods.

Tables 6.5 and 6.6 show the Spearman's rank correlations, averaged over 1000 runs, of the values found using each imputation method and the actual values randomly deleted from the complete G×E matrix. The variance of the 1000 correlation coefficients for each run is given in parentheses in Tables 6.5 and 6.6 to provide an indication of the consistency in these results. Although the correlations for each data set with three and five points removed were calculated, they have been discarded as correlations found on such small samples do not follow the same distribution (Siegel, 1956; Gibbons, 1970).

The correlations in Table 6.5 should generally be high, due to differences in environmental means. Consequently, no significance test was worth performing, as the basic assumptions of the null hypothesis would not be valid. That is, the expected value of the correlation coefficients would not be zero, even if no benefit was to be gained from imputed values. The use of the random method of imputation was therefore useful as it provided a baseline from which to gauge the performance of the three other methods.

Any differences in environment means were removed by the standardization of data sets. The relevant correlations presented in Table 6.6 are expected to be nearer zero, especially those based on the random imputation method. Indeed, the correlation for twenty

data set	Number of points removed	Two-stage	Nearest cluster	Closest observation	Random
Flores 1	10	0.800 (0.023)	0.720 (0.039)	0.710 (0.041)	0.599 (0.059)
	15	0.809 (0.013)	0.739 (0.021)	0.732 (0.023)	0.638 (0.038)
	20	0.814 (0.009)	0.748 (0.015)	0.741 (0.015)	0.644 (0.022)
	25	0.814 (0.007)	0.750 (0.011)	0.741 (0.011)	0.652 (0.017)
	30	0.811 (0.006)	0.752 (0.009)	0.740 (0.010)	0.642 (0.016)
	35	0.811 (0.005)	0.758 (0.007)	0.747 (0.007)	0.651 (0.012)
Flores 2	10	0.794 (0.027)	0.762 (0.031)	0.768 (0.028)	0.723 (0.037)
	15	0.806 (0.015)	0.774 (0.019)	0.777 (0.018)	0.749 (0.019)
	20	0.810 (0.010)	0.780 (0.012)	0.784 (0.012)	0.755 (0.014)
	25	0.807 (0.008)	0.779 (0.009)	0.789 (0.009)	0.755 (0.010)
	30	0.811 (0.007)	0.779 (0.009)	0.786 (0.008)	0.759 (0.009)
	35	0.811 (0.006)	0.782 (0.007)	0.790 (0.006)	0.761 (0.007)
	40	0.814 (0.004)	0.784 (0.006)	0.794 (0.005)	0.766 (0.006)
	45	0.808 (0.004)	0.783 (0.005)	0.789 (0.004)	0.766 (0.004)
	50	0.810 (0.003)	0.783 (0.004)	0.788 (0.004)	0.768 (0.004)
	55	0.809 (0.003)	0.783 (0.004)	0.790 (0.004)	0.767 (0.004)
	60	0.810 (0.003)	0.784 (0.003)	0.789 (0.003)	0.768 (0.004)
Gauch	10	0.860 (0.013)	0.797 (0.016)	0.823 (0.014)	0.673 (0.044)
Ramey	10	0.859 (0.012)	0.818 (0.019)	0.823 (0.018)	0.772 (0.022)
	15	0.868 (0.007)	0.832 (0.010)	0.839 (0.010)	0.783 (0.013)

Table 6.5: Mean correlation coefficients, over 1000 runs, of the actual deleted values and imputed values found using the four imputation methods. The four data sets used have not been transformed. The variance of each coefficient is given in parentheses.

points removed from the Mungomery *et al.* (1974) data was negative (–0.024), but there was insufficient evidence to suggest that it is significantly different from zero. This was a common occurrence across the correlations of deleted values and random imputation for the standardized data. The variances given in Table 6.6 indicated that these correlations were not always significantly positive for any of the imputation methods.

Two-stage imputation always showed the greatest correlation with the deleted values, and the correlations for the two Euclidean distance based methods were similar to each other. The three methods showed relatively little difference in their overall performance. The variance of the 1000 correlations for each amount of randomly deleted data showed that the consistency of the correlations was generally the same. The variances would be expected to decrease as the correlation increased, but rounding to three decimal places meant that the differences could not be seen clearly. They seemed trivial however, and not worth more in-depth investigation.

The benefits of standardizing the two data sets from Flores *et al.* (1998) were then considered. There was little difference in the patterns of results for the raw data (Table 6.3) and the standardized data (Table 6.4) for either data set, compared to the results from other data sets. There was a slight advantage in using standardized data from the second

Data set	Number of points removed	Two-stage	Nearest cluster	Closest observation	Random
Flores 1	10	0.439 (0.067)	0.319 (0.074)	0.233 (0.082)	−0.060 (0.107)
	15	0.440 (0.040)	0.303 (0.052)	0.213 (0.057)	−0.044 (0.068)
	20	0.446 (0.030)	0.316 (0.037)	0.227 (0.041)	−0.061 (0.015)
	25	0.441 (0.022)	0.306 (0.028)	0.227 (0.031)	−0.061 (0.041)
	30	0.432 (0.020)	0.300 (0.024)	0.224 (0.026)	−0.058 (0.035)
	35	0.424 (0.018)	0.294 (0.022)	0.224 (0.023)	−0.062 (0.029)
Flores 2	10	0.276 (0.096)	0.193 (0.101)	0.107 (0.096)	−0.083 (0.096)
	15	0.247 (0.064)	0.064 (0.163)	0.069 (0.099)	−0.090 (0.069)
	20	0.228 (0.049)	0.149 (0.050)	0.087 (0.046)	−0.103 (0.049)
	25	0.215 (0.039)	0.142 (0.040)	0.081 (0.034)	−0.099 (0.038)
	30	0.181 (0.034)	0.107 (0.031)	0.056 (0.030)	−0.099 (0.032)
	35	0.165 (0.028)	0.095 (0.030)	0.055 (0.028)	−0.101 (0.027)
	40	0.156 (0.024)	0.085 (0.025)	0.046 (0.023)	−0.103 (0.024)
	45	0.132 (0.021)	0.064 (0.022)	0.040 (0.021)	−0.090 (0.021)
	50	0.126 (0.019)	0.051 (0.022)	0.030 (0.019)	−0.090 (0.019)
	55	0.113 (0.017)	0.045 (0.018)	0.022 (0.016)	−0.095 (0.016)
	60	0.101 (0.015)	0.029 (0.017)	0.017 (0.016)	−0.090 (0.015)
Fox	10	0.188 (0.106)	0.138 (0.105)	0.077 (0.103)	−0.034 (0.110)
	15	0.184 (0.068)	0.120 (0.069)	0.104 (0.062)	−0.046 (0.069)
	20	0.197 (0.044)	0.121 (0.048)	0.099 (0.047)	−0.045 (0.050)
	25	0.202 (0.035)	0.129 (0.039)	0.094 (0.039)	−0.033 (0.040)
	30	0.202 (0.029)	0.131 (0.029)	0.098 (0.027)	−0.026 (0.035)
	35	0.203 (0.024)	0.122 (0.025)	0.096 (0.027)	−0.038 (0.029)
	40	0.201 (0.020)	0.126 (0.021)	0.097 (0.022)	−0.013 (0.025)
	45	0.205 (0.017)	0.123 (0.017)	0.097 (0.017)	−0.037 (0.022)
	50	0.202 (0.017)	0.123 (0.018)	0.098 (0.018)	−0.034 (0.020)
	55	0.207 (0.014)	0.128 (0.016)	0.099 (0.016)	−0.042 (0.018)
	60	0.202 (0.015)	0.124 (0.015)	0.089 (0.015)	−0.041 (0.017)
	65	0.204 (0.013)	0.123 (0.013)	0.098 (0.013)	−0.041 (0.015)
	70	0.203 (0.011)	0.128 (0.012)	0.095 (0.012)	−0.038 (0.015)
	75	0.199 (0.011)	0.119 (0.012)	0.097 (0.011)	−0.047 (0.013)
Mungomery	10	0.625 (0.051)	0.525 (0.069)	0.566 (0.063)	0.000 (0.112)
	15	0.635 (0.033)	0.530 (0.043)	0.571 (0.040)	0.003 (0.078)
	20	0.631 (0.025)	0.536 (0.031)	0.569 (0.029)	−0.024 (0.053)

Table 6.6: Mean correlation coefficients, over 1000 runs, of the actual deleted values and imputed values found using the four imputation methods. The four data sets used have been standardized within environments. The variance of each coefficient is given in parentheses.

set, while the first data set had slightly better results for the raw data. A homogeneity of variance test for these data sets showed that the first did not need standardizing (Levene test of 1.182, p-value of 0.303), whereas the second should have been transformed (Levene test of 3.816, p-value of 0.000). It was difficult to gauge the effectiveness of transformation of the data sets using the correlations presented in Tables 6.5 and 6.6 because of the high correlation the imputed values had with the environment means. The reduction in correlations when data are standardized was greater for the second data set than it was for the first Flores *et al.* (1998) data set.

No discernible trends with the G×E interaction were identified in any of the results. This was contrary to the initial belief that a method which separates the interaction component would increase in effectiveness as the relative size of the G×E interaction increases. Identification of other reasons for the success of the two-stage method would require further simulation testing with more data sets and possibly greater amounts of missing data.

6.5 Application of two-stage imputation to the trials programme data

The imputation methodology presented in this chapter is now applied to the principal data introduced in Chapter 3. The two subsets of the full data will be imputed using the process described in Section 6.2 above.

Two-stage imputation uses the output from first stage clustering to find groups of genotypes that have similar G×E interaction profiles across environments. Recall from Section 5.4 that there are 24 and 22 first stage clusters in Onion Data I and II respectively; this can be seen in Figures 5.8 and 5.9 on pages 134 and 136.

It was interesting to note that only 34.0% and 35.9% of missing entries were imputed using first stage cluster data for Onion Data I and II respectively. The remaining 6434 and 4572 entries were found using data from genotypes that did have yields in the environment where the missing entry lay, and which first clustered with the genotype whose yield was missing at a level beyond the truncation level. The quality of these imputed values could not be assured, as two-stage imputation allows for yields to be imputed using data from genotypes that are not in the same first stage cluster, and may therefore, not have truly similar interaction profiles.

The last step of two-stage imputation is to replace unrealistic imputed values with a yield that is within the observed range of yields for each environment. This 'trimming' altered 12.0% and 12.5% of imputed values in Onion Data I and II respectively. Table 6.7 shows in detail how many imputed yields were outside the observed range of yields for environments. Histograms of range standardized imputed yields for Onion Data I and II are presented in Figure 6.5, which clearly shows the impact of the trimming of unrealistic

Data	Number of imputations	Number trimmed	Per cent trimmed	Trimmed up	Trimmed down
Onion Data I	9753	1168	12.0	611	557
Onion Data II	7135	891	12.5	421	470

Table 6.7: Details of the number of imputed values that were trimmed during the final step of two-stage imputation for Onion Data I and II. This step ensures that imputed yields that fall outside the observed range of yields from an environment are replaced by the environment maximum or minimum as required.

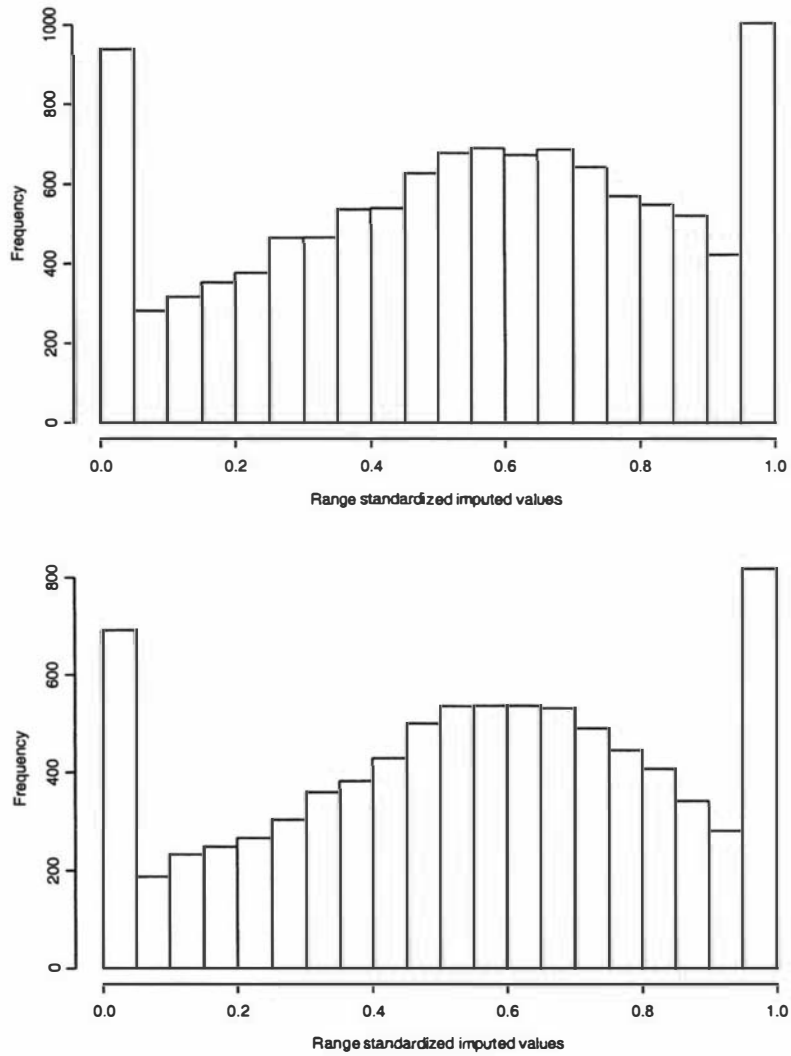


Figure 6.5: Histograms of range standardized imputed yields for Onion Data I (top) and Onion Data II (bottom). Values have been found by scaling all imputed yields within each environment to the range from zero (for the observed minimum) to one (for the observed maximum).

yields on the distribution of final imputed yields. In these histograms, imputed yields are transformed so that the observed minimum and maximum yields for an environment are given a score of zero and one respectively, while all other yields are scaled to this range.

Further investigation into the large number of trimmed imputed values showed that over half of them were originally estimated using data from outside the first stage cluster. For Onion Data I, 191 and 243 values were trimmed up to the observed minimum, and down to the observed maximum respectively; the corresponding numbers for Onion Data II are 126 and 190. The use of data from genotypes outside first stage clusters should therefore be questioned. Results need to be sought using imputed values from only genotypes within the same first stage cluster, as well as from all genotypes.

A disadvantage of the trimming became evident when the number of trimmed values in each environment was checked. On occasion a substantial number of imputed values were trimmed down to the maximum observed value. For example, trial X04601 from Cameroon has seven varieties in Onion Data I, and the fully imputed data for this environment had the same value for 43 varieties. This was clearly a very large number of varieties to recommend for future testing. Testing subsets of these 43 varieties in future trials may provide greater differentiation of their estimated performance in environment X04601, because addition of more data will improve the imputed values.

A second concern with the trimming process was also evident from this environment. Because two of the test varieties in X04601, were not grown in enough environments, their data were omitted from Onion Data I. One of these was the variety that gave the highest yield in the trial, and the trimmed imputed values were therefore not equal to the maximum observed yield from that environment. In situations where extra information is available it could be incorporated into results.

The correlation coefficient for the imputed values from Onion Data II with the same G×E combinations in Onion Data I was 0.695. Figure 6.6 shows that correlation of results within environments was generally high, especially when greater numbers of genotypes were tested at each location. A more in-depth investigation of this will be carried out in Chapter 8, which will consider the dependence of genotype inter-relationships on the data used for imputing missing yields.

The genotypes of Onion Data I and II were re-clustered to see what impact two-stage imputation had on first stage clustering. These dendrograms appear in Figures 6.7 and 6.8, but it is very difficult to compare these with the dendrograms that appear in Figures 5.8 and 5.9 for sparse yield data. While a detailed examination of the differences between these pairs of figures is left until Chapter 8, the most noticeable change is that there are now six fewer first stage genotype clusters in both Onion Data I and II.

One explanation for the reduced number of clusters can be given by considering a set of genotypes that would have similar G×E interaction profiles if they had been tested in all environments. If they were grown in two distinct sets of environments they would

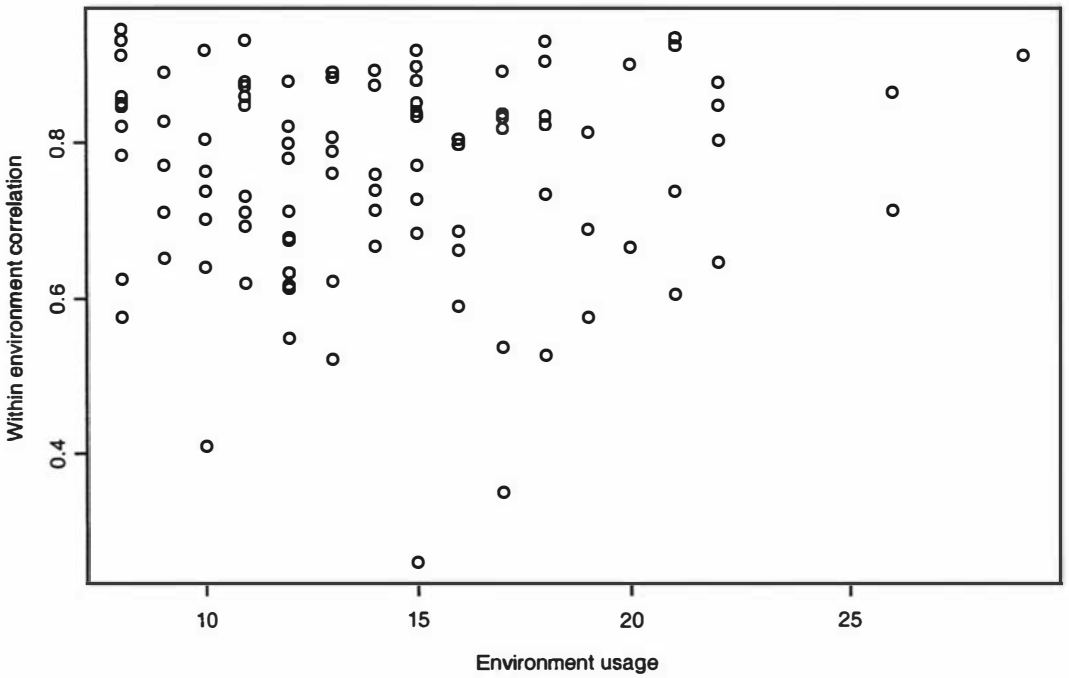


Figure 6.6: Within environment correlation of fully imputed Onion Data I and II plotted against the number of genotypes tested in each environment.

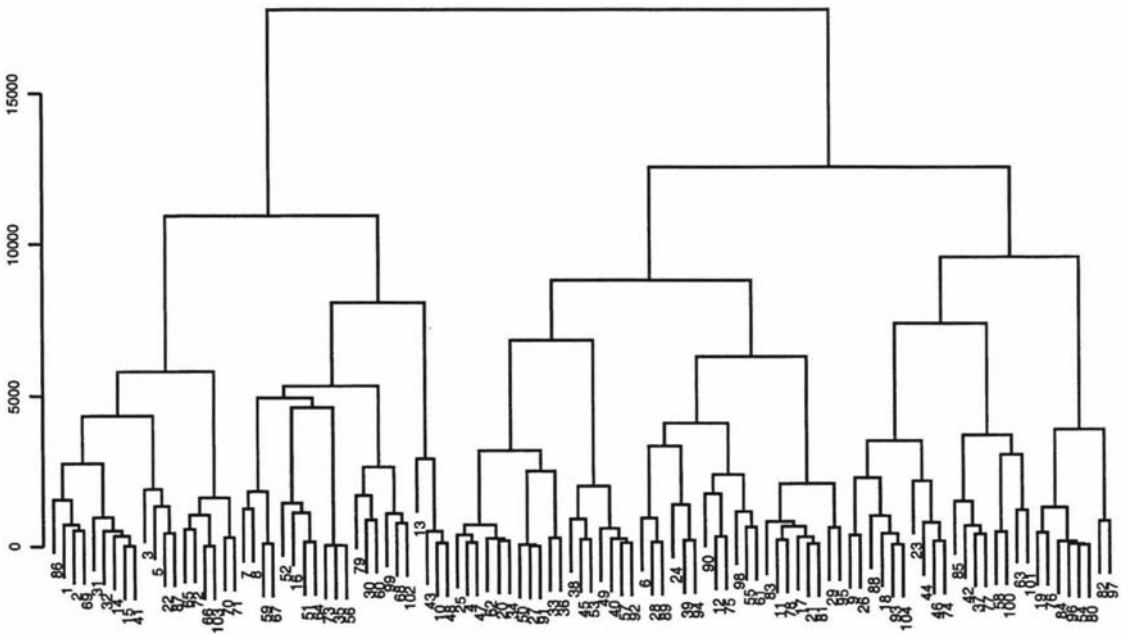


Figure 6.7: Dendrogram of the genotypes of Onion Data I. Interaction distance was applied to the data after two-stage imputation. Clustering was truncated at the level 3505.26, forming 18 first stage clusters.

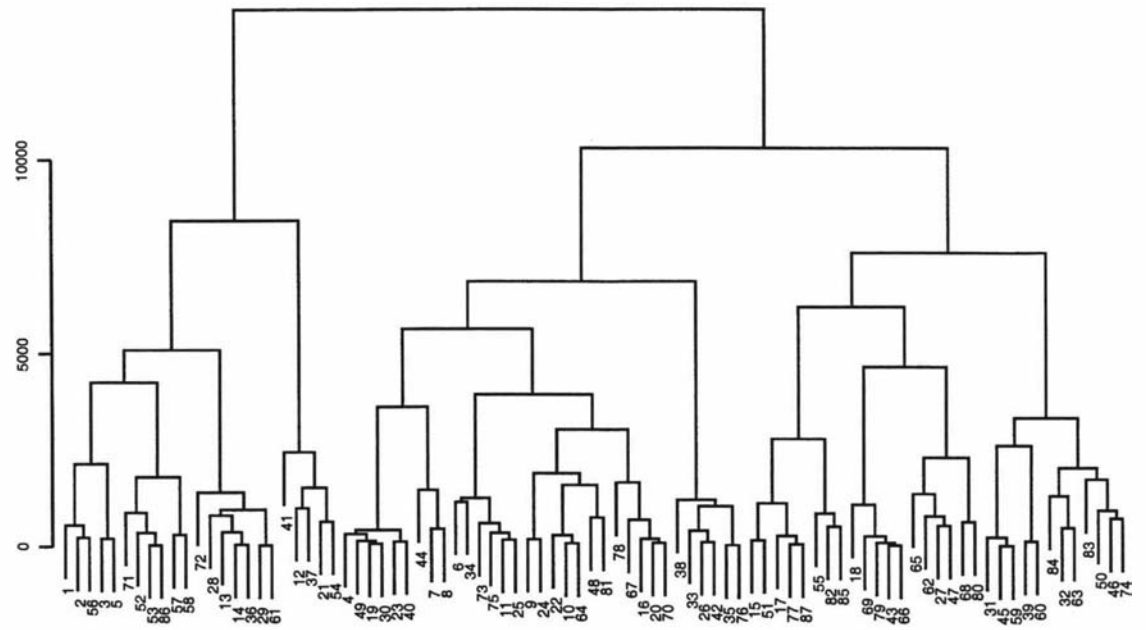


Figure 6.8: Dendrogram of the genotypes of Onion Data II. Interaction distance was applied to the data after two-stage imputation. Clustering was truncated at the level 2598, forming 16 first stage clusters.

be grouped into two first stage clusters. These genotypes would cluster together, using a distance based on imputed and observed yields, if two-stage imputation reconstituted the missing data accurately. Recall that when data is sparse, some distances are estimated using the method described in Section 4.6, which is likely to result in estimated distances that are greater than observed distances. Thus, if two-stage imputation was accurately estimating untested G×E combinations the estimated distances between genotypes would be improved. If this was so, and data were missing at random, the new clustering using the imputed data could confidently be assumed to be the same as the clustering that would result if complete data were available.

This assertion is not restricted to two-stage imputation. It was also tested on the nearest cluster imputation method proposed by Drake (1981). Figures 6.9 and 6.10 show the clustering of Onion Data I and II genotypes using Euclidean distance. These dendrograms were constructed using the same data as was used for Figures 5.8 and 5.9, for comparability. There were 26 and 20 clusters of genotypes in Onion Data I and II respectively when sparse data were clustered. When the nearest cluster imputed data was used to cluster Onion Data I and II genotypes there were 14 and 12 clusters respectively (as seen in Figures 6.11 and 6.12) for these data sets. Once again, these comparisons are given a more detailed inspection in Chapter 8.

Cross validation is one technique that could be employed to gauge the stability of imputed values. 'Stability' is used to mean how robust the imputed values are to the presence or absence of other yields. Given the high number of genotype clusters in the two

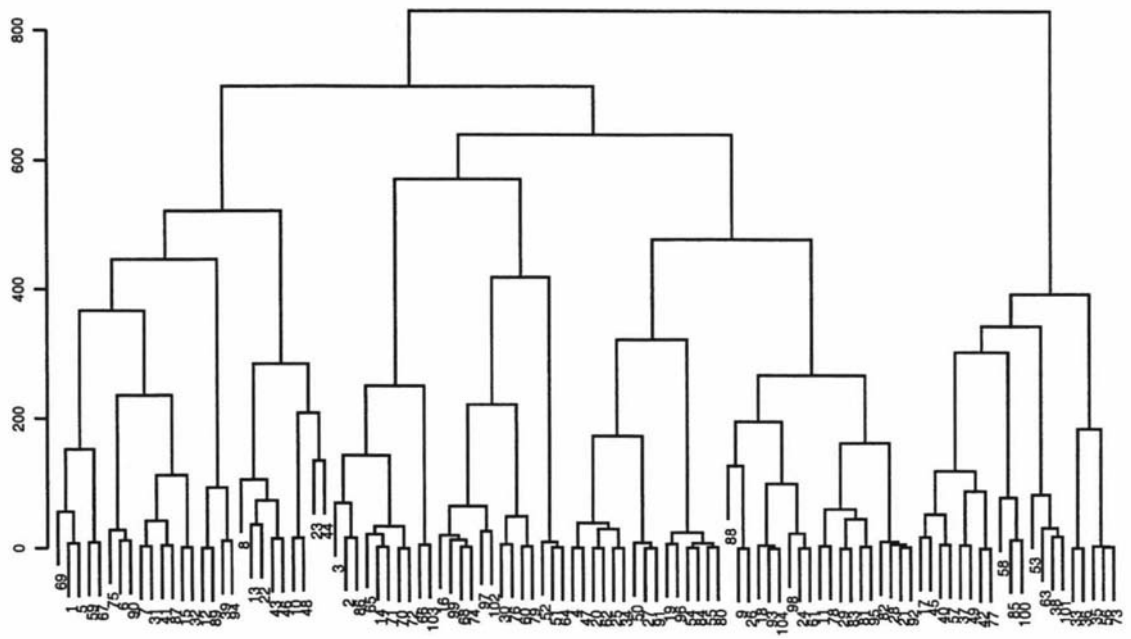


Figure 6.9: Dendrogram of the genotypes of Onion Data I, clustered using Euclidean distance applied to the sparse data. Clustering was truncated at the level 135.36, forming 26 clusters.

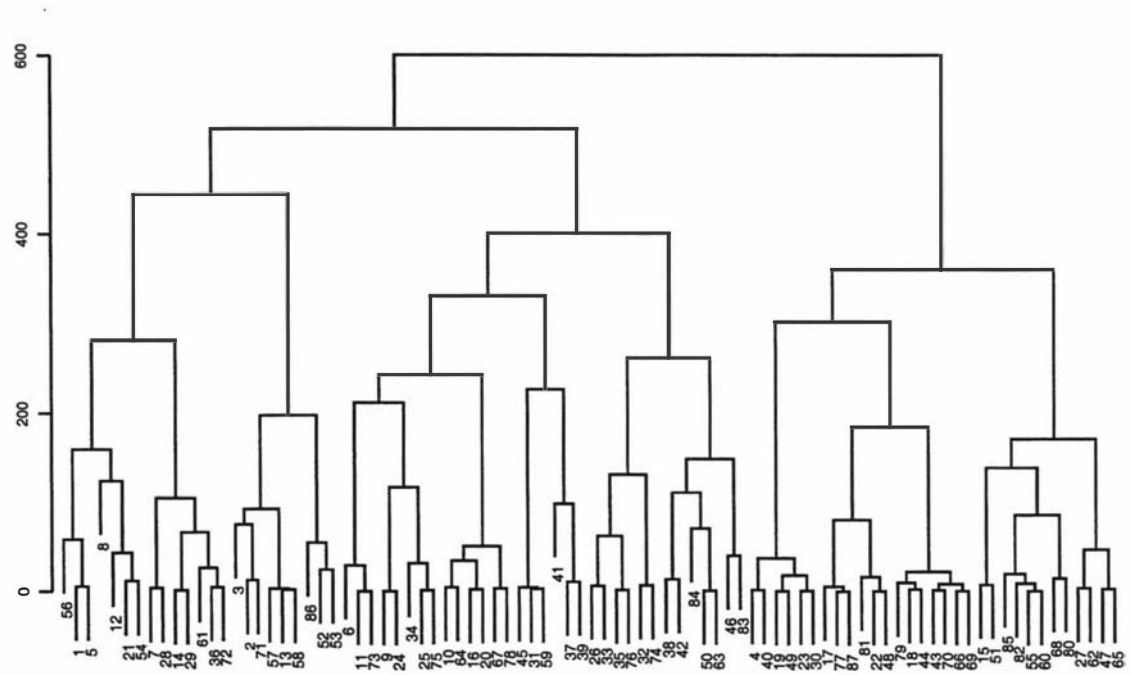


Figure 6.10: Dendrogram of the genotypes of Onion Data II, clustered using Euclidean distance applied to the sparse data. Clustering was truncated at the level 124.04, forming 20 clusters.

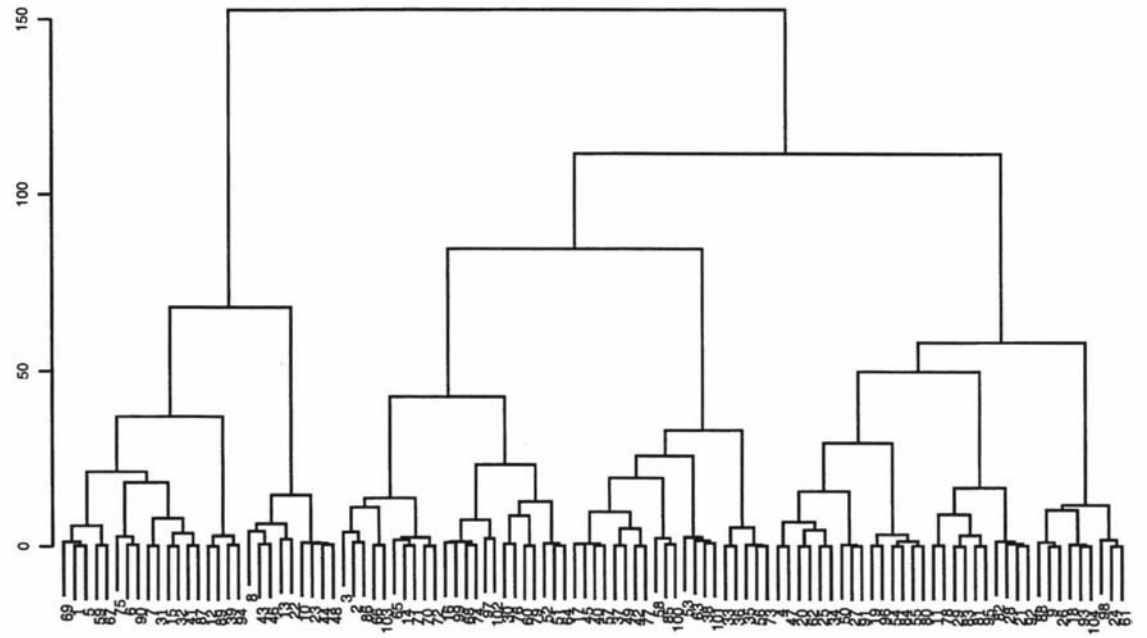


Figure 6.11: Dendrogram of the genotypes of Onion Data I, clustered using Euclidean distance. Fully imputed data was created using the nearest cluster method. Clustering was truncated at the level 19.45, forming 14 clusters.

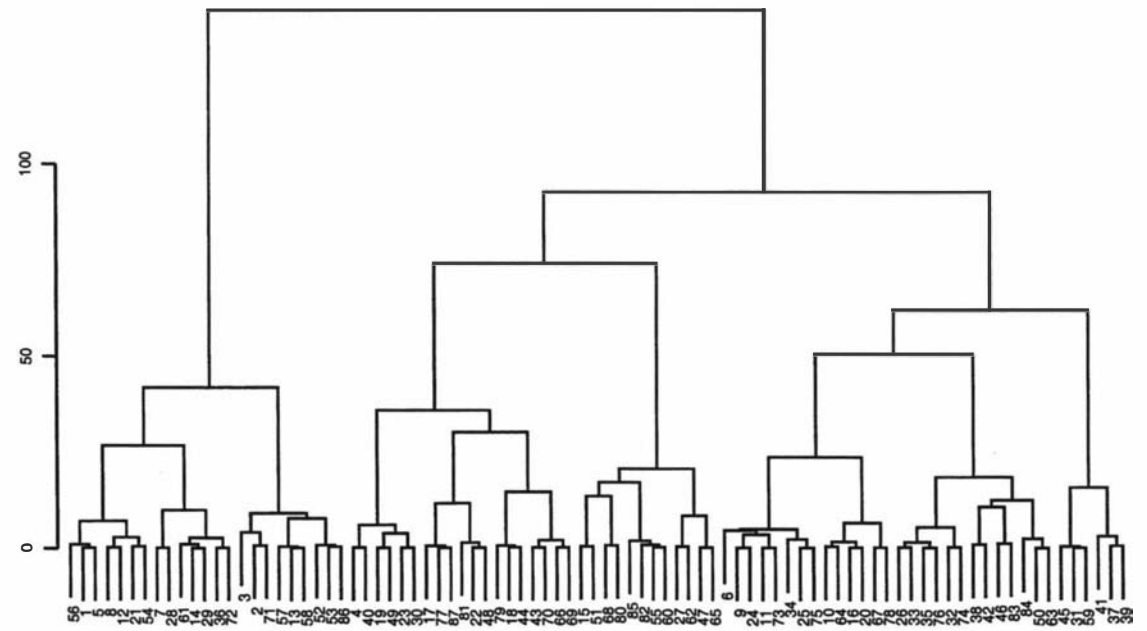


Figure 6.12: Dendrogram of the genotypes of Onion Data II, clustered using Euclidean distance. Fully imputed data was created using the nearest cluster method. Clustering was truncated at the level 18.55, forming 12 clusters.

data sets, compared to the number of observations, it was likely that two-stage imputed values would be unstable because first stage clusters were formed using so few observations. Clustering of genotypes using the sparse data resulted in clusters that had few genotypes in them irrespective of the distance measure used for clustering. It is therefore expected that neither method has a clear stability advantage.

6.6 Further ideas for two-stage imputation

This section highlights future avenues for investigation that have arisen through the development of two-stage imputation.

Two-stage imputation has been presented using genotypes as observations. There is no reason why the transpose of the G×E matrix could not be used in order to impute missing G×E combinations. There is also the possibility that genotype based imputed values could be averaged with those based on environment similarity. In either case, the observed minimum and maximum yields from each environment would still need to be used in the trimming stage.

In situations where no data is available for a certain environment within a first stage cluster of genotypes, two-stage imputation looks for the cluster that is most similar which has data in that environment. There is some potential for imputed values in this instance to be based on dubious similarity, especially when data is extremely sparse. If only first stage clusters were used to find imputed values, many G×E observations in Onion Data I and II would not have been imputed. Imputations based on first stage clustering could be used as the input for a second round of imputation based on a new first stage clustering. This process could be repeated until all missing G×E combinations were imputed, but would be unsuccessful in the event that a single genotype remained alone in a first stage cluster.

No cross validation approaches have been investigated at this point. Ideally some G×E combinations could be omitted from the data and the subsequent imputed values compared. The correlation between the common G×E combinations of Onion Data I and II in the previous section is an example of what is possible. The sparsity of Onion Data I and II made this task less possible, because some G×E combinations are crucial for maintaining the ability to link data from trials together. Such an exercise could be used to provide an interval estimate for imputed values rather than the current point estimates. Note that this interval estimation of imputed values could not be undertaken for all missing entries in Onion Data I and II if the minimum data constraints were to remain intact.

Another means of establishing interval estimates for imputed values would be to use the two-stage clustering model given in (5.6) to impute missing values as part of a multiple imputation strategy. The process presented in Section 6.1 can now be enhanced to include

the G×E interaction effects, determined by first stage clustering, to give:

1. Perform first stage clustering.
2. Create an explanatory indicator variable $G_{f(i)}$ to indicate first stage cluster membership.
3. Fit the model
$$Y_{ik} = \mu + G_i + GE_{f(i)k} + E_k + \epsilon_{ik} \tag{6.2}$$
to the data. The interaction term uses the term for genotype cluster membership, not the term for genotype main effect.
4. Use the parameter values from this model to determine the expected values of the missing entries.
5. Use the error MS from this model as the variance of the distributions used to impute missing entries.

If a first stage cluster has no data in an environment, there will be no data to estimate parameter values for the corresponding $GE_{f(i)k}$ terms. These parameters could be estimated using their expected value (zero), but will limit the usefulness of the approach in choosing the best genotypes to grow in each type of environment.

6.7 Summary

This chapter presented an imputation method that arose as a consequence of the two-stage clustering method developed in Chapter 5. The ability to determine sets of genotypes that perform similarly across environments and subsequently to take advantage of differences in their mean performance allows incomplete G×E matrices to be made complete.

Existing imputation methodology was shown to be of limited use when working with G×E data. Model-based imputation strategies were discounted for use with data as sparse as that arising from the Onion Trials Programme. Simulation testing found that the two-stage imputed values were better than those found using other clustering-based imputation strategies. Testing was done using data sets from the G×E literature which were all complete and differed in size.

Application of two-stage imputation to Onion Data I and II gave consistent results, especially for the imputed values for environments in which greater numbers of genotypes were tested.

Some further ideas for the future development of two-stage imputation were introduced in Section 6.6. Regardless of the imputation method employed, there is a need to link variety selections to the different types of environment within the data. In Chapter 7 these groups of environments will be found so that genotype success in new environments can be predicted.

Chapter 7

Determining mega-environments

7.1 Introduction

An international trials programme covers an extensive range of environments. To gauge specific adaptation to certain types of environments it is necessary to think about how to define these types, or 'mega-environments' as they will now be called. Mega-environments, as defined by Gauch and Zobel (1997), are groups of environments which have similar characteristics in terms of genotypic performance, environmental factors, geographic co-location, and even economic conditions. They included the following in the benefits of establishing mega-environments:

1. Improved resource allocation in a research programme.
2. Rationalization of germplasm and information exchanges between breeding programmes.
3. Increased efficiency of breeding programmes.
4. Targeting appropriate genotypes to environments.

The second of these points is of particular importance in this work as it will, according to Gauch and Zobel (1997), allow "even small programmes to progress by focusing on the most promising material".

Establishing mega-environments is a cornerstone of the entire project as it will allow the transfer of results from the trials programme to new environments. The ability to bring the theoretical results from Chapters 4 to 6 together so that they can be used in the future to guide researchers is paramount. A means of gauging success in this venture must, therefore, be developed to validate results. The comparison of results is, however, left to Chapter 8 as the tools are as yet undeveloped.

In the meantime, various strategies for finding these mega-environments, need to be considered for this investigation, including:

1. Cluster analysis using available yield data,
2. Cluster analysis using imputed yield data,
3. Cluster analysis using covariate information,
4. AMMI used in conjunction with the EM algorithm (Gauch and Zobel, 1990), and
5. Discriminant analysis using covariate information.

We could also use, given complete data:

6. Results from AMMI or other principal component modelling,
7. Discriminant analysis using yield data, or
8. Cluster analysis using a mixture of yield data and covariate information.

Theory developed in Chapters 4 and 5 allows the clustering of environments using available yield data, and is presented in Section 7.2. The fully imputed yield data, found in Section 6.5, can also be used to form mega-environments, and will be used in Section 7.3.

It has been argued that observed yields, or more specifically, the deviations from environment mean performance, are indicative of the expected effects of covariate factors occurring at an environment (Abou-El-Fittouh *et al.*, 1969; Ivory *et al.*, 1991). Clustering using covariates was therefore investigated using data from the 101 environments for which Covariate information was available. Section 7.4 shows how a consistent set of covariates was found and used to create mega-environments.

The use of model-based imputation was questioned in Section 6.2. These concerns hold when any data is missing, and the amount of sparsity in the Onion Trial Data exacerbated them. Applications of the EM algorithm will not therefore be used in this chapter.

Discriminant analysis can be used to determine group membership for a set of observations. Its implementation is well suited to situations where the groups are easily defined. It can also be used as an extension of cluster analysis, where the groups have been found. The benefit of discriminant analysis comes when a new observation is added to the existing set, and can be assigned membership to an existing group on the basis of probability of belonging; or in this context a new environment is given a probability of being in each mega-environment. If this is the primary motivator for the use of discriminant analysis, then use of such modern clustering ideas as 'fuzzy clustering' would be able to give the same information

The use of principal component models (such as AMMI) on the fully imputed data was deemed unnecessary at this point. These models provide the means by which the varieties that should be grown in each tested environment can be determined. Extrapolation to new environments will still rely on the ability to first identify mega-environments, and then to decide to which mega-environment the new environment belongs.

The above list is not meant to be exhaustive. Cluster analysis methods are well-known to researchers in the G×E discipline. Their application and limitations are understood. They provide a simple analysis that is visually interpretable through the creation of dendrograms. This practicality is the prime motivation for their use in creating mega-environments in this chapter. The creation of mega-environments using available yield data, imputed data, and covariate information is now considered. Comparisons between these methods are left until Chapter 8.

7.2 Use of available yield data to cluster environments

The theory developed in Chapters 4 and 5 is now used to cluster environments on their G×E interaction similarity. Interaction distance I_{ij} was applied to the incomplete yield data of three sets of environments:

1. The environments from the entire trials programme,
2. Those in Onion Data I , and
3. Those in Onion Data II.

The cluster analyses presented in this section are complementary to the first stage cluster analyses for genotypes given in Section 5.4. As with that analysis, the impact that variables would have on the outcome was considered before clustering observations.

Heterogeneity of genotype variance was presented in Figure 3.1, which showed the mean and standard deviation of the untransformed genotype performances. Specifically, it showed that the within-genotype variability was extremely wide ranging; this can be partially attributed to the selection of environments for testing each genotype. Using the untransformed data to cluster environments would give results that were influenced by the selection of genotypes for environments. On the other hand, if the selections were completely random, use of the untransformed data would be more appropriate because the genotypes that have higher G×E interaction would have greater influence on the clustering of environments. For this reason, transformation of the data to counter the impact of genotype selection was clearly necessary. Within-genotype standardization was considered first.

Histograms of the available yields (after within-genotype standardization) for the three sets of interest are shown in Figure 7.1. The shape of these distributions gave concern. The heavy positive skewness meant that there were many values that reflect mediocre performances. The similarity of environments should be independent of their mean performance. If positively skewed data is used, pairs of low performing environments would cluster together sooner than pairs of high performing environments, even though the relevant qualitative G×E interaction structures were effectively the same. With complete data, this problem could be overcome by using ranks within genotypes. Transformation

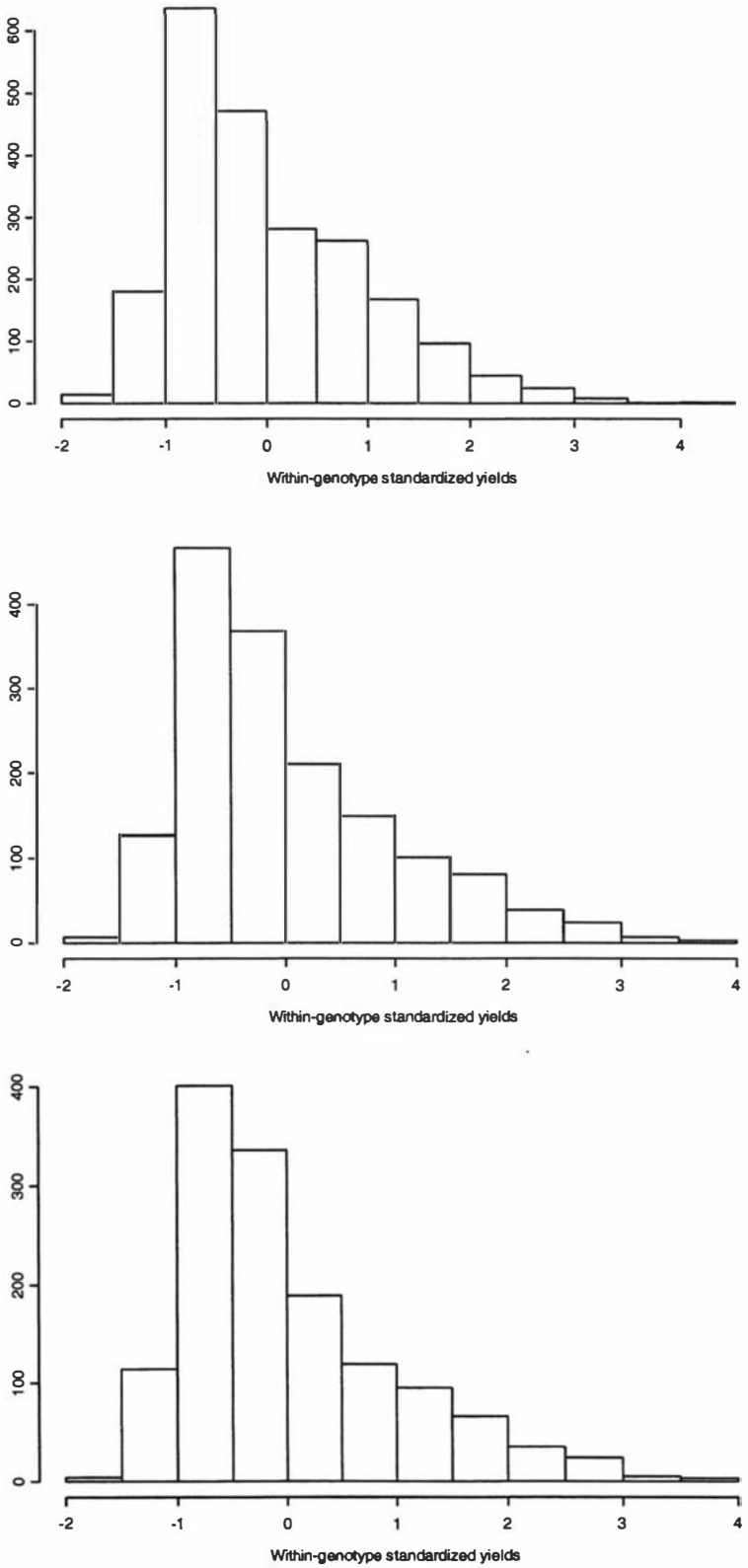


Figure 7.1: Histograms for within-genotype standardized yields for the three data sets to be clustered. The entire data from the Onion Trials Programme (top), are presented along with the data from the two reduced data sets - Onion Data I (middle) and Onion Data II (bottom).

to ranks would introduce other problems when working with incomplete data, so a search for a more symmetric distribution was preferred. In Chapter 3 the need to use the square roots of yields to remove the problem of increasing variance with increasing mean was established; this transformation also improved the symmetry of the data. The option to standardize these square roots within genotypes was investigated and, as shown in Figure 7.2, gave a better distribution of values to use for clustering environments. In Section 5.4, this transformation was applied to environments for clustering genotypes.

Figure 7.3 shows the dendrogram created using all 123 environments of the trials programme, with the incremental sum of squares method and interaction distance I_{ij} developed in Section 4. The square root yield data were standardized within genotypes to ensure that each genotype contributed to the distance measures evenly, in a complementary way to the clustering of genotypes discussed in Chapter 5. The stopping criterion introduced in Section 5.2 shows that clustering should be truncated at the level 1285, where thirty clusters of environments had been determined. The memberships of these clusters are shown in Table 7.1.

Figure 7.4 shows the dendrogram created using the 109 environments which formed Onion Data I , with the same distance measure, clustering method, and data transformations as described previously in this section. The stopping criterion shows that clustering should be truncated at the level 536, which created 29 clusters of environments. These cluster memberships are presented in Table 7.2.

Figure 7.5 shows the dendrogram created using the 98 environments of Onion Data II , the incremental sum of squares method and interaction distance I_{ij} as described previously. The stopping criterion shows that clustering environments of Onion Data II should be truncated at the level 491, which created 25 mega-environments. Table 7.3 shows the cluster memberships in more detail.

The number of different environments that are being clustered has an impact on more than just the cluster membership. The stopping criterion truncated the clustering at different heights in Figures 7.3, 7.4, and 7.5 with the clusters formed in each dendrogram differing in size. The truncation height indicated the within-cluster sum of squared distances of the last cluster to be formed in each dendrogram. The largest cluster in Figure 7.3 was more than twice the size of the largest cluster in either of Figures 7.4 or 7.5. This can be explained by the need to impute a greater number of distances using the methodology of Section 4.6, which adds existing distances together to give imputed distances. These imputed distances were generally greater than existing distances, and as the number of missing distances increased, the imputed distances had a greater impact on the truncation level.

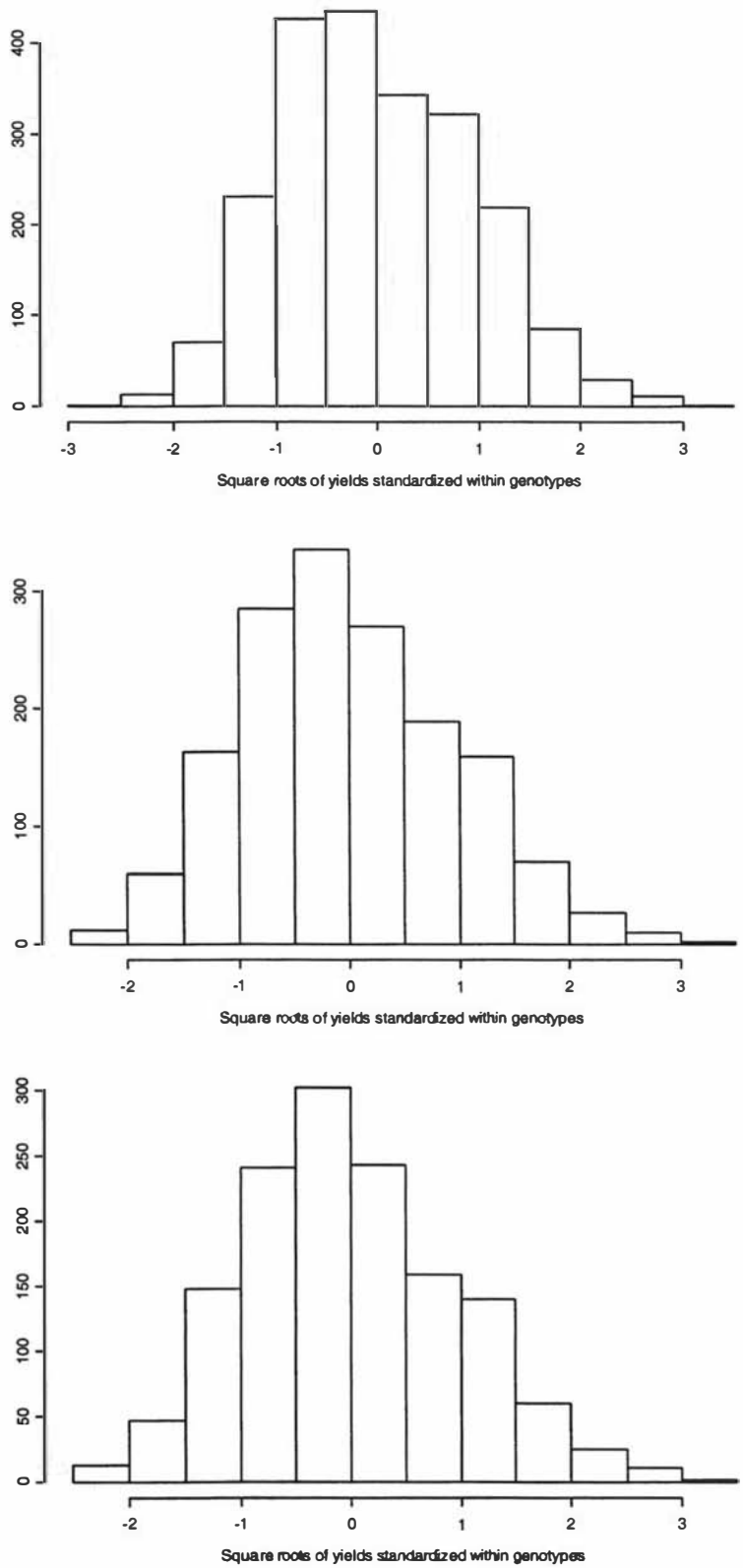


Figure 7.2: Histograms for within-genotype standardized square roots of yields for the three data sets to be clustered. The entire data from the Onion Trials Programme (top), are presented along with the data from the two reduced data sets - Onion Data I (middle) and Onion Data II (bottom).

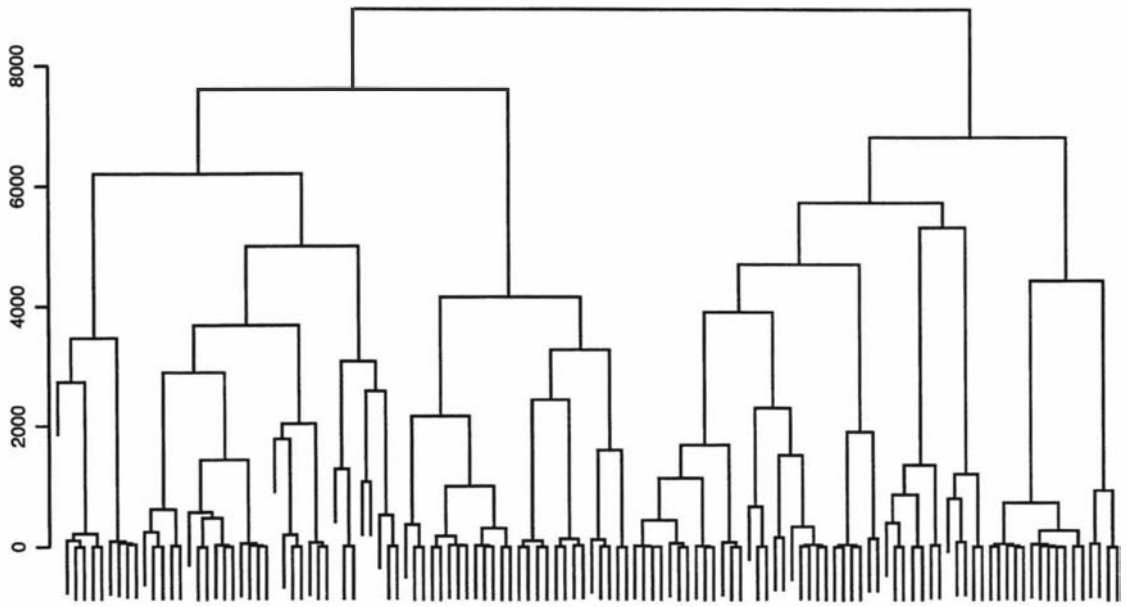


Figure 7.3: Dendrogram of all 123 environments of the Onion Trials Programme. Interaction distance and incremental sum of squares clustering have been applied to the within-genotype standardized square roots of available yields. The stopping criterion created 30 mega-environments at the level 1285.

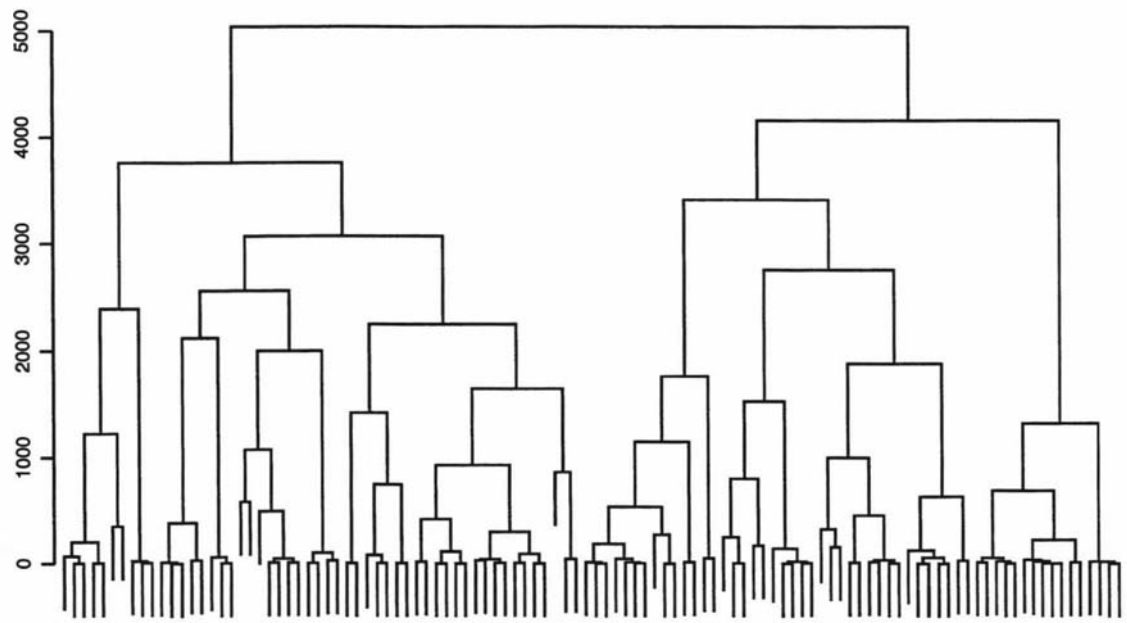


Figure 7.4: Dendrogram of the 109 environments which form Onion Data I . Interaction distance I_{ij} and the incremental sum of squares method have been applied to the within-genotype standardized square roots of available yields. The stopping criterion stopped clustering at the level 536 and created 29 mega-environments.

Cluster	Members
1	D02501, Egypt, 1996; Y13102, St Helena, 1997; Z03901, Ghana , 1997
2	A01604, Yemen, 1993; A05902, Fiji, 1991; C04402, PNG, 1993; J08801, Mauritius, 1998; O02707, Sri Lanka, 1994
3	A00501, Nigeria, 1991; A04102, Bangladesh, 1991; A04103, Bangladesh, 1992; L03701, Lesotho, 1992; O02700, Sri Lanka, 1991; O02701, Sri Lanka, 1992; W22801, Italy, 1990; X07201, Nepal, 1993; X15107, Guinee, 1995; Z03402, Pakistan, 1995
4	J08001, Benin, 1998; Z12201, Malawi, 1996
5	A04001, India, 1993; A05901, Fiji, 1991; F02405, Botswana, 1993; F02406, Botswana, 1993; F02701, Tanzania, 1992; F04006, Uganda, 1993; O07601, Mozambique, 1991; X10801, P R China , 1998; X15105, Guinee, 1995; Y01401, Uruguay, 1993
6	W17501, Greece, 1993; W17502, Greece, 1993; W17505, Greece, 1995 W17506, Greece, 1995
7	A01601, Yemen, 1992; A01603, Yemen, 1993; A01606, Yemen, 1994; J00301, India, 1994; W23001, Ethiopia, 1993
8	D03602, Zimbabwe, 1990; P07200, Nepal, 1990; P07201, Nepal, 1991; P07202, Nepal, 1991; P07203, Nepal, 1991; P07204, Nepal, 1991; P07205, Nepal, 1991; P07206, Nepal, 1991; W13204, Nepal, 1992; W17503, Greece, 1994; W17504, Greece, 1994; W23003, Ethiopia, 1993
9	F02702, Tanzania, 1994; L01101, Kenya, 1997; L01103, Kenya, 1998
10	A03602, Australia, 1996; O07603, Mozambique, 1997; W16508, Brazil, 1995; X13901, Senegal, 1994; Y07501, Honduras, 1996
11	A04202, Mauritania, 1994; D03605, Zimbabwe, 1991; O07604, Mozambique, 1997; Y01302, Argentina, 1997; Y07403, Taiwan, 1993; Z09603, Korea, 1995
12	R00401, Tunisia, 1999; Y01501, Kenya, 1992; Y13101, St Helena, 1997
13	A03401, Belize, 1990; A04101, Bangladesh, 1990; Z00101, Thailand, 1991
14	R00701, India, 1998; W16503, Brazil, 1993; X05701, Nigeria, 1994; X15101, Guinee, 1993; Z13501, Nigeria, 1997
15	A00301, Barbados, 1995; A03603, Australia, 1997; R00101, Australia, 1996; R00102, Australia, 1997; Y07405, Taiwan, 1996
16	D03610, Zimbabwe, 1999; W23009, Ethiopia, 1997; X15103, Guinee, 1994
17	O11601, Senegal, 1993; Y13901, Nigeria, 1995
18	X04601, Cameroon, 1994; X07202, Nepal, 1994
19	L05101, South Africa, 1993; X10901, New Caledonia, 1994; Z03501, Philippines, 1996; Z09601, Korea, 1994
20	D03601, Zimbabwe, 1989; D03603, Zimbabwe, 1990; D03604, Zimbabwe, 1991

Table 7.1: Members of the 30 mega-environments determined by the clustering presented in Figure 7.3. All 123 environments of the Onion Trials Programme are used in this clustering. *Continued on page 181.*

Cluster	Members
21	C04401, PNG, 1993; F02408, Botswana, 1994; X04804, Zambia, 1993; Y07401, Taiwan, 1992
22	A03101, Malaysia, 1998; X13801, Mali, 1997; Y03404, C Ivoire, 1998; Z13503, Nigeria, 1997
23	O02702, Sri Lanka, 1992; O02709, Sri Lanka, 1994; X15106, Guinee, 1995
24	W05201, Bulgaria, 1995; Y03402, C Ivoire, 1994; Z10501, Cape Verde, 1996
25	F02407, Botswana, 1994; R00702, India, 1998; W16502, Brazil, 1992; W16506, Brazil, 1994
26	W23005, Ethiopia, 1994; W23007, Ethiopia, 1996; Z03401, Pakistan, 1994 Z03903, Ghana , 1998
27	D07101, Burkina, 1993; O02703, Sri Lanka, 1993
28	D08401, Mexico, 1992; F02404, Botswana, 1993
29	F02703, Tanzania, 1996
30	W23008, Ethiopia, 1996

Table 7.1: *Continued from page 180.* Members of the 30 mega-environments determined by the clustering presented in Figure 7.3. All 123 environments of the Onion Trials Programme are used in this clustering.

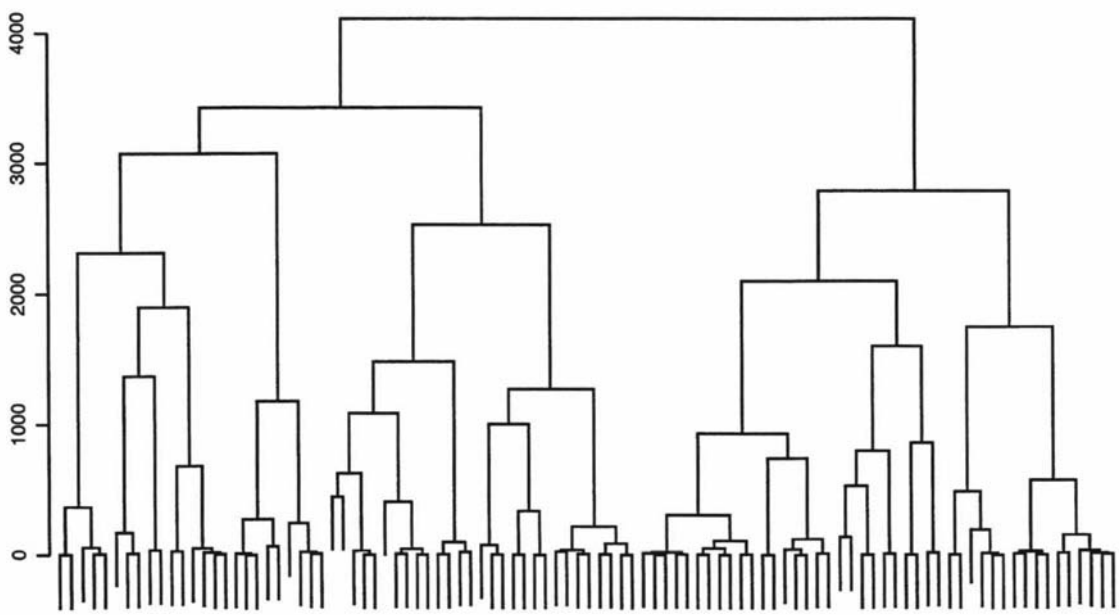


Figure 7.5: Dendrogram of the 98 environments which form Onion Data II , formed using the within-genotype standardized square roots of available yields, interaction distance I_{ij}, and the incremental sum of squares method of forming new clusters. The stopping criterion indicates that clustering should be truncated at the level 491, where 25 mega-environments remain.

Cluster	Members
1	A00501, Nigeria, 1991; A03401, Belize, 1990; A04001, India, 1993; A04101, Bangladesh, 1990; A04102, Bangladesh, 1991; F02701, Tanzania, 1992; O02700, Sri Lanka, 1991; O02701, Sri Lanka, 1992; W23001, Ethiopia, 1993; Z00101, Thailand, 1991
2	R00702, India, 1998; W16502, Brazil, 1992; W16506, Brazil, 1994; W16508, Brazil, 1995; Y07501, Honduras, 1996
3	F02702, Tanzania, 1994; L01101, Kenya, 1997; R00701, India, 1998 W16503, Brazil, 1993; X05701, Nigeria, 1994; X15101, Guinee, 1993
4	D02501, Egypt, 1996; F02406, Botswana, 1993; X15105, Guinee, 1995; Y01401, Uruguay, 1993; Y07403, Taiwan, 1993; Z09603, Korea, 1995
5	A04202, Mauritania, 1994; O07604, Mozambique, 1997; W23009, Ethiopia, 1997; X15103, Guinee, 1994; Y01302, Argentina, 1997
6	A03602, Australia, 1996; X13801, Mali, 1997
7	A01604, Yemen, 1993; O11601, Senegal, 1993; Y13901, Nigeria, 1995
8	D03610, Zimbabwe, 1999; F02405, Botswana, 1993; F02407, Botswana, 1994; F02408, Botswana, 1994; O07601, Mozambique, 1991; X04804, Zambia, 1993; X10801, P R China , 1998; Y07401, Taiwan, 1992
9	A01601, Yemen, 1992; D07101, Burkina, 1993; J00301, India, 1994
10	P07202, Nepal, 1991; P07203, Nepal, 1991; P07204, Nepal, 1991; P07206, Nepal, 1991; W13204, Nepal, 1992; W17504, Greece, 1994; W23003, Ethiopia, 1993
11	A00301, Barbados, 1995; A03603, Australia, 1997; R00101, Australia, 1996; R00102, Australia, 1997; Y07405, Taiwan, 1996
12	X04601, Cameroon, 1994; X07202, Nepal, 1994
13	W23005, Ethiopia, 1994; W23007, Ethiopia, 1996; Z03401, Pakistan, 1994; Z03402, Pakistan, 1995; Z03903, Ghana , 1998
14	A04103, Bangladesh, 1992; O02702, Sri Lanka, 1992; O02709, Sri Lanka, 1994; X15106, Guinee, 1995; X15107, Guinee, 1995
15	L05101, South Africa, 1993; X10901, New Caledonia, 1994; Z03501, Philippines, 1996; Z09601, Korea, 1994
16	D03601, Zimbabwe, 1989; D03603, Zimbabwe, 1990; D03604, Zimbabwe, 1991
17	W05201, Bulgaria, 1995; Y03402, C Ivoire, 1994; Z10501, Cape Verde, 1996
18	D03602, Zimbabwe, 1990; P07200, Nepal, 1990; P07201, Nepal, 1991; P07205, Nepal, 1991; W17503, Greece, 1994
19	W17501, Greece, 1993; W17502, Greece, 1993
20	O07603, Mozambique, 1997; R00401, Tunisia, 1999
21	O02703, Sri Lanka, 1993; X13901, Senegal, 1994;
22	A03101, Malaysia, 1998; Y03404, C Ivoire, 1998; Z13503, Nigeria, 1997
23	L03701, Lesotho, 1992; W22801, Italy, 1990; X07201, Nepal, 1993; Y01501, Kenya, 1992
24	F04006, Uganda, 1993; Z03901, Ghana , 1997
25	A05902, Fiji, 1991; J08801, Mauritius, 1998
26	D08401, Mexico, 1992; F02404, Botswana, 1993
27	D03605, Zimbabwe, 1991
28	Y13101, St Helena, 1997
29	Y13102, St Helena, 1997

Table 7.2: Members of the 29 mega-environments determined by the clustering presented in Figure 7.4. Only the 109 environments of Onion Data I are used in this clustering.

Cluster	Members
1	A01601, Yemen, 1992; D07101, Burkina, 1993; J00301, India, 1994; W22801, Italy, 1990; X07201, Nepal, 1993; Y01501, Kenya, 1992
2	A03602, Australia, 1996; D03605, Zimbabwe, 1991
3	R00702, India, 1998; W16502, Brazil, 1992; W16506, Brazil, 1994; W16508, Brazil, 1995; Y07501, Honduras, 1996
4	A00301, Barbados, 1995; A03603, Australia, 1997; R00101, Australia, 1996; R00102, Australia, 1997; Y07405, Taiwan, 1996
5	D08401, Mexico, 1992; F02404, Botswana, 1993; X15105, Guinee, 1995 Y01401, Uruguay, 1993
6	A00501, Nigeria, 1991; A04001, India, 1993; A04102, Bangladesh, 1991; A04103, Bangladesh, 1992; O02700, Sri Lanka, 1991; O02701, Sri Lanka, 1992; O02702, Sri Lanka, 1992; X05701, Nigeria, 1994; X15101, Guinee, 1993; X15106, Guinee, 1995; X15107, Guinee, 1995
7	A04202, Mauritania, 1994; O02709, Sri Lanka, 1994; O07604, Mozambique, 1997; W23009, Ethiopia, 1997; X15103, Guinee, 1994
8	W23005, Ethiopia, 1994; W23007, Ethiopia, 1996; X07202, Nepal, 1994; Z03401, Pakistan, 1994
9	D03610, Zimbabwe, 1999; F02405, Botswana, 1993; F02407, Botswana, 1994; F02408, Botswana, 1994; O07601, Mozambique, 1991; X04804, Zambia, 1993; X10801, P R China , 1998; Y07401, Taiwan, 1992
10	A03101, Malaysia, 1998; X13801, Mali, 1997; Z13503, Nigeria, 1997
11	L03701, Lesotho, 1992; P07202, Nepal, 1991; P07203, Nepal, 1991; P07204, Nepal, 1991; P07206, Nepal, 1991; W23003, Ethiopia, 1993
12	A01604, Yemen, 1993; Y13901, Nigeria, 1995
13	F02406, Botswana, 1993; F02701, Tanzania, 1992; F04006, Uganda, 1993; W23001, Ethiopia, 1993; Z03903, Ghana , 1998
14	L05101, South Africa, 1993; X10901, New Caledonia, 1994; Z03501, Philippines, 1996; Z09601, Korea, 1994
15	D03601, Zimbabwe, 1989; D03603, Zimbabwe, 1990; D03604, Zimbabwe, 1991
16	R00701, India, 1998; W05201, Bulgaria, 1995; Y03402, C Ivoire, 1994; Z10501, Cape Verde, 1996
17	O07603, Mozambique, 1997; R00401, Tunisia, 1999
18	Y01302, Argentina, 1997; Y07403, Taiwan, 1993; Z09603, Korea, 1995
19	D02501, Egypt, 1996; Z03402, Pakistan, 1995
20	D03602, Zimbabwe, 1990; P07201, Nepal, 1991; P07205, Nepal, 1991; W17503, Greece, 1994
21	O02703, Sri Lanka, 1993; X13901, Senegal, 1994
22	F02702, Tanzania, 1994; L01101, Kenya, 1997
23	A05902, Fiji, 1991; J08801, Mauritius, 1998
24	O11601, Senegal, 1993; W17504, Greece, 1994
25	A03401, Belize, 1990; A04101, Bangladesh, 1990

Table 7.3: Members of the 25 mega-environments determined by the clustering presented in Figure 7.5. Only the 98 environments of Onion Data II are used in this clustering.

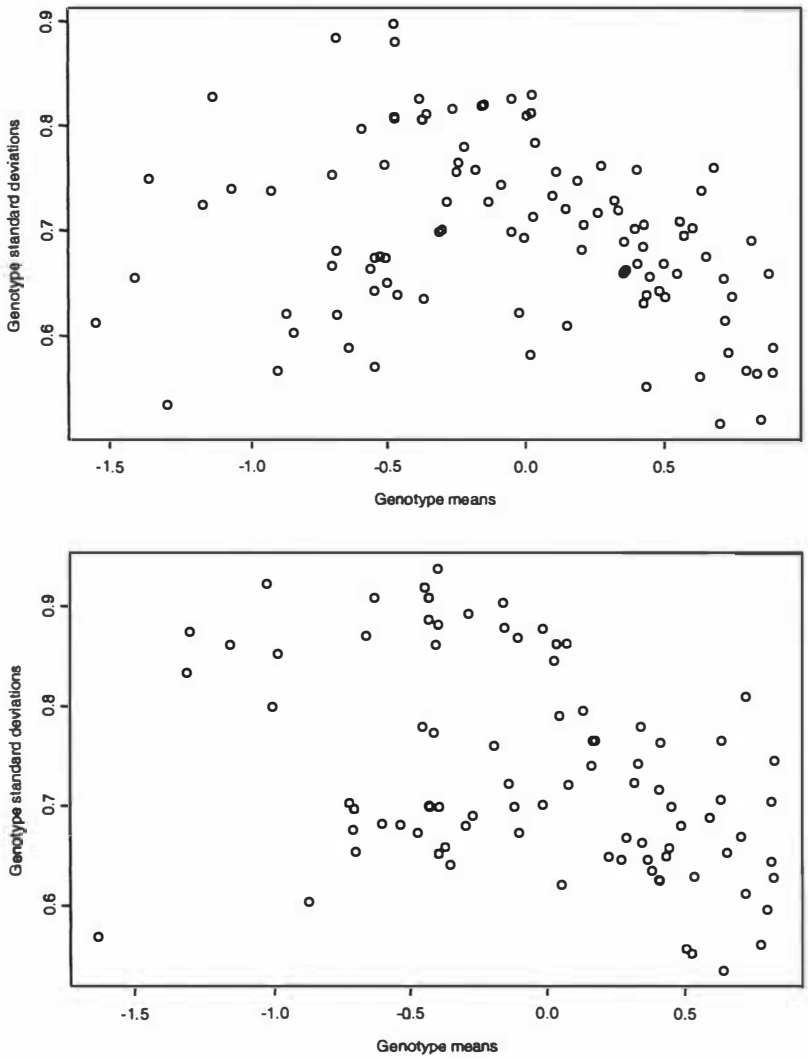


Figure 7.6: Genotype standard deviations plotted against genotype means for Onion Data I and II (top and bottom respectively), after full imputation via two-stage imputation in Section 6.5.

7.3 Use of fully imputed yield data to cluster environments

In this section the fully imputed matrices found in Section 6.5 are used to cluster environments. The output matrices from Section 6.5 were yields in terms of environmentally standardized performances. As in the previous section, differences in the within-genotype variation need to be considered. Figure 7.6 shows the within-genotype means and standard deviations plotted against one another for the imputed G×E matrices of Onion Data I and II. The within-genotype standard deviations differed markedly, so standardization of the within-genotype data was undertaken for both data sets.

Within-genotype standardization for these G×E matrices resulted in doubly standardized entries. Figure 7.7 shows the effects on the within-environment means and standard deviations for Onion Data I and II, of the within-genotype standardization. A new problem arose as a consequence of this second standardization. Using row and column transformations cannot provide homoscedasticity over both rows and columns simultaneously. If row (genotype) homoscedasticity is obtained, equal weighting of the data from each genotype

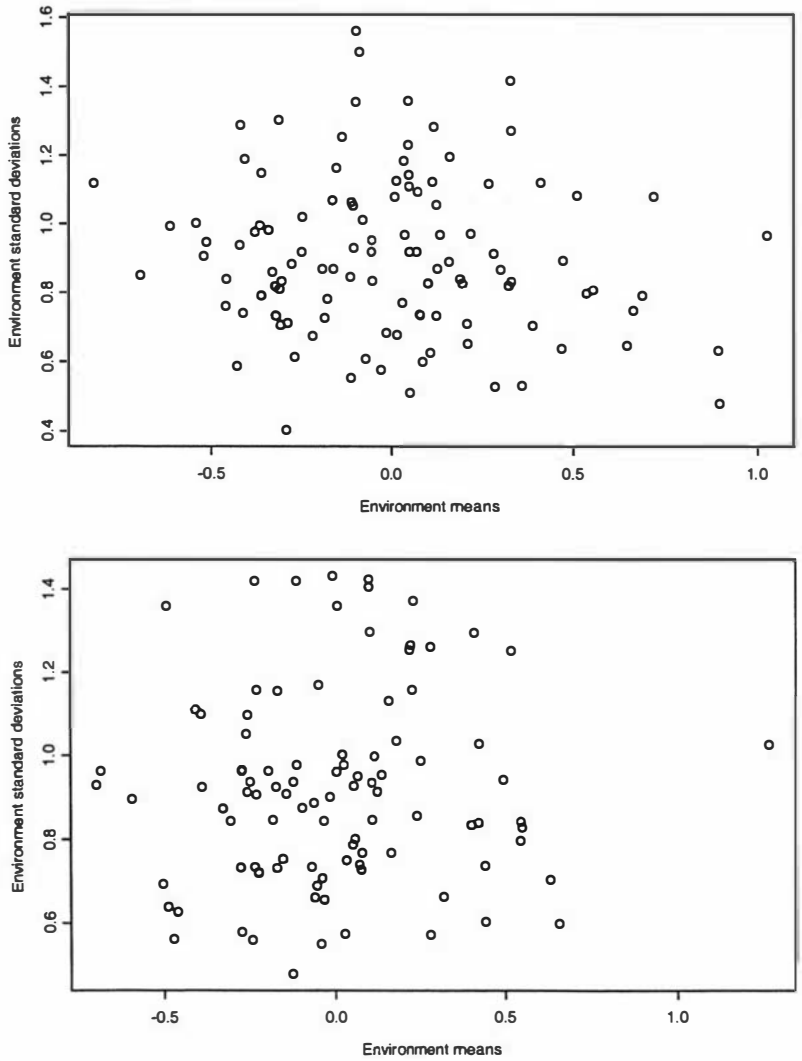


Figure 7.7: Environment standard deviations plotted against environment means for Onion Data I and II, after within-genotype standardization of the fully imputed G×E matrices found in Section 6.5.

will be assured.

The fact that the heteroscedasticity across environments after within-genotype standardization was greater than the heteroscedasticity across genotypes before it, highlighted the differences of the environments on trial. Environments with similar variation were likely to cluster sooner than ones with dissimilar within-environment variation. The richness of the data sets would therefore contribute to the clustering, if the second transformation was applied. The decision was taken to cluster environments using within-genotype standardized data after two-stage imputation of each of Onion Data I and II in this presentation, for comparability to the clusterings presented in the previous section.

Clustering could also have been performed on the raw data or using within-environment transformations as recommended by Fox and Rosielle (1982) and Cooper *et al.* (1993). Use of these transformations will need to be given further consideration as part of the ongoing work of this investigation.

Figure 7.8 shows the dendrogram for the 109 environments of Onion Data I, with the

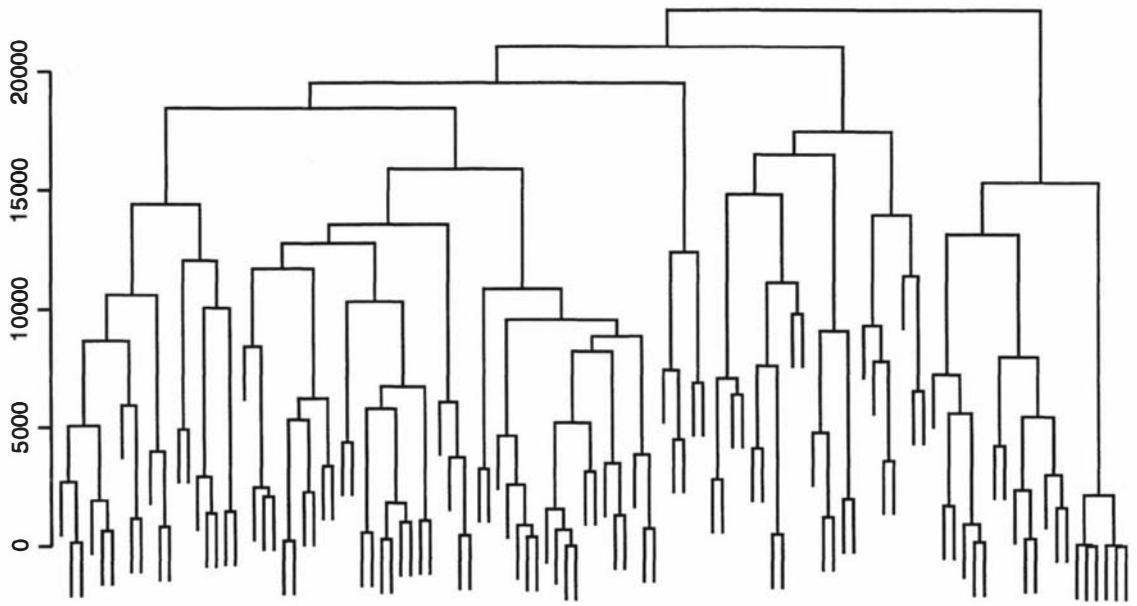


Figure 7.8: Dendrogram for the 109 environments of Onion Data I formed using the within-genotype standardized data found after two-stage imputation in Section 6.5. The stopping criterion truncates clustering at the level 8426 determining that 30 mega-environments exist.

cluster memberships presented in Table 7.4. Interaction distance I_{ij} and the incremental sum of squares method of forming new clusters were applied. At the level 8426, the stopping criterion determined that clustering should be truncated, resulting in 30 mega-environments.

Figure 7.9 shows the dendrogram of the 98 environments of Onion Data II after it was fully imputed using two-stage imputation. Interaction distance I_{ij} and the incremental sum of squares method of forming clusters were applied. The stopping criterion truncated clustering at the level 6195, and determined that 26 mega-environments existed. The memberships of the clusters in Figure 7.9 are presented in Table 7.5. In Chapter 8, these clusterings will be compared to one another, and to those of the previous section.

7.4 Use of covariates to cluster environments

Factors affecting the growth of crops are numerous, and observed G×E interactions result from the combination of these factors on tested varieties at each location. Examination of results from clustering environments using the collected covariate information from the Onion Trials Programme data, versus the possible cluster groupings offered in the previous sections will show that either:

1. The available or fully imputed yield data can substitute for collection of covariate information, thus saving time and effort of collaborators and organizers; or

2. Results from clustering based on yield data is at odds with the clustering based on covariate information. This may be because the collected covariate information

Cluster	Members
1	A00501, Nigeria, 1991; A03602, Australia, 1996; D08401, Mexico, 1992; X15101, Guinee, 1993.
2	D03602, Zimbabwe, 1990; D03603, Zimbabwe, 1990; F02407, Botswana, 1994; F02408, Botswana, 1994; F02702, Tanzania, 1994; O11601, Senegal, 1993; X10901, New Caledonia, 1994; Y07501, Honduras, 1996; Z13503, Nigeria, 1997.
3	D03610, Zimbabwe, 1999; D07101, Burkina, 1993; L05101, South Africa, 1993; P07200, Nepal, 1990; W16503, Brazil, 1993; W23003, Ethiopia, 1993; W23009, Ethiopia, 1997; X05701, Nigeria, 1994.
4	A04103, Bangladesh, 1992; F02701, Tanzania, 1992; Z10501, Cape Verde, 1996.
5	J08801, Mauritius, 1998; X15106, Guinee, 1995; X15107, Guinee, 1995; Z03501, Philippines, 1996.
6	A03401, Belize, 1990; O02701, Sri Lanka, 1992; O02709, Sri Lanka, 1994.
7	A01604, Yemen, 1993; O07601, Mozambique, 1991; W23005, Ethiopia, 1994; X04601, Cameroon, 1994; X07201, Nepal, 1993; X07202, Nepal, 1994.
8	A04101, Bangladesh, 1990; Y01401, Uruguay, 1993; Y03402, C Ivoire, 1994; Z03401, Pakistan, 1994.
9	W17502, Greece, 1993; Z09601, Korea, 1994.
10	D03604, Zimbabwe, 1991; F02404, Botswana, 1993; W17501, Greece, 1993; W17503, Greece, 1994; W17504, Greece, 1994; X13901, Senegal, 1994; Y07401, Taiwan, 1992; Z00101, Thailand, 1991.
11	W23007, Ethiopia, 1996; Z03402, Pakistan, 1995.
12	D03601, Zimbabwe, 1989; O02700, Sri Lanka, 1991; R00401, Tunisia, 1999. W16506, Brazil, 1994; Y01302, Argentina, 1997; Z03901, Ghana, 1997.
13	D03605, Zimbabwe, 1991; F02405, Botswana, 1993; F02406, Botswana, 1993; Y01501, Kenya, 1992.
14	W16508, Brazil, 1995; X04804, Zambia, 1993; Y13102, St Helena, 1997.
15	A00301, Barbados, 1995; D02501, Egypt, 1996; O07603, Mozambique, 1997; W16502, Brazil, 1992; W22801, Italy, 1990; Y03404, C Ivoire, 1998.
16	F04006, Uganda, 1993; R00702, India, 1998.
17	O02702, Sri Lanka, 1992; O02703, Sri Lanka, 1993; W23001, Ethiopia, 1993.
18	A03603, Australia, 1997; A04001, India, 1993; A04102, Bangladesh, 1991; A04202, Mauritania, 1994; Y13101, St Helena, 1997.
19	A01601, Yemen, 1992; L03701, Lesotho, 1992.
20	J00301, India, 1994; L01101, Kenya, 1997; Y07403, Taiwan, 1993.
21	X10801, P R China, 1998; Y07405, Taiwan, 1996; Z03903, Ghana, 1998.
22	A03101, Malaysia, 1998; X15103, Guinee, 1994.
23	O07604, Mozambique, 1997; X13801, Mali, 1997; Z09603, Korea, 1995.
24	P07201, Nepal, 1991; P07202, Nepal, 1991; P07203, Nepal, 1991; P07204, Nepal, 1991; P07205, Nepal, 1991; P07206, Nepal, 1991.
25	W13204, Nepal, 1992; X15105, Guinee, 1995.
26	R00101, Australia, 1996; R00102, Australia, 1997.
27	A05902, Fiji, 1991.
28	R00701, India, 1998.
29	W05201, Bulgaria, 1995.
30	Y13901, Nigeria, 1995.

Table 7.4: Cluster memberships for the 109 environments of Onion Data I, clustered using the fully imputed data, presented in Figure 7.8.

Cluster	Members
1	O02701, Sri Lanka, 1992; X13801, Mali, 1997.
2	L03701, Lesotho, 1992; Y01401, Uruguay, 1993; Y03402, C Ivoire, 1994.
3	D03604, Zimbabwe, 1991; D03605, Zimbabwe, 1991; O07603, Mozambique, 1997; R00702, India, 1998.
4	A00501, Nigeria, 1991; D07101, Burkina, 1993; O02702, Sri Lanka, 1992; O02703, Sri Lanka, 1993; W23001, Ethiopia, 1993; X15101, Guinee, 1993; X15105, Guinee, 1995.
5	D03603, Zimbabwe, 1990; F04006, Uganda, 1993; L01101, Kenya, 1997; Y01501, Kenya, 1992; Y07403, Taiwan, 1993; Z09603, Korea, 1995.
6	A00301, Barbados, 1995; A04202, Mauritania, 1994; F02404, Botswana, 1993; F02407, Botswana, 1994; F02408, Botswana, 1994; O02700, Sri Lanka, 1991; R00101, Australia, 1996; R00102, Australia, 1997; R00401, Tunisia, 1999; W16506, Brazil, 1994; W22801, Italy, 1990; Y07401, Taiwan, 1992; Z03903, Ghana , 1998.
7	A05902, Fiji, 1991; W23009, Ethiopia, 1997; Y13901, Nigeria, 1995.
8	D03610, Zimbabwe, 1999; O02709, Sri Lanka, 1994; R00701, India, 1998; W23005, Ethiopia, 1994; X05701, Nigeria, 1994; X15106, Guinee, 1995.
9	D03601, Zimbabwe, 1989; D03602, Zimbabwe, 1990; D08401, Mexico, 1992; X10801, P R China , 1998; X15103, Guinee, 1994; Z13503, Nigeria, 1997.
10	A04103, Bangladesh, 1992; Z09601, Korea, 1994.
11	A03401, Belize, 1990; D02501, Egypt, 1996.
12	P07201, Nepal, 1991; P07202, Nepal, 1991; P07203, Nepal, 1991; P07204, Nepal, 1991; P07205, Nepal, 1991; P07206, Nepal, 1991; W16502, Brazil, 1992; W16508, Brazil, 1995.
13	J08801, Mauritius, 1998; O07604, Mozambique, 1997; O11601, Senegal, 1993.
14	F02702, Tanzania, 1994; O07601, Mozambique, 1991; Z03402, Pakistan, 1995; Z03501, Philippines, 1996.
15	A03101, Malaysia, 1998; A03603, Australia, 1997; X10901, New Caledonia, 1994; X13901, Senegal, 1994; Y01302, Argentina, 1997.
16	F02701, Tanzania, 1992; Y07405, Taiwan, 1996; Y07501, Honduras, 1996; Z03401, Pakistan, 1994; Z10501, Cape Verde, 1996.
17	A03602, Australia, 1996; X04804, Zambia, 1993.
18	F02405, Botswana, 1993; F02406, Botswana, 1993.
19	A01604, Yemen, 1993; J00301, India, 1994; W23003, Ethiopia, 1993.
20	A04001, India, 1993; W17503, Greece, 1994; W17504, Greece, 1994.
21	A01601, Yemen, 1992; A04102, Bangladesh, 1991; L05101, South Africa, 1993; X15107, Guinee, 1995.
22	A04101, Bangladesh, 1990.
23	W05201, Bulgaria, 1995.
24	W23007, Ethiopia, 1996.
25	X07201, Nepal, 1993.
26	X07202, Nepal, 1994.

Table 7.5: Cluster memberships for the 98 environments of Onion Data II, clustered using the fully imputed yield data in Figure 7.9.

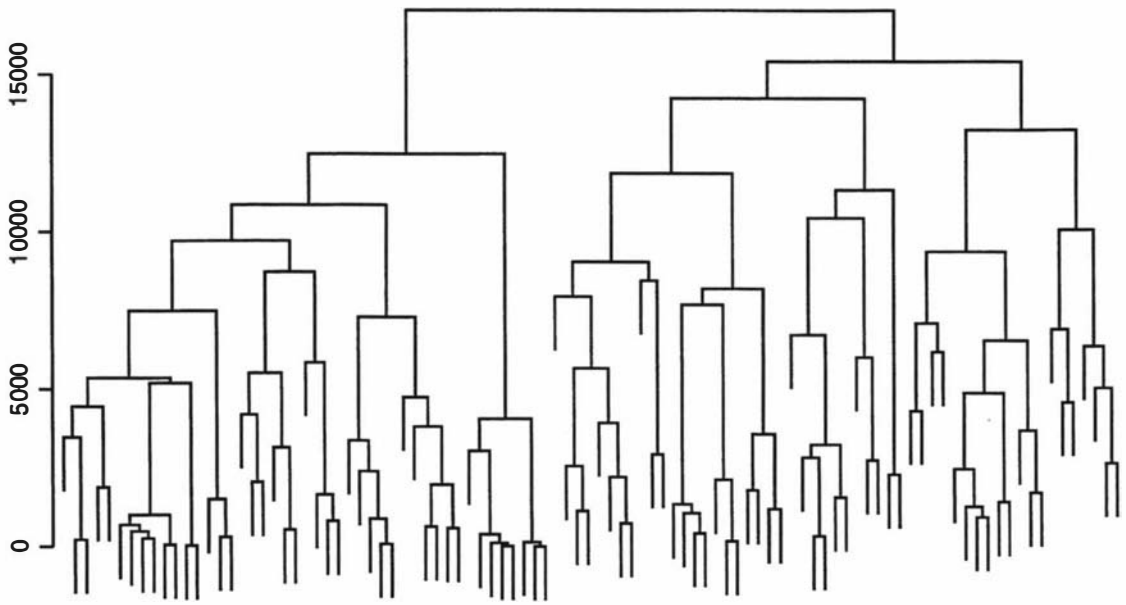


Figure 7.9: Dendrogram for the 98 environments of Onion Data II formed using the within-genotype standardized data found after two-stage imputation in Section 6.5. The stopping criterion truncated clustering at the level 6195, determining that 26 mega-environments existed.

does not include all influential factors from an environment; or it may disprove the original notion.

In this section, environments are clustered using the available covariate information, while comparison of results from separate cluster analyses is left to the next chapter.

Available covariate information specific to environments was limited to latitude, altitude, and initial sowing date of crops, but sowing dates were not directly comparable. Environments from different hemispheres with similar growing conditions should be grouped together despite differences in sowing dates (which are chosen to make the most of the location conditions). While in general, a northern hemisphere environment would have a sowing date approximately six months different from an environment in the southern hemisphere that has the same climatic conditions, tropical environments do not necessarily follow this rule. No adjustment for hemisphere was therefore undertaken, and the sowing date for each environment was not used in the clustering that follows.

Further covariate information was available for each genotype-environment combination; information collected on temperatures and photoperiods varied according to the growing period for each genotype. It is well known that success of onion growth is dependent on the amount of sunlight and the temperature crops are subjected to over their growth period (Brewster,1994). This notion led to photoperiod and temperature data for each G×E combination being recorded by many collaborators. Of the 123 environments, only 101 had full temperature and photoperiod data available. Clustering for this set of environments, along with the subset of these environments that were used in Onion Data I

and II, described in Section 3.6 are presented in this section. The strategy used to form a set of variables that represented each environment is now presented.

Photoperiod is known to be dependent on latitude, and temperature dependent on both latitude and altitude. Latitude and altitude were therefore removed from further consideration. Environments cannot be clustered using the recorded photoperiod and temperature data directly because the length of growing periods differed between and within environments. This can be seen most easily in Figure 3.12.

The strategy that was followed is to create proxy variables that were based on 'heat units' and photoperiods. These variables must be comparable over environments, so they needed to be found across pre-determined portions of the growing period. Each environment had its own default season length which depended on the local practices using varieties common to that environment. An arbitrary choice was made to use four time periods for each of the two sets of proxy variables.

Heat unit and photoperiod values for each G×E combination were found by averaging these variables within each of the four time periods. The values used for each environment were then found by averaging values from relevant G×E combinations. Eight indices were thus provided for each environment, and although not without faults they sufficed in the absence of more consistent covariate information. A particular concern was that the indices were dependent on the genotypes grown in each environment. There was potential for an environment's growing period to be determined by tested genotypes, rather than the growing periods commonly used in such environments. For some environments this was obviated, as the collaborator did not allow the genotypes to grow over differing lengths of time. This in itself was problematic, as the genotypes were not necessarily allowed to perform optimally. This is yet another concern that is addressed in Chapter 10 in greater detail.

The method for finding these two sets of proxy variables is now described in detail. Minimum and maximum temperatures for each week over the growing period of each G×E combination were used to create heat unit variables in the following way:

1. Establish the total growing period in weeks for each G×E combination.

2. Decide how many portions the total growing period will be broken into. Note that the number of weeks in a portion of the growing period will change over both genotypes and environments.

3. Identify the weekly minimum and maximum temperatures for each G×E combination in each portion of the growing period.

4. Calculate the heat units for each portion of the growing period using

$$H_{ikp} = \frac{\sum_{w=1}^{g_{ikp}} \left(T_{ikw}^{\min} + T_{ikw}^{\max}\right)/2 - 5}{g_{ikp}} \tag{7.1}$$

where g_{ikp} is the number of weeks in the p th portion of the total growing period for genotype i in environment k, while $T_{ikw}^{\min}$ and $T_{ikw}^{\max}$ are the minimum and maximum temperatures in week w for genotype i in environment k. This has been modified from the method proposed by de Ruiter (1986) which did not cater for differing total growing periods for genotypes. This heat unit calculation also takes into account the lack of any benefit to the growth of plants from temperatures below five degrees.

5. Take averages over genotypes in each environment of the heat unit indices H_{ikp} to give $\bar{H}_{.kp}$.

6. Replace any negative heat unit values ($\bar{H}_{.kp} < 0$) with zero heat units, which indicates no benefit from heat in that growing period (de Ruiter, 1986).

The photoperiod variables were found in a similar way, using the following process:

1. Establish the total growing period in weeks for each G×E combination.

2. Decide how many portions the total growing period will be broken into. Note that the number of weeks in a portion of the growing period will change over both genotypes and environments.

3. Identify the weekly average photoperiods for each G×E combination in each portion of the total growing period.

4. Calculate the average photoperiod for each portion of the total growing period using

$$\bar{P}_{ikp} = \frac{\sum_{w=1}^{g_{ikp}} P_{ikw}}{g_{ikp}} \tag{7.2}$$

where P_{ikw} is the average photoperiod in week w for genotype i in environment k.

5. Take the average of the photoperiod indices $\bar{P}_{ikp}$ over the genotypes in each environment k to give $\bar{P}_{.kp}$.

The values found for these proxy variables are compiled in a data set on the CD-ROM accompanying this volume.

The use of other multivariate methods, such as discriminant analysis, based on these covariates is difficult to present as the number of environments is large and the number of

Correlation	Heat Units 1	Heat Units 2	Heat Units 3	Heat Units 4
Photoperiod 1	0.6291	0.6714	0.5106	0.1530
Photoperiod 2	0.4009	0.7068	0.7072	0.3147
Photoperiod 3	−0.2306	0.1282	0.3151	0.1796
Photoperiod 4	−0.7403	−0.6931	−0.4934	−0.1999
Heat Units 1		0.8655	0.7159	0.5741
Heat Units 2			0.9085	0.6267
Heat Units 3				0.7962
		Photoperiod 2	Photoperiod 3	Photoperiod 4
Photoperiod 1		0.7635	0.1043	−0.6581
Photoperiod 2			0.6305	−0.3757
Photoperiod 3				0.4535

Table 7.6: Correlations of proxy variables for environments, calculated using (7.1) and (7.2), for the 101 environments for which full covariate information was available.

mega-environments in the data has not been established. Once clustering has been completed, and therefore the number of mega-environments, a set of decision rules to classify any new environments into one of the current mega-environments could be created. The use of principal components to visually display the environment groupings is a technique that will be considered further in Chapter 9.

Before moving on to the clustering of environments, a preliminary investigation of the proxy variables is presented. The value of using these variables to cluster environments was questionable. They need to show enough different information to warrant their use as variables for clustering. Correlations, calculated within and between the two sets of variables, are presented in Table 7.6. Figure 7.10 plots the four heat unit variables against one another, while Figure 7.11 gives the corresponding plots for the four photoperiod variables.

The heat unit variables showed strong positive correlation between all pairs of indices. It was more interesting to note that the photoperiod variables displayed both strong positive and strong negative correlation, and that some pattern existed in these correlations. Photoperiod 1 scores show strong negative correlation with photoperiod 4 scores. The correlations between successive periods is high, while in general the further apart the two periods were the less they were correlated. The only correlation that is not significant among the photoperiod variables is that between photoperiods 1 and 3 of 0.1043. A regression of each of the proxy variables against the environment latitude and altitude, summarized in Table 7.7, showed that each of the Heat Unit variables was explained well by the latitude and altitude of the environments, and also that the photoperiods were dependent on latitude alone. These regression models were indicative of the relationship, and were not tested for their optimality in terms of finding the best relationship for use as a predictive model. Residuals were in general symmetric about zero, and as shown by the R^2 values the models were clearly useful.

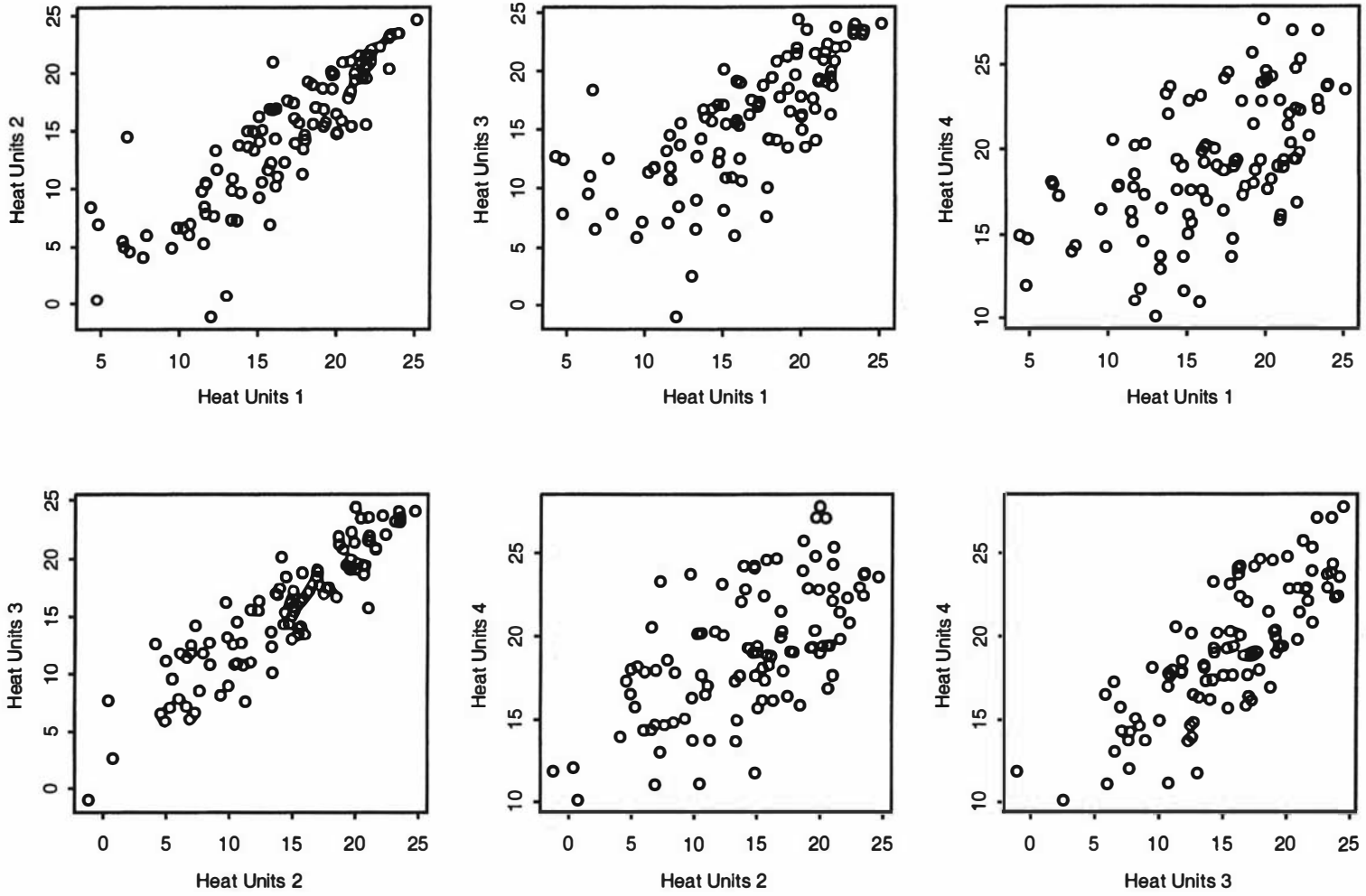


Figure 7.10: Heat unit variables plotted against one another. The values for these variables were found using the formula presented in (7.1).

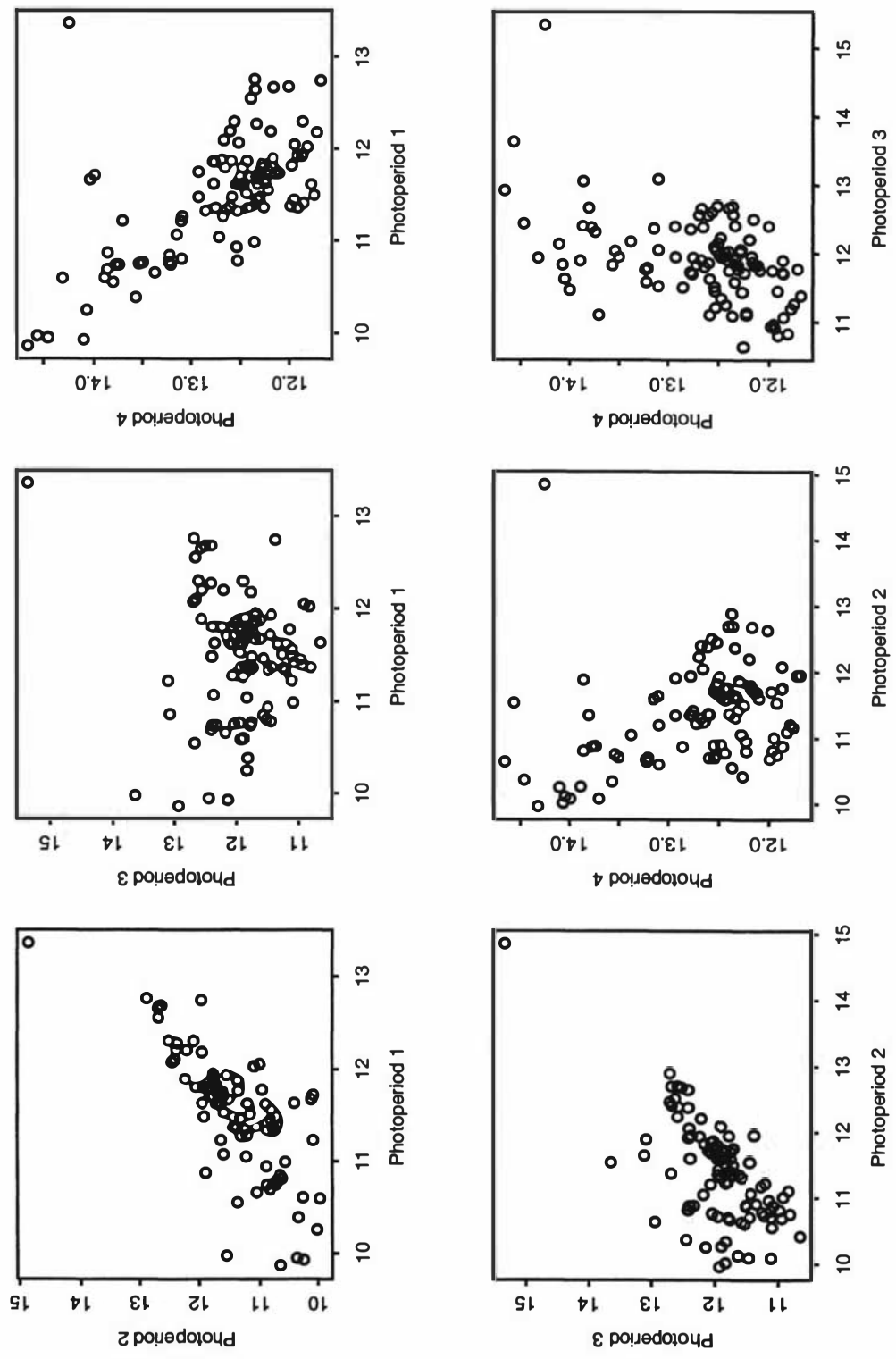


Figure 7.11: Plots of photoperiod variables against one another. The values for these variables were found using the formula given in (7.2).

Proxy variable	Regression coefficient latitude	altitude	p values latitude	altitude	Model R^2
Heat Unit 1	−0.369(0.026)	−0.005(0.000)	0.000	0.000	73.8%
Heat Unit 2	−0.519(0.023)	−0.005(0.000)	0.000	0.000	85.0%
Heat Unit 3	−0.0428(0.030)	−0.003(0.001)	0.000	0.000	69.9%
Heat Unit 4	−0.182(0.033)	−0.003(0.001)	0.000	0.000	33.6%
Photoperiod 1	−0.052(0.004)	0.000(0.000)	0.000	0.231	66.4%
Photoperiod 2	−0.060(0.003)	0.000(0.000)	0.000	0.724	79.0%
Photoperiod 3	−0.011(0.005)	0.000(0.000)	0.052	0.255	61.4%
Photoperiod 4	−0.053(0.005)	0.000(0.000)	0.000	0.118	54.3%

Table 7.7: Summary of the regression models using each of the eight proxy variables constructed to summarize environments as response variables, and the latitude and altitude of environments as explanatory variables. The numbers in brackets after the regression coefficients are their standard errors, while the p values are those found testing the coefficients against the null hypothesis that they are equal to zero.

If the set of proxy variables are to explain the environments fully, they should be capable of replacing the geographic information. Table 7.8 summarizes the best subsets regression output for latitude and altitude against the constructed variables. It shows that a very good estimate of the latitude could be found if the potential environmental conditions were known. The altitude of an environment could also be predicted using these variables, so they could be used to determine the missing altitude data from some environments. On the whole this analysis showed that the proxy variables serve well for the cluster analyses that follow.

Now that variables on which to base clustering have been obtained, it was necessary to ensure that each contributed to the determination of mega-environments equally. The need for standardization of these variables was investigated. The variability of the raw data could impact on the outcome of clustering. If left untransformed the heat unit variables would contribute significantly more than the photoperiod variables simply because of their scale. Two standardizations were considered in this instance:

1. Standardization into units of standard deviation:

$$z_{ik} = \frac{x_{ik} - \bar{x}_i}{s_i} \tag{7.3}$$

where x_{ik} is the untransformed covariate, z_{ik} the transformed covariate, $\bar{x}_i$ the mean value of covariate i, and s_i the standard deviation of covariate i; or

2. Standardization into a $[0, 1]$ range:

$$z_{ik} = \frac{x_{ik} - x_i^{\min}}{x_i^{\max} - x_i^{\min}} \tag{7.4}$$

where $x_i^{\max}$ and $x_i^{\min}$ refer to the maximum and minimum values for covariate i.

Response variable	No. of Variables	Model R^2	Mallow's C_k	Variables included in model
Latitude	1	79.0%	96.3	Photoperiod 2
	1	66.3%	212.6	Heat Unit 2
	2	87.6%	19.5	Photoperiods 2 & 4
	2	86.5%	29.1	Photoperiods 2 & 3
	3	88.2%	15.5	Heat Unit 4, Photoperiods 2 & 4
	3	88.2%	15.6	Heat Unit 3 Photoperiods 2 & 4
	4	88.9%	11.4	Heat Units 3, Photoperiods 1, 3, & 4
	4	88.8%	12.1	Heat Unit 4, Photoperiods 1, 3, & 4
	5	89.8%	4.5	Heat Units 1 & 3, Photoperiods 1, 3, & 4
	5	89.5%	7.7	Heat Units 1 & 4, Photoperiods 1, 3, & 4
	6	89.9%	5.5	Heat Units 1, 3, & 4, Photoperiods 1, 3, & 4
	6	89.9%	6.4	Heat Units 1, 2, & 3, Photoperiods 1, 3, & 4
	7	90.0%	7.0	All but Photoperiod 2
	7	89.9%	7.5	All but Heat Units 2
	8	90.0%	9.0	All eight variables
Altitude	1	17.1%	112.3	Heat Units 1
	1	12.3%	124.4	Heat Units 4
	2	48.7%	35.2	Heat Units 1, Photoperiod 1
	2	46.1%	41.7	Heat Units 2, Photoperiod 2
	3	57.2%	16.0	Heat Units 1, Photoperiods 2 & 3
	3	56.9%	16.9	Heat Units 2, Photoperiod 2 & 4
	4	61.9%	6.4	Heat Units 1 & 2, Photoperiods 2 & 4
	4	60.8%	9.1	Heat Units 1 & 2, Photoperiods 2 & 3
	5	63.3%	4.7	Heat Units 1, 2, & 4, Photoperiods 2 & 4
	5	63.2%	5.1	Heat Units 1, 2, & 3, Photoperiods 2 & 4
	6	63.8%	5.6	Heat Units 1, 2, & 4, Photoperiods 1, 2, & 4
	6	63.7%	5.8	Heat Units 1, 2, & 3, Photoperiods 1, 2, & 4
	7	64.0%	7.0	All but Photoperiod 3
	7	63.8%	7.6	All but Heat Units 3
	8	64.0%	9.0	All eight variables

Table 7.8: Summary of best subsets regression using the latitude and altitude of an environment as response variables, and the eight proxy variables as explanatory variables.

This will provide values that fall in the range $0 \leqslant z_{ik} \leqslant 1$.

Both sets of transformed variables were checked for normality and compared for distributional equality. Figure 7.12 shows the normal probability plots of the amalgamated proxy variables and boxplots of each proxy variable for the 101 environments that had full covariate information available. It showed that the better transformation to use in this instance was standardization. There was too much heteroscedasticity across the range transformed variables, shown by the boxplots in Figure 7.12. The only concern in using standardization would be outliers in some variables having excess influence on results; there were no outliers evident in this case.

Figure 7.13 shows the dendrogram of the 101 environments, for which full covariate information was available, clustered using the incremental sum of squares method (described in Section 5.2) and Euclidean distance (described in Chapter 4). The stopping criterion described in Section 5.2 truncated clustering at a level of 24.21 and determined that there are fifteen clusters of environments. Table 7.9 displays the members of these fifteen mega-environments. Appendix A presents a more detailed list of the attributes of the environments.

The clustering process performed for all 101 environments with full covariate information was then repeated for the subsets of these environments that were part of Onion Data I and II. The checks for the normality and homoscedasticity of the covariates are presented in Figures 7.14 and 7.15. The same conclusion was drawn from these investigations as before. The normal standardized covariates were therefore used to cluster these two subsets of data in Figures 7.16 and 7.17. These dendrograms were created using the same methodology as Figure 7.13, including the stopping criterion which yielded different results. The reduced number of environments that were included in each of these figures were also clustered differently to Figure 7.13 above. Clustering was truncated at the levels 24.40 and 23.36 in these figures, yielding thirteen and twelve mega-environment groupings respectively. These memberships are displayed in Tables 7.10 and 7.11 which follow. Section 8.2 will investigate the inter-relationships of these three different dendrograms in more detail.

7.5 Summary

In this chapter, mega-environments were found in three ways. The observed (sparse) yield data of the Onion Trials Programme were used first, and then the imputed values for Onion Data I and II. The third method used proxy variables that were created for each environment using the covariate information collected by collaborators for each G×E combination.

There is a need to put all the findings of this chapter together and develop one method that will be used to determine the mega-environments in the data. The covariate clustering

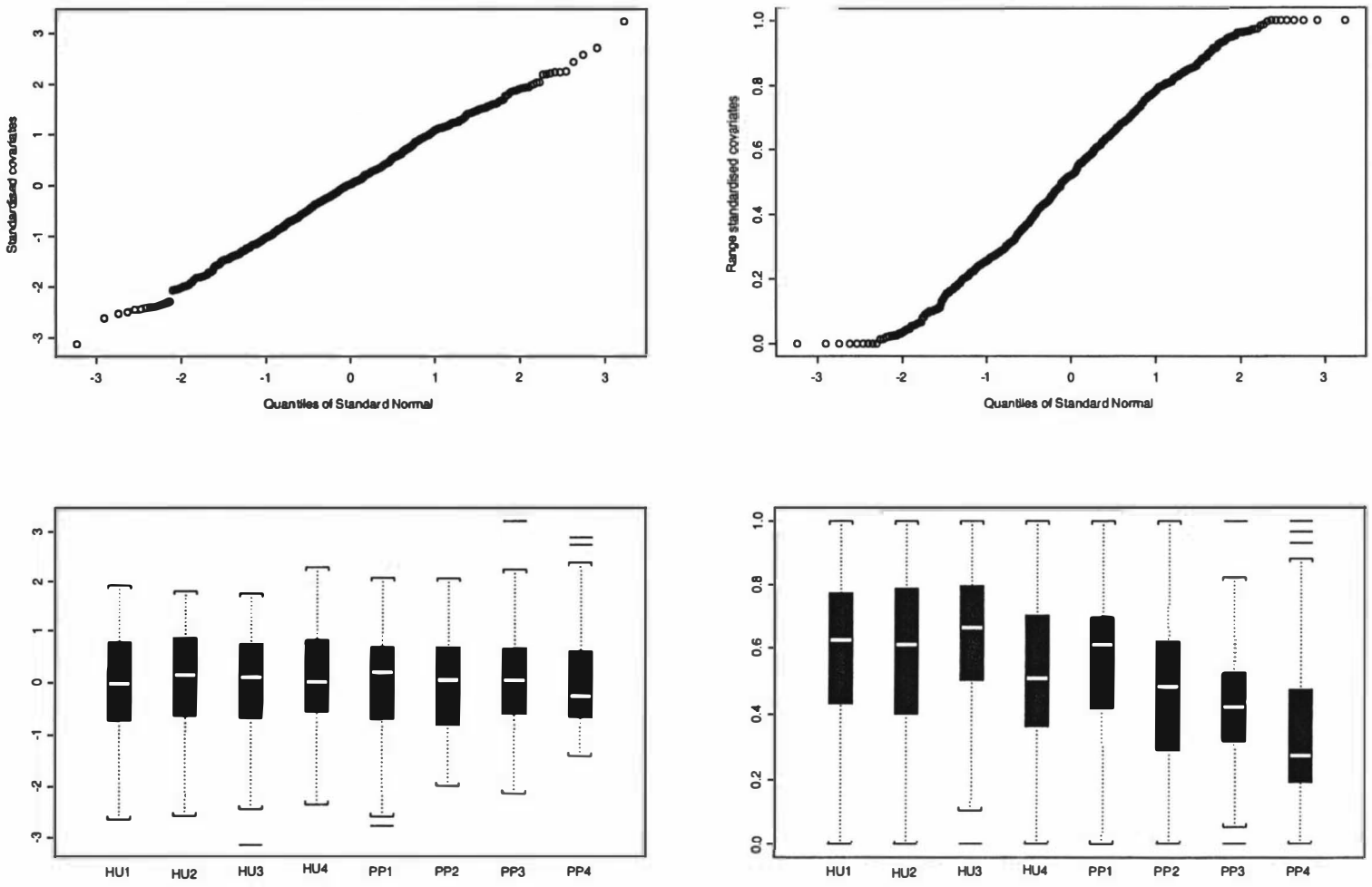


Figure 7.12: Normal probability plots and boxplots of transformed proxy variables for all 101 environments that had full covariate information available. The two transformations applied were standardization (left) and the range standardization (right) as described by (7.3) and (7.4) respectively.

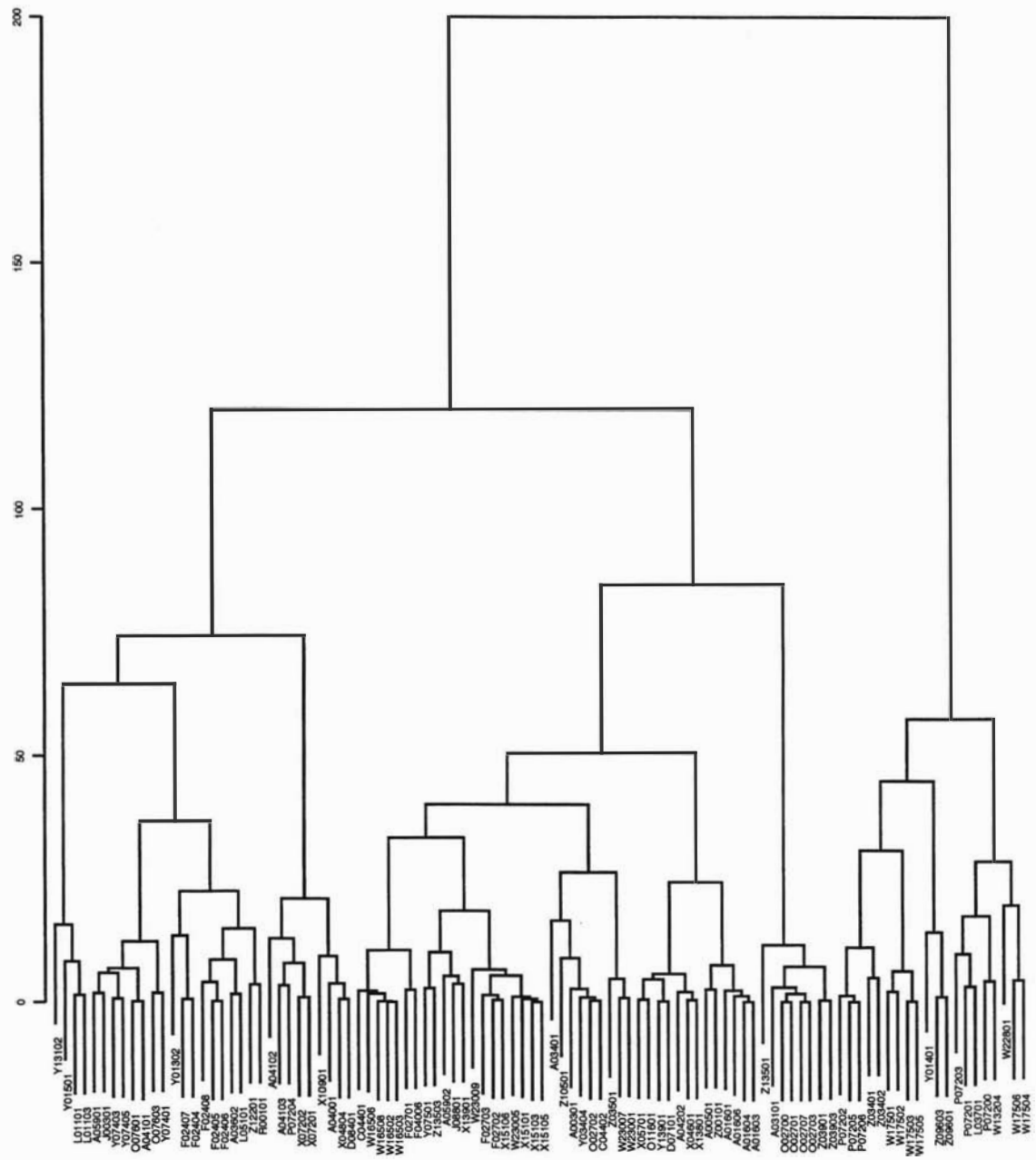


Figure 7.13: Dendrogram of 101 environments clustered using proxy variables, formed using (7.1) and (7.2), for all the environments that have full covariate information available. The stopping criterion used creates fifteen clusters in this data, with clustering truncated at the level of 24.21.

Cluster	Members
1	F02701, Tanzania, 1992; L01103, Kenya, 1998; L03701, Lesotho, 1992; O02700, Sri Lanka, 1991; O02701, Sri Lanka, 1992; P07200, Nepal, 1990; P07203, Nepal, 1991; W17504, Greece, 1994; W17505, Greece, 1995; W23005, Ethiopia, 1994; X04804, Zambia, 1993; X07201, Nepal, 1993; X10901, New Caledonia, 1994
2	A01604, Yemen, 1993; A04101, Bangladesh, 1990; A04102, Bangladesh, 1991; F02407, Botswana, 1994; F02702, Tanzania, 1994; O02707, Sri Lanka, 1994; O07601, Mozambique, 1991; P07201, Nepal, 1991; W16506, Brazil, 1994; W17501, Greece, 1993
3	W16508, Brazil, 1995; W23009, Ethiopia, 1997; X04601, Cameroon, 1994; X13901, Senegal, 1994; X15101, Guinee, 1993; X15105, Guinee, 1995; X15106, Guinee, 1995; Y01302, Argentina, 1997; Y01501, Kenya, 1992
4	Z12201, Malawi, 1996; Z13501, Nigeria, 1997; Z13503, Nigeria, 1997
5	A04001, India, 1993; A05902, Fiji, 1991; D07101, Burkina, 1993; F02703, Tanzania, 1996; J00301, India, 1994; J08801, Mauritius, 1998; L01101, Kenya, 1997; P07202, Nepal, 1991; P07204, Nepal, 1991; P07206, Nepal, 1991; R00101, Australia, 1996; W16503, Brazil, 1993; W22801, Italy, 1990
6	Y07501, Honduras, 1996; Y13102, St Helena, 1997; Y13901, Nigeria, 1995; Z00101, Thailand, 1991; Z03401, Pakistan, 1994
7	F04006, Uganda, 1993; O11601, Senegal, 1993; W13204, Nepal, 1992; X05701, Nigeria, 1994; X13801, Mali, 1997; X15103, Guinee, 1994
8	A00301, Barbados, 1995; F02404, Botswana, 1993; F02405, Botswana, 1993; F02406, Botswana, 1993
9	Y07405, Taiwan, 1996; Z03501, Philippines, 1996; Z03901, Ghana , 1997
10	A01601, Yemen, 1992; A01603, Yemen, 1993; A03602, Australia, 1996; A04202, Mauritania, 1994; F02408, Botswana, 1994; P07205, Nepal, 1991; W17503, Greece, 1994; X07202, Nepal, 1994
11	A05901, Fiji, 1991; C04402, PNG, 1993; L05101, South Africa, 1993; O02702, Sri Lanka, 1992; O02703, Sri Lanka, 1993; O07603, Mozambique, 1997; W17502, Greece, 1993; W17506, Greece, 1995
12	Y01401, Uruguay, 1993; Y03404, C Ivoire, 1998; Y07401, Taiwan, 1992; Y07403, Taiwan, 1993; Z03402, Pakistan, 1995
13	A00501, Nigeria, 1991; A01606, Yemen, 1994; A03101, Malaysia, 1998; A03401, Belize, 1990; A04103, Bangladesh, 1992; C04401, PNG, 1993; D08401, Mexico, 1992
14	Z03903, Ghana , 1998; Z09601, Korea, 1994; Z09603, Korea, 1995; Z10501, Cape Verde, 1996
15	W16502, Brazil, 1992; W23001, Ethiopia, 1993; W23007, Ethiopia, 1996;

Table 7.9: Cluster memberships for the 101 environments of the Onion Trials Programme that had sufficient covariate information, shown by the dendrogram in Figure 7.13.

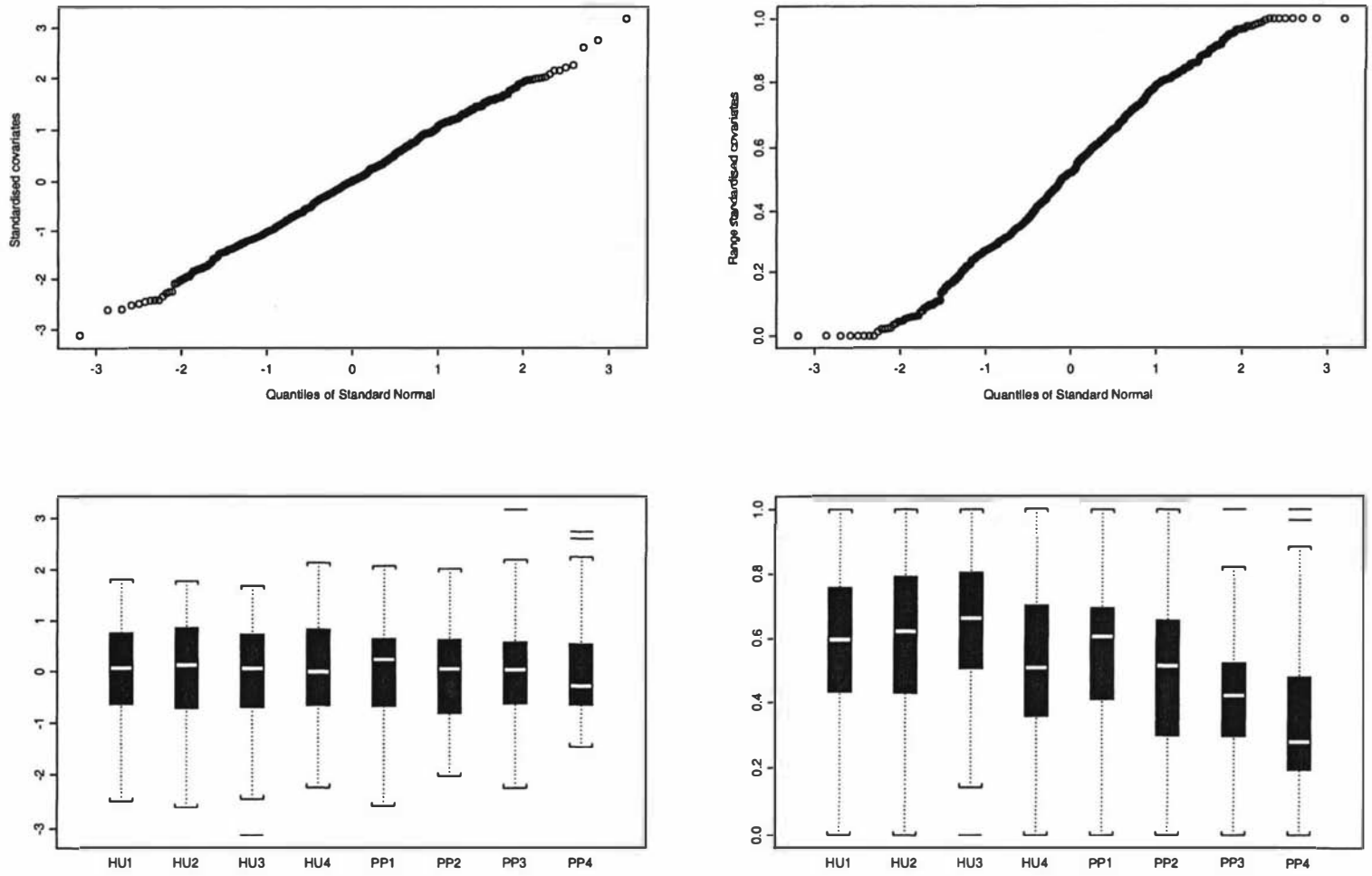


Figure 7.14: Normal probability plots and boxplots of transformed covariates for 89 environments that had full covariate information available, and were part of Onion Data I described in Section 3.6. The two transformations applied were standardization (left) and the range standardization (right) as described by (7.3) and (7.4) respectively.

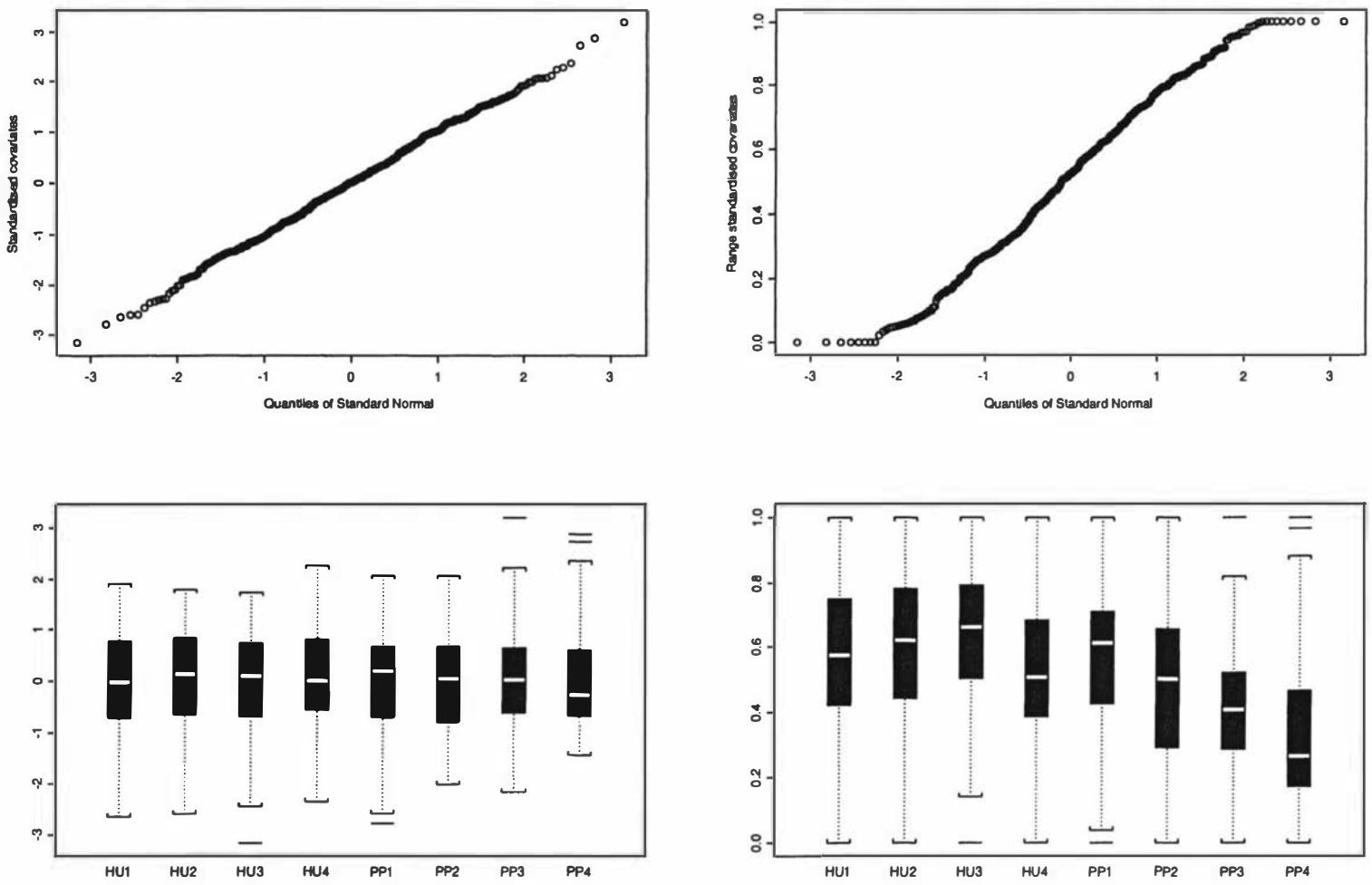


Figure 7.15: Normal probability plots and boxplots of transformed covariates for 79 environments that had full covariate information available, and are part of Onion Data II described in Section 3.6. The two transformations applied were standardization (left) and the range standardization (right) as described by (7.3) and (7.4) respectively.

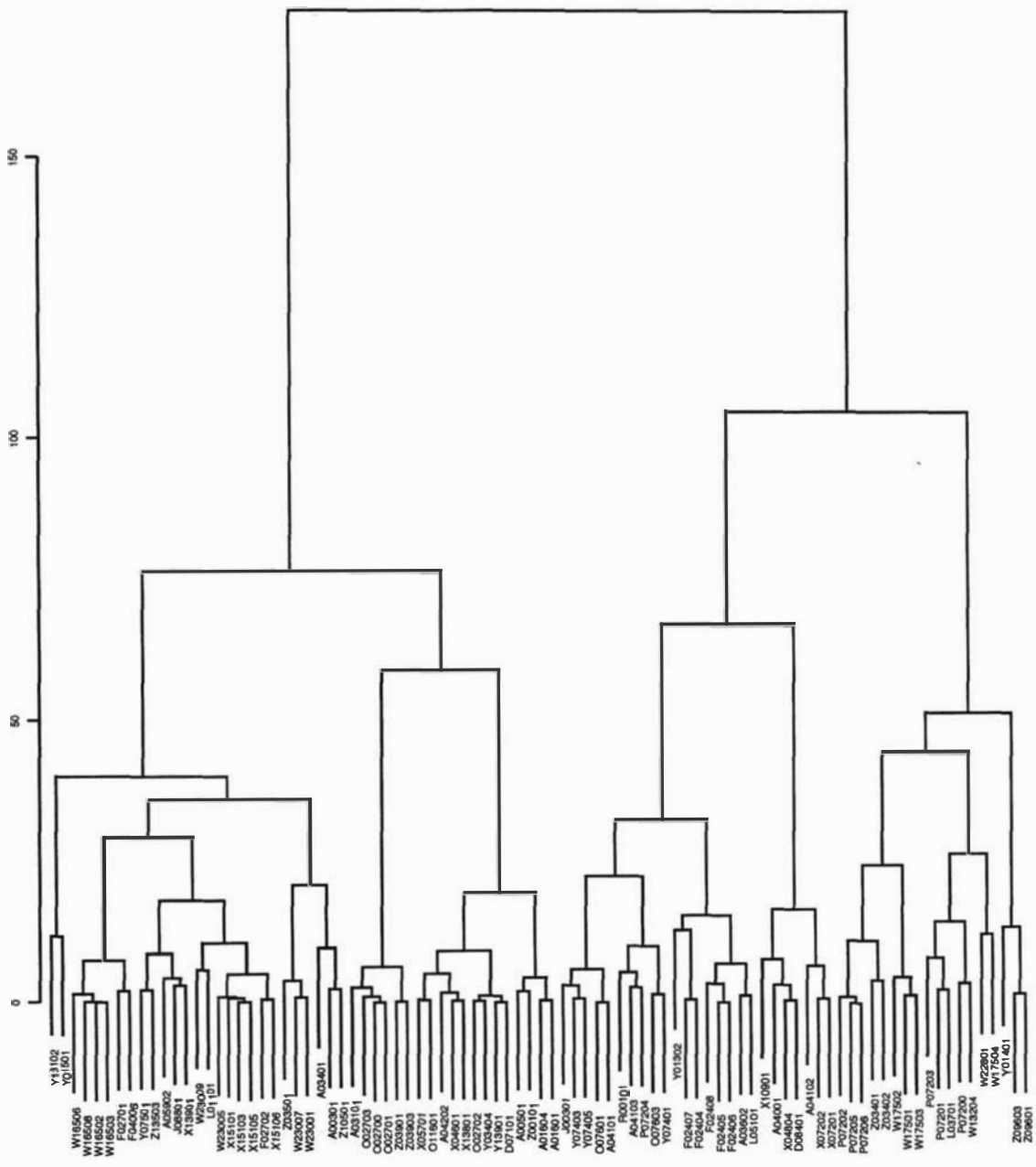


Figure 7.16: Dendrogram of 89 environments clustered using proxy variables, formed using (7.1) and (7.2), for the environments used in Onion Data I, described in Section 3.6 above. The stopping criterion used creates thirteen clusters in this data, with clustering truncated at the level of 24.40.

Cluster	Members
1	Y01401, Uruguay, 1993; Y03404, C Ivoire, 1998; Y07401, Taiwan, 1992; Y07403, Taiwan, 1993; Z03402, Pakistan, 1995; Z03903, Ghana , 1998; Z09601, Korea, 1994; Z09603, Korea, 1995
2	A01601, Yemen, 1992; A04001, India, 1993; A05902, Fiji, 1991; F02702, Tanzania, 1994; P07206, Nepal, 1991; W16508, Brazil, 1995; W17501, Greece, 1993; W17503, Greece, 1994; X07202, Nepal, 1994; X15101, Guinee, 1993
3	W16502, Brazil, 1992; W23005, Ethiopia, 1994; W23009, Ethiopia, 1997; X07201, Nepal, 1993; X13801, Mali, 1997; X15103, Guinee, 1994
4	F04006, Uganda, 1993; L01101, Kenya, 1997; O02700, Sri Lanka, 1991; O02701, Sri Lanka, 1992;O02703, Sri Lanka, 1993; O07601, Mozambique, 1991; P07200, Nepal, 1990; P07201, Nepal, 1991; P07204, Nepal, 1991; W17504, Greece, 1994; W23007, Ethiopia, 1996; X05701, Nigeria, 1994; X10901, New Caledonia, 1994
5	A04101, Bangladesh, 1990; D08401, Mexico, 1992; F02405, Botswana, 1993; F02407, Botswana, 1994; J08801, Mauritius, 1998; L03701, Lesotho, 1992; L05101, South Africa, 1993; P07203, Nepal, 1991; P07205, Nepal, 1991; R00101, Australia, 1996; W13204, Nepal, 1992; W16503, Brazil, 1993; W23001, Ethiopia, 1993
6	X04601, Cameroon, 1994; X04804, Zambia, 1993; X13901, Senegal, 1994; X15105, Guinee, 1995; X15106, Guinee, 1995; Y01302, Argentina, 1997; Y01501, Kenya, 1992
7	A01604, Yemen, 1993; A04102, Bangladesh, 1991; A04103, Bangladesh, 1992; F02701, Tanzania, 1992; J00301, India, 1994; O07603, Mozambique, 1997; P07202, Nepal, 1991; W16506, Brazil, 1994
8	Y07501, Honduras, 1996; Y13102, St Helena, 1997; Y13901, Nigeria, 1995; Z00101, Thailand, 1991; Z03401, Pakistan, 1994
9	Y07405, Taiwan, 1996; Z03501, Philippines, 1996; Z03901, Ghana , 1997
10	Z10501, Cape Verde, 1996; Z13503, Nigeria, 1997
11	A00301, Barbados, 1995; F02408, Botswana, 1994
12	A00501, Nigeria, 1991; A03101, Malaysia, 1998; A03401, Belize, 1990; A03602, Australia, 1996; A04202, Mauritania, 1994; F02406, Botswana, 1993
13	D07101, Burkina, 1993; F02404, Botswana, 1993; O02702, Sri Lanka, 1992; O11601, Senegal, 1993; W17502, Greece, 1993; W22801, Italy, 1990

Table 7.10: Cluster memberships for the dendrogram presented in Figure 7.16.

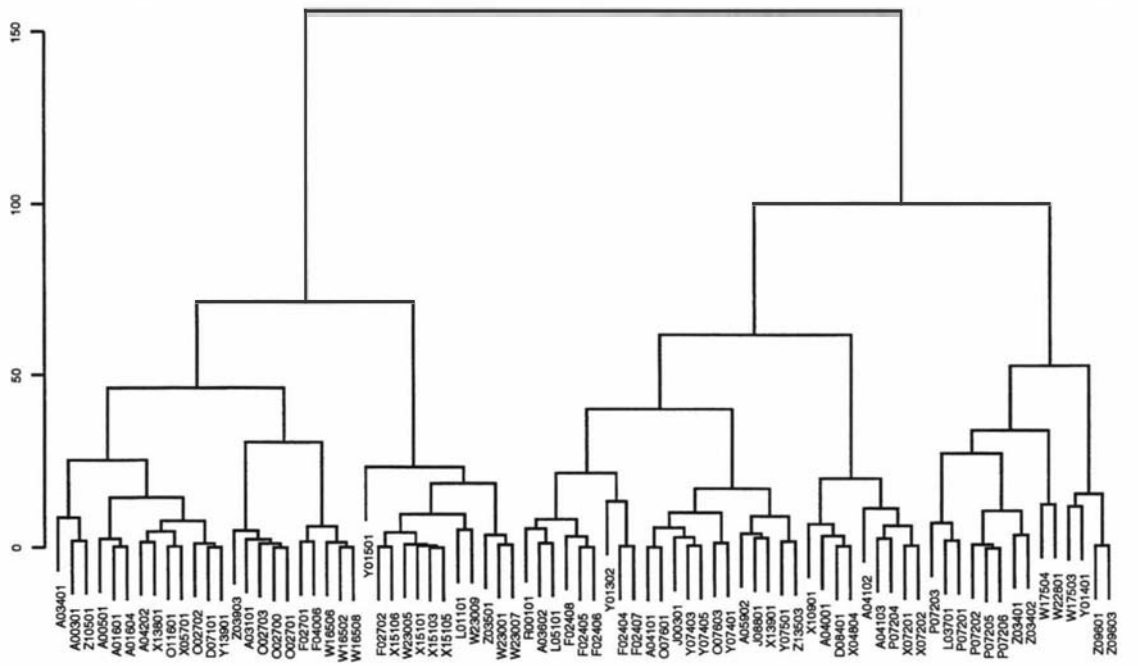


Figure 7.17: Dendrogram of 79 environments clustered using proxy variables, formed using (7.1) and (7.2), for the environments used in Onion Data II, described in Section 3.6. The stopping criterion used created twelve mega-environments in this data, with clustering truncated at the level of 23.36.

presented in Section 7.4 showed that there were, at most, fifteen mega-environments. On the other hand, the cluster analyses in Sections 7.2 and 7.3, using sparse and fully imputed yield data respectively, suggested that more than 25 mega-environments existed. If yield results amplify the results found using covariate information, clustering based on yields may have actually truncated the process prematurely. Careful examination of the tables of cluster memberships showed that this was not the case.

At some point, the most suitable of these methods (among others) needs to be determined. Clustering of sparse yield data relied on imputation of distances, but the fully imputed data depended on the imputation method applied. The ability to calculate distance measures directly would be determined by the G×E combinations collaborators and organizers chose to test. Trials managed by the same collaborator may in fact have had similarity of the varieties chosen. Trials with similar variety selection would have a greater opportunity to be determined similar, compared to those that had little or no commonality of tested varieties. The clustering of Section 7.2 must therefore be seriously questioned.

The necessity for imputation results to be independent of the selections of genotypes for each environment must also be considered. The next chapter presents some tools for determining the similarity of results from two cluster analyses. These and other issues can then be investigated further.

Cluster	Members
1	F02702, Tanzania, 1994; L01101, Kenya, 1997; W23001, Ethiopia, 1993; W23005, Ethiopia, 1994; W23007, Ethiopia, 1996; W23009, Ethiopia, 1997; X15101, Guinee, 1993; X15103, Guinee, 1994; X15105, Guinee, 1995; X15106, Guinee, 1995; Y01501, Kenya, 1992; Z03501, Philippines, 1996
2	A03602, Australia, 1996; F02404, Botswana, 1993; F02405, Botswana, 1993; F02406, Botswana, 1993; F02407, Botswana, 1994; F02408, Botswana, 1994; L05101, South Africa, 1993; R00101, Australia, 1996; Y01302, Argentina, 1997
3	A04001, India, 1993; A04102, Bangladesh, 1991; A04103, Bangladesh, 1992; D08401, Mexico, 1992; P07204, Nepal, 1991; X04804, Zambia, 1993; X07201, Nepal, 1993; X07202, Nepal, 1994; X10901, New Caledonia, 1994
4	A04101, Bangladesh, 1990; A05902, Fiji, 1991; J00301, India, 1994; J08801, Mauritius, 1998; O07601, Mozambique, 1991; O07603, Mozambique, 1997; X13901, Senegal, 1994; Y07401, Taiwan, 1992; Y07403, Taiwan, 1993; Y07405, Taiwan, 1996; Y07501, Honduras, 1996; Z13503, Nigeria, 1997
5	W17503, Greece, 1994; Y01401, Uruguay, 1993; Z09601, Korea, 1994; Z09603, Korea, 1995
6	A00501, Nigeria, 1991; A01601, Yemen, 1992; A01604, Yemen, 1993; A04202, Mauritania, 1994; D07101, Burkina, 1993; O02702, Sri Lanka, 1992; O11601, Senegal, 1993; X05701, Nigeria, 1994; X13801, Mali, 1997; Y13901, Nigeria, 1995
7	W17504, Greece, 1994; W22801, Italy, 1990
8	P07202, Nepal, 1991; P07205, Nepal, 1991; P07206, Nepal, 1991; Z03401, Pakistan, 1994; Z03402, Pakistan, 1995
9	A00301, Barbados, 1995; A03401, Belize, 1990; Z10501, Cape Verde, 1996
10	L03701, Lesotho, 1992; P07201, Nepal, 1991; P07203, Nepal, 1991
11	F02701, Tanzania, 1992; F04006, Uganda, 1993; W16502, Brazil, 1992; W16506, Brazil, 1994; W16508, Brazil, 1995
12	A03101, Malaysia, 1998; O02700, Sri Lanka, 1991; O02701, Sri Lanka, 1992; O02703, Sri Lanka, 1993; Z03903, Ghana, 1998

Table 7.11: Cluster memberships for the dendrogram presented in Figure 7.17.

Chapter 8

Comparing cluster analyses

8.1 The need to compare cluster analyses

Three different ways of determining mega-environments were presented in the previous chapter. Comparison of the results from each of these analyses was difficult because of the high number of observations and clusters. In Section 5.3 two clusterings were compared using a two-way table with rows and columns representing different sets of clusters. In that section, 'distortions' were loosely defined as causes of difference between the results from using incomplete data instead of complete data. Distortions will be used to mean a cause of difference throughout this chapter, and will not reflect the number of observations affected.

'Distortions' in a clustering reflect the number of major changes affecting the outcome of adding data in most of the situations discussed in this chapter. The best that can be expected is no distortions, which would arise when every pair of observations clustered together in one clustering, remain paired in a second clustering. A single distortion arises when a subset of a cluster, whether a single observation or group of observations, 'breaks away' from a cluster to form a new cluster or to join another cluster of observations.

The reason for counting distortions is that if a pair of observations is no longer within a cluster and they now reside in another cluster together, they are still clustered together; counting the number of observations that altered their cluster membership would discredit the fact that these observations were still considered similar enough to stay together in the new clustering. Distortions therefore can be counted to give a measure of the strength of the relationship between two clusterings, as they can show the level of dependence that one clustering has on another.

Table 5.2 was simple to use as there were only two distortions when the members of nine clusters were rearranged into ten clusters. Rows and columns of that table were rearranged to make it easier to work with. For larger tables this rearrangement can be laborious, and the table may well remain difficult to interpret due to the larger number of entries.

Rand (1971) proposed measuring the probability that a pair of observations are treated in the same way from one clustering to another. All pairs of observations were compared as he argued that observations not in a cluster determined its nature, as much as those observations that were in the cluster. Rand's (1971) measure, denoted R_g, will be used in this investigation, although it does not allow for a directional relationship between two clusterings. Cophenetic correlation is defined as the correlation of observed distances (similarities) between pairs of observations and the levels at which they fuse together in an agglomerative clustering (Everitt, 1993). Fowlkes and Mallows (1983) proposed a graphical approach for comparing hierarchical clusterings. Their measure directly compares cluster memberships as the number of clusters changes. These methods concentrate on the clustering process rather than the outcome of clustering. The Fowlkes and Mallows (1983) approach is also inappropriate in this investigation as the number of clusters differs from one clustering to another.

A simple graphical summary has been developed to show transfers of cluster membership effectively. 'Cluster influence diagrams', explained in the following section, show differences between the observational relationships within each of two multi-dimensional sets of data on a two-dimensional page. Numerical tools that measure the strength of relationships between two nominal variables were also applied in order to quantify this qualitative graphical tool for ease of comparison. The Goodman and Kruskal λ and Theil's uncertainty coefficient U will be introduced in Section 8.3.

These tools could be used to consider the effects of applying different distance measures or cluster formation strategies; instead, they were used to:

1. Illustrate the effect of adding observations to a data set (Sections 8.2 and 8.4 in particular).

2. Investigate the similarity of two-stage imputed values for Onion Data I and II (Section 8.5).

3. Compare the different methods of determining mega-environments presented in the previous chapter (Sections 8.5 and 8.6).

4. Determine if imputation altered the qualitative G×E interaction structure of the data, in terms of the way genotypes clustered together (Section 8.7).

5. Find out if imputations were dependent on inter-relationships between the test environments used for each genotype (Section 8.8).

The ultimate aim of this chapter is to identify how consistent results from various cluster analyses actually are. 'Consistency in cluster memberships' follows the work of Rand (1971), and is used to mean that pairs of observations that are within a cluster in one analysis, should be clustered together in another cluster analysis; those that are determined

to be different from one another in one cluster analysis, should remain separate in other cluster analyses.

8.2 Cluster influence diagrams

The need to uncover the number and cause of distortions that result in differences between two cluster analyses led to development of a new graphical tool, which will be referred to as a 'cluster influence diagram'. Specific features of cluster influence diagrams for the three dendrograms in Figures 7.13, 7.16, and 7.17 will be discussed after some introductory explanation. These figures show the clustering of environments based on the covariate information available, and can be found in Section 7.4.

Output from a pair of cluster analyses are represented by two sets of circular nodes on a cluster influence diagram; see Figure 8.1 for instance. These nodes are labelled by S-PLUS, with the most recently formed cluster being given a lower number as its label. The nodes were re-ordered to improve clarity, because it was easier to show the one-to-one relationships in the clusterings by separating them from the other clusters. Clusters of the subset data are shown by nodes on the left-hand-side, with the number of observations in each cluster placed alongside. The nodes on the right-hand-side represent output from clustering the superset data. The lines between these sets of nodes show how observations move from one cluster formation to another, with dotted, dashed, and solid lines representing the movement of one, two, and three or more observations respectively. The number of observations in each RHS cluster is shown to the right of these nodes as the sum of two numbers; the first is the number of observations that come from the subset data into the cluster, and the second is the number of new observations added to that cluster.

If two cluster analyses provide identical results, the cluster influence diagram would contain two equally sized sets of nodes, with lines joining one LHS cluster with one RHS cluster. The worst case is that no two observations clustered together on the LHS remain clustered together, in which case there would be a single dotted line for every observation in the data, with n dotted lines emanating from a LHS cluster with n observations. Having less 'clutter' on a cluster influence diagram indicates greater similarity of results. If subset data is representative of superset data, few distortions would be caused by the addition of observations to the set being clustered. The collection of more data should have confirmed findings, not significantly altered them, although some slight 'adjustments' might have occurred.

Figure 8.1 shows how twelve clusters formed by the 79 environments in Onion Data II using the available covariate information, alter their cluster membership when another ten environments were added to give Onion Data I. Reviewing the relevant dendrograms (Figures 7.16 and 7.17) and the details of the cluster members shown in Tables 7.10

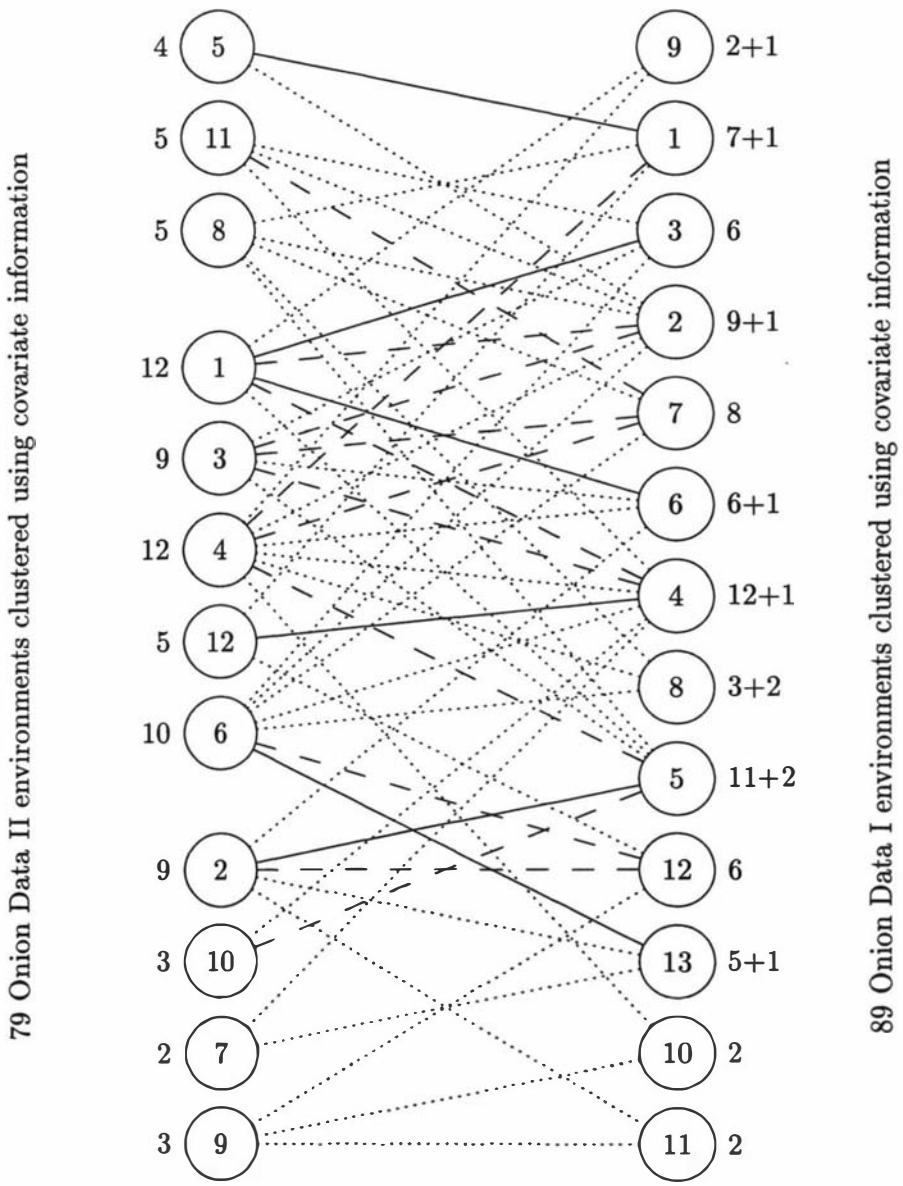


Figure 8.1: Cluster influence diagram for the dendrograms presented in Figures 7.16 (right) and 7.17 (left). In both cluster analyses available covariate information was used to form clusters, as described in Section 7.4. Left-hand-side clusters are of the 79 environments of Onion Data II with full covariate information, while right-hand-side clusters are for the 89 environments of Onion Data I.

and 7.11 shows a diverse range of geographical locations encompassed by these clusters. In the grand scheme of the trials programme these environments may be more similar than their locations suggest; names of countries do not provide enough information to understand the output presented in Figures 7.16 and 7.17.

This cluster influence diagram shows that every LHS cluster is affected by the addition of the ten environments. For example, LHS Cluster 4 is shattered when additional observations are introduced, with its members being rearranged into nine different RHS clusters.

Cluster influence diagrams are presented in Figures 8.2 and 8.3 to compare each of the dendrograms of Figures 7.16 and 7.17 to that of Figure 7.13. Recall that Figure 7.13 shows the dendrogram for all 101 environments with full covariate information. Figures 8.2 and 8.3 therefore show how the subsets of environments in Onion Data I and II reflect the wider set of environments tested by the Onion Trials Programme.

Two LHS clusters in Figure 8.2 are rearranged to form three RHS clusters as a result of the addition of a pair of environments to RHS Cluster 4. These two new environments are more similar to one of the LHS Cluster 10 than it was to the other environment in LHS Cluster 10. The flow-on effect of this is that the eight environments in LHS Cluster 1, the last step of clustering in that dendrogram, are now left as two RHS clusters (12 and 14). There are two clusters in this figure that remain unchanged. LHS Cluster 9 is also labelled Cluster 9 on the RHS, but LHS Cluster 8 is re-labelled RHS Cluster 6. Rather than identify each change, it is the number of changes or 'distortions' that occur in each cluster influence diagram that was of interest. The number of distortions in these cluster influence diagrams were counted, by adding the number of lines to the number of RHS clusters created entirely by added observations, and subtracting the number of LHS clusters. Recall that lines are counted rather than the number of observations because a distortion that affected two or more observations in conjunction, was counted in the same way as one that affected a single observation.

Figure 8.3 is more cluttered than Figure 8.2; it should be expected that the addition of 22 observations has a greater impact than the addition of 12 observations. Figures 8.1 to 8.3 have 42, 31, and 44 distortions respectively. Measuring the probability that pairs of observations are treated the same by a pair of cluster analyses in the manner of Rand (1971), gives scores of 0.840, 0.881, and 0.853 respectively. While these scores seem high, they are in fact quite predictable given the findings of simulations undertaken to understand the range of possible R_g scores.

A note on the range of R_g

It was discovered that the R_g scores observed throughout this investigation fell into a range that was considerably smaller than the zero to one range that is implied by the definition. It has been noted that the upper bound of one occurs when the clusterings

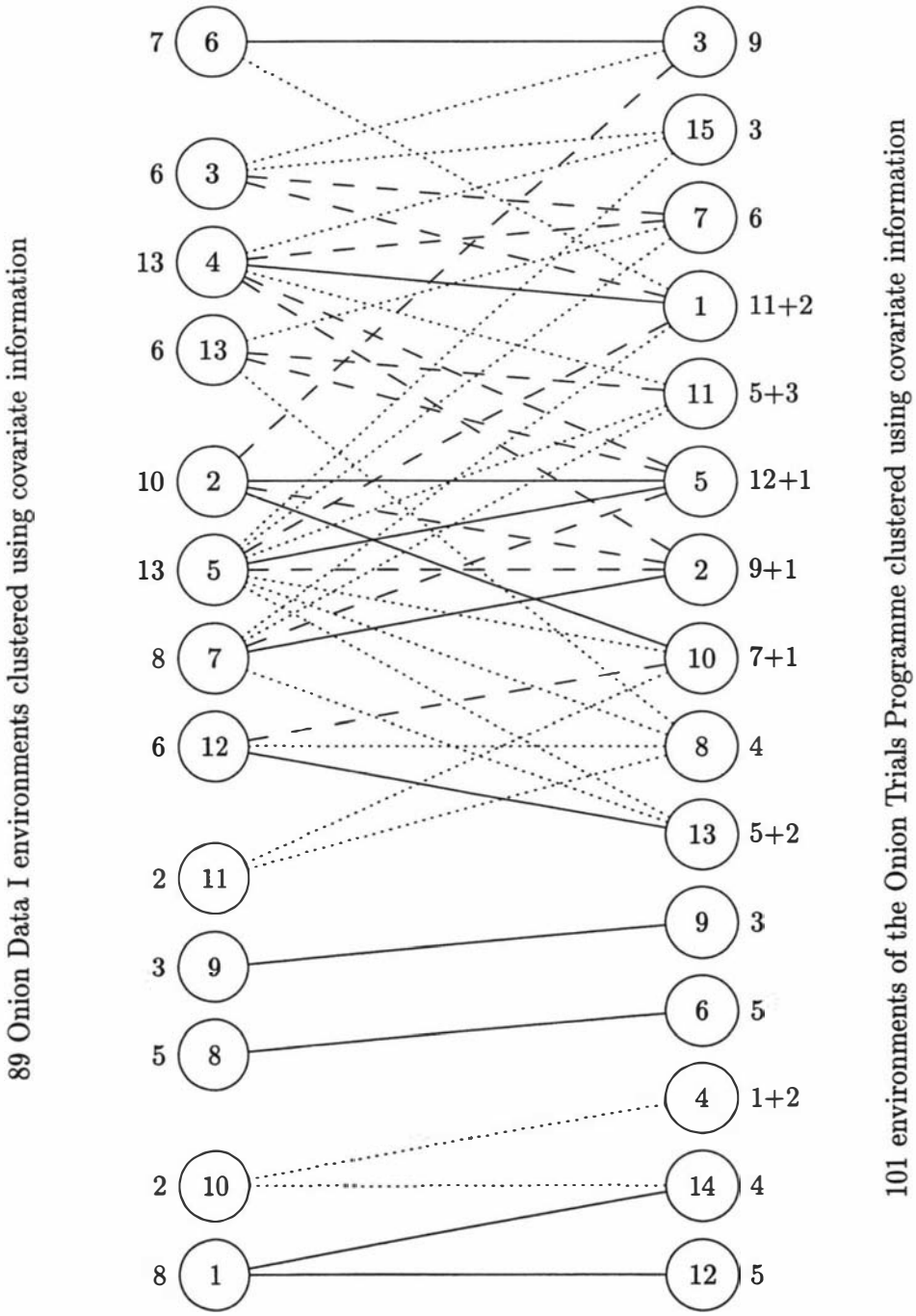


Figure 8.2: Cluster influence diagram for the dendrograms presented in Figures 7.13 (right) and 7.16 (left). Covariate information was used to form the clusters in both cases, as described in Section 7.4. Left-hand-side clusters are formed from the 89 environments of Onion Data I, while all 101 environments from the Onion Trials Programme with full covariate information were used to form right-hand-side clusters.

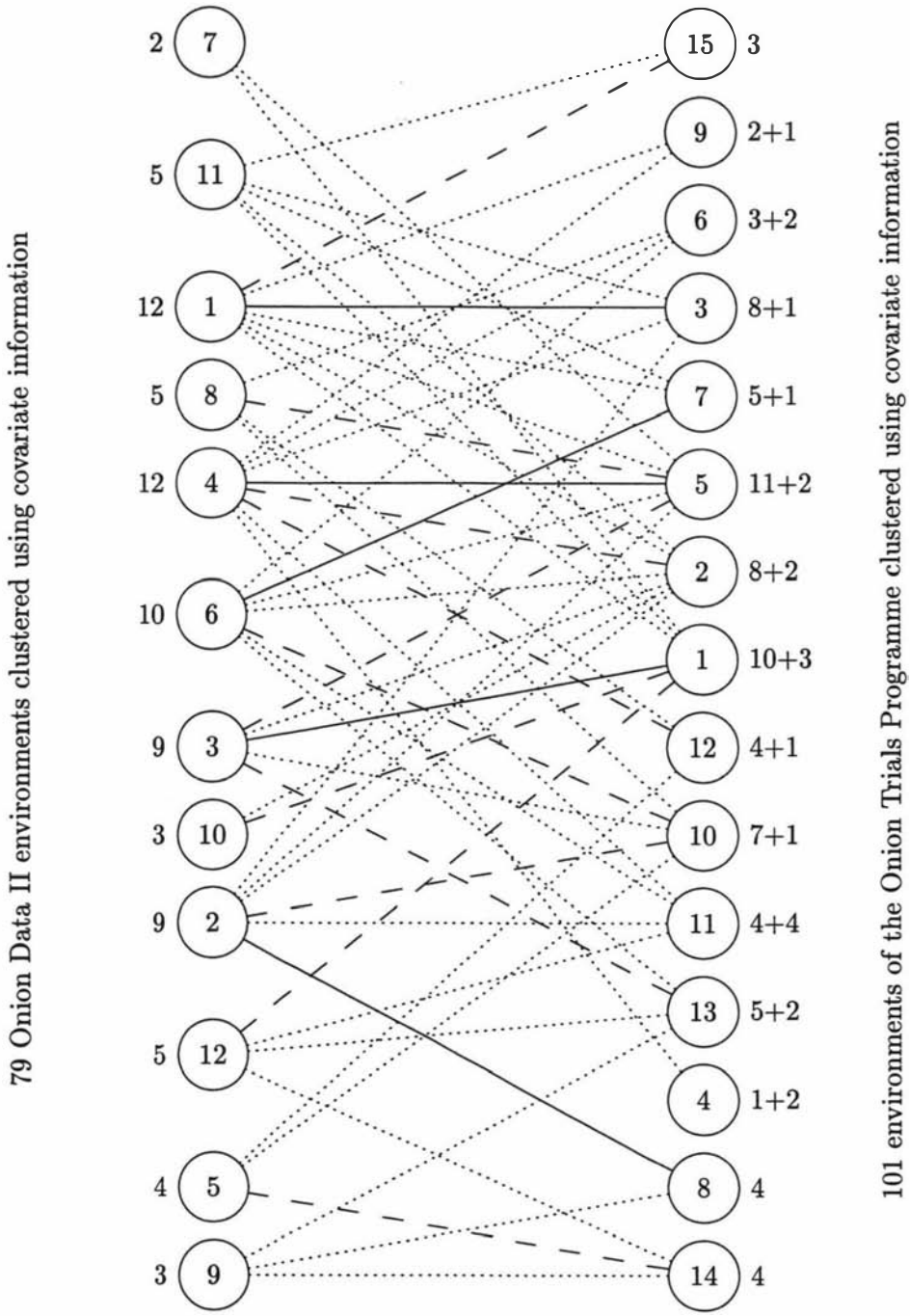


Figure 8.3: Cluster influence diagram for the dendrograms presented in Figures 7.13 (right) and 7.17 (left). Section 7.4 presents these cluster analyses which use the available covariate information in their construction. Left-hand-side clusters are of the 79 environments in Onion Data II with full covariate information, while all 101 environments from the Onion Trials Programme form right-hand-side clusters.

under comparison are identical, but this is only possible when the two clusterings have the same number of clusters. The only possible way to get an R_g of zero is to have a single cluster of n observations split into n singleton clusters. The majority of comparisons between clusterings in general, and all of those found in this chapter, fall well within these extremes.

Rand (1971) gave examples of how R_g for k equally sized clusters would be affected by a small selection of changes to the clustering, but this clustering was felt restrictive and unlikely in practice. Simulations were, therefore, carried out to better understand the behaviour of R_g, and additional theoretical values of R_g were obtained by considering the minimal alterations possible for a certain clustering.

Approximately 100 simulations were made for every combination of the number of clusters between ten and thirty. Cluster memberships were randomly assigned using discrete uniform and triangular distributions. The number of observations used was allowed to vary over the range 60 to 100, in steps of 10, in order to cover the range of scenarios found in this chapter.

The findings from simulations included:

1. The greater the number of clusters in the smaller clustering, the higher the minimum observed R_g score.

2. The smaller the difference in the number of clusters between clusterings, the greater the maximum observed R_g score.

3. The greater the number of observations, the greater the R_g scores.

4. The R_g scores found for the uniformly distributed cluster memberships were in a smaller range than those for triangular distributed cluster memberships. Maxima were lower, and minima were higher.

A result of the first and second observations was that a pair of clusterings that had widely different numbers of clusters had a much smaller range of observed R_g scores. In general the minimum R_g scores were greater than 0.80, but the maxima were not as close to 1 as expected. The pattern of observed minimum R_g scores was also much smoother than that for the observed maxima.

The smallest possible change in R_g occurs when singleton clusters merge, or when a cluster splits into singletons. This results in an increase or decrease in the number of clusters of k_s. If there were n_t observations in all, the upper bound of R_g would become

$$R_g = 1 - \frac{k_s\,(k_s + 1)}{n_t\,(n_t - 1)} \tag{8.1}$$

This simplifies an expression given by Rand (1971). Theoretical minimum R_g lower bounds were also found, but are not presented here because the likelihood of their occurrence was

felt to be too small. These minima relied on a clustering of observations into $k-2$ singleton clusters and two other clusters having the remaining $n_t - k + 2$ observations.

In general, the observed R_g scores from the various simulations covered similar ranges, but depended on the number of clusters in each clustering and the dispersion of the cluster memberships. R_g scores presented in this chapter could be compared to one another, but should not be relied upon to provide a truly meaningful indication of the similarity of pairs of clusterings.

On the whole, it can be said that in terms of environments with full covariate information, the environments of Onion Data I and II do not represent the environments of the entire trials programme. In the next section these summaries are further quantified by use of tools that measure the strength of the relationship between two nominal variables.

8.3 Gauging strength of the relationship between two cluster analyses

Output from cluster analyses usually includes the memberships of each cluster formed. Clusters are given integer valued labels by S-PLUS, which order the size of the multi-dimensional space spanned by clusters. The comparison of two sets of these nominal labels needed to be quantified to back up the graphical tool presented in the previous section.

Examination of a relationship between two nominal variables is limited to a χ^2-test for independence in many statistical packages. SPSS and SAS software offer users additional options that measure the strength of this relationship.

The Goodman and Kruskal λ statistic was developed to show the ability to accurately predict the value of one nominal variable having knowledge of another (Goodman and Kruskal, 1954). In this case it is known as asymmetric and is given as

$$\lambda = \frac{\sum_{i=1}^{I} \max_k \{x_{ik}\} - \max_i \left\{ \sum_{k=1}^{K} x_{ik} \right\}}{\sum_{i=1}^{I}\sum_{k=1}^{K} x_{ik} - \max_i \left\{ \sum_{k=1}^{K} x_{ik} \right\}} \tag{8.2}$$

where x_{ik} is the number of observations that are categorized by the ith and kth levels of the dependent and independent variables respectively. This measure is based on the idea of reducing the error rate of making a prediction of the dependent variable which, given no information, would be predicted as the most common outcome of that variable for each level of the independent variable. This is known as 'proportional reduction in error' because it minimizes the number of wrong guesses, given knowledge of the independent variable. If no directional relationship is known to exist between the variables (8.2) is

combined with its complement to give the symmetric form

$$\lambda = \frac{\sum_{i=1}^{I} \max_k \{x_{ik}\} + \sum_{k=1}^{K} \max_i \{x_{ik}\} - \max_i \left\{ \sum_{k=1}^{K} x_{ik} \right\} - \max_k \left\{ \sum_{i=1}^{I} x_{ik} \right\}}{2 \sum_{i=1}^{I} \sum_{k=1}^{K} x_{ik} - \max_i \left\{ \sum_{k=1}^{K} x_{ik} \right\} - \max_k \left\{ \sum_{i=1}^{I} x_{ik} \right\}} \tag{8.3}$$

Note that the symmetric form is not the simple average of asymmetric complements. For both cases, $0 \leqslant \lambda \leqslant 1$, with perfect predictability reflected by $\lambda = 1$.

When applying the Goodman and Kruskal λ to the outcome of a pair of cluster analyses, x_{ik} refers to the number of observations clustered into the ith cluster by one clustering, and the kth cluster of the other clustering. As in the previous section, the ability of a subset of data to provide the same information as its superset can be established using the asymmetric measure. The left-hand-sides of these cluster influence diagrams show the clustering of the subset and are therefore considered the independent variable in (8.2) above. In some situations where no directional relationship can be justified, asymmetric λ cannot be used. In such cases, the symmetric statistic given in (8.3) will be used to compare categorization of observations by two cluster analyses.

The major criticism of the λ measure is that it concentrates on the mode, ignoring variability within the joint distribution. Theil's uncertainty coefficient is similar to λ in that it looks for a reduction in the error of prediction. More specifically it uses the notion of entropy of nominal variables, instead of variance for continuous variables, as the quantity to be reduced via knowledge of another nominal variable (Theil, 1972). The asymmetric uncertainty coefficient is

$$U = \frac{\sum_{i=1}^{I} \pi_{i\cdot} \ln(\pi_{i\cdot}) + \sum_{i=1}^{I} \sum_{k=1}^{K} \pi_{ik} \ln\left(\frac{\pi_{ik}}{\pi_{\cdot k}}\right)}{\sum_{i=1}^{I} \pi_{i\cdot} \ln(\pi_{i\cdot})} \tag{8.4}$$

where π_{ik} is the proportion of data that are categorized by the ith and kth levels of the dependent and independent variables respectively; marginal distributions are expressed using $\pi_{i\cdot}$ and $\pi_{\cdot k}$. The symmetric version of (8.4) is

$$U = 2\frac{\sum_{i=1}^{I} \pi_{i\cdot} \ln(\pi_{i\cdot}) + \sum_{k=1}^{K} \pi_{\cdot k} \ln(\pi_{\cdot k}) + \sum_{i=1}^{I} \sum_{k=1}^{K} \pi_{ik} \ln(\pi_{ik})}{\sum_{k=1}^{K} \pi_{\cdot k} \ln(\pi_{\cdot k}) + \sum_{i=1}^{I} \pi_{i\cdot} \ln(\pi_{i\cdot})} \tag{8.5}$$

As for λ in (8.2) and (8.3) above, $0 \leqslant U \leqslant 1$, and $U = 1$ indicates perfect predictability. Theil (1972) proved that the numerator of (8.4) is half the numerator of (8.5); this proof

clarified the confusion caused by changes in notation across sources for this measure.

When λ and U coefficients are calculated for two separate cluster analyses that have different numbers of clusters, the range of possible values alters. A coefficient of one is only possible for the asymmetric cases when the dependent clustering has no more clusters than the independent clustering. It follows that symmetric λ and U can only have a value of one if the cluster analyses have the same number of clusters.

Table 8.1 shows the strength of commonality between the clusterings given in Figures 7.13, 7.16, and 7.17 presented in Section 7.4. These values were found using SPSS, as it offers the user asymptotic estimates of standard errors, for both λ and U.

In each of these examples the asymmetric λ and U coefficients were less than their symmetric counterparts. The corresponding cluster influence diagrams show that the subset data formed a smaller number of clusters than the superset data. In unbalanced situations like this, the ability to predict the value of the nominal variable with more levels will (in general) be harder than prediction of the variable with fewer levels.

Standard errors presented in Table 8.1 show that null hypotheses that $\lambda, U = 0$ would be rejected in every case. It is reasonable to expect the environments from Onion Data II to represent the environments of Onion Data I, which in turn should represent the total set of environments. The environments of Onion Data II should represent the environments of Onion Data I better than they represent the total set of environments if adding more data increases the number of distortions. Quantification of these relationships suggests that the environments of Onion Data II are marginally better at representing the overall set than they are at representing the environments of Onion Data I. This is refuted by considering the difference in the values of λ and U in light of their standard errors. We may argue however, that Onion Data I's environments better represent the overall set of environments than do the environments of Onion Data II, based on the difference in U

Cluster influence diagram	Dendrograms in Figures	Asymmetric		Symmetric	
		λ	U	λ	U
8.1	7.16 and 7.17	0.224 (0.073)	0.366 (0.028)	0.239 (0.056)	0.372 (0.027)
	79 Onion Data II environments and 89 Onion Data I environments				
8.2	7.13 and 7.16	0.390 (0.063)	0.534 (0.033)	0.412 (0.059)	0.549 (0.032)
	89 Onion Data I environments and all 123 environments				
8.3	7.13 and 7.17	0.250 (0.069)	0.391 (0.025)	0.281 (0.056)	0.409 (0.025)
	79 Onion Data II environments and all 123 environments				

Table 8.1: λ and U coefficients corresponding to cluster influence diagrams presented in Figures 8.1, 8.2, and 8.3; which compare environment clustering based on covariate information for three sets of environments presented in Section 7.4. Formulae for asymmetric and symmetric cases of λ and U are presented in (8.2) to (8.5). Asymptotic standard errors are provided in brackets.

values.

8.4 The consistency of available data

In Section 8.2, the three different cluster analyses based on covariate information were compared using cluster influence diagrams. As environments were added the cluster memberships altered. It was also evident that there were alterations in the cluster groupings when clusters were formed using available yield data in Section 7.2. This section determines the consistency of mega-environments based on available yield data using cluster influence diagrams and the numerical summary statistics introduced in preceding sections.

The cluster influence diagram presented in Figure 8.4 compares mega-environments found using the entire data set and Onion Data I (Figures 7.3 and 7.4, respectively), based on available yield data. It therefore shows the effect of adding more genotypes and environments to the data that were clustered; this means that not only were observations added to the data, but that variables those observations were measured over were also added. Adding both genotype and environment data triggers a discussion point that will be addressed later in this section, namely, "Why not use all the genotype information available to cluster particular sets of environments?".

Of particular interest in this diagram are the three new clusters formed when all 123 environments of the trials programme were clustered together. These clusters were formed from four environments that were seemingly quite distinct from the other 119 environments. RHS Cluster 2 is formed by two environments that are not in Onion Data I, two environments from LHS Cluster 25, and another from LHS Cluster 7. The two remaining environments in LHS Cluster 7 now form RHS Cluster 17. Of the seven clusters that remained intact, only one had any environments added into it, and there was only one instance where a LHS cluster was split cleanly into two or more RHS clusters.

This cluster influence diagram has 24 distortions, which is comparatively low given that there are 29 LHS clusters and Figures 8.1, 8.2, and 8.3 had many more distortions than LHS clusters.

Table 8.2 shows λ and U values for the comparisons made in this section, while Table 8.8 shows these numerical summary statistics for all cluster influence diagrams in this chapter. Both asymmetric and symmetric λ and U values are given in Table 8.8 for ease of comparison of these scores. The asymmetric forms of λ and U are the appropriate measures in this instance, so the symmetric scores have not been presented in Table 8.2. Asymmetric λ for this pair of cluster analyses was comparatively high at 0.691, and asymmetric U was also quite high at 0.834; standard errors provided indicate that these relationships are significant. The values for Rand's R_g presented in Table 8.2 are better than those for the clusterings based on covariate information presented previously.

The consistency of results from Onion Data II to the other cluster analyses based on

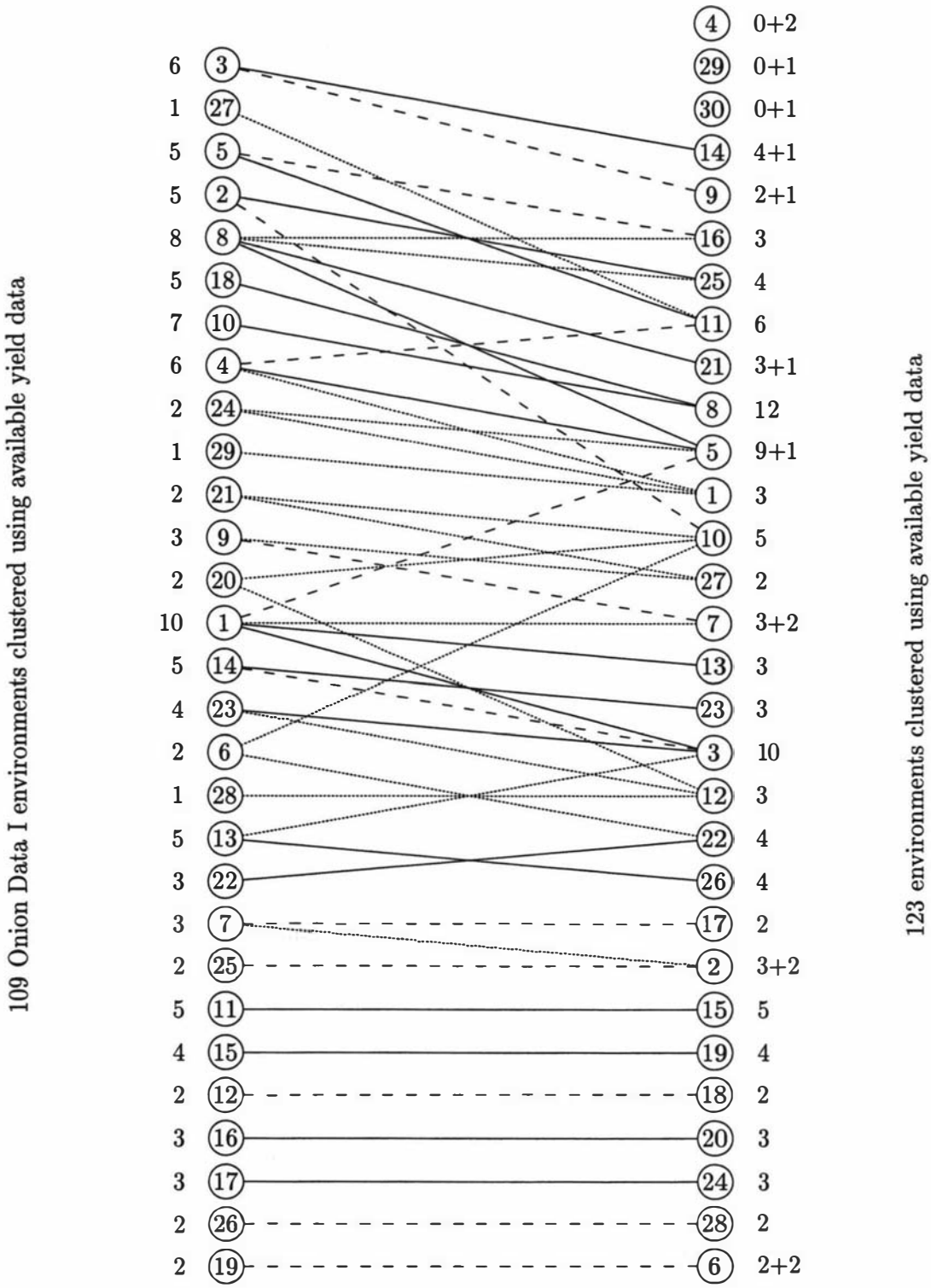


Figure 8.4: Cluster influence diagram for the dendrograms presented in Figures 7.3 (right) and 7.4 (left). Left-hand-side clusters are formed using available yield data from Onion Data I, while right-hand-side clusters are formed using the entire Onion Trials Programme data. Details of cluster formation can be found in Section 7.2.

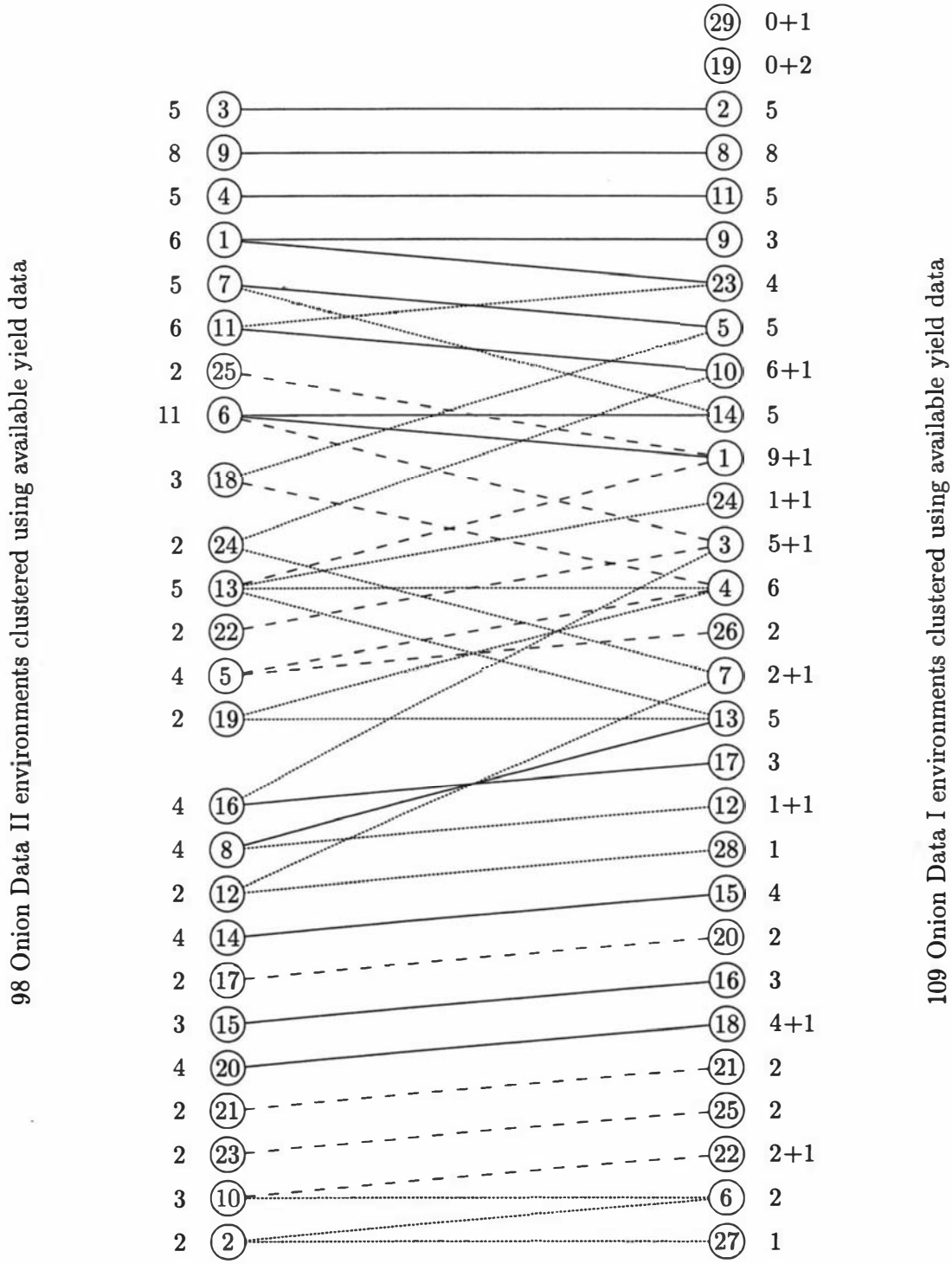


Figure 8.5: Cluster influence diagram for the dendrograms presented in Figures 7.4 (right) and 7.5 (left). Left-hand-side clusters are formed using available yield data from Onion Data II, while right-hand-side clusters are formed using Onion Data I. Full details of these clusterings can be found in Section 7.2.

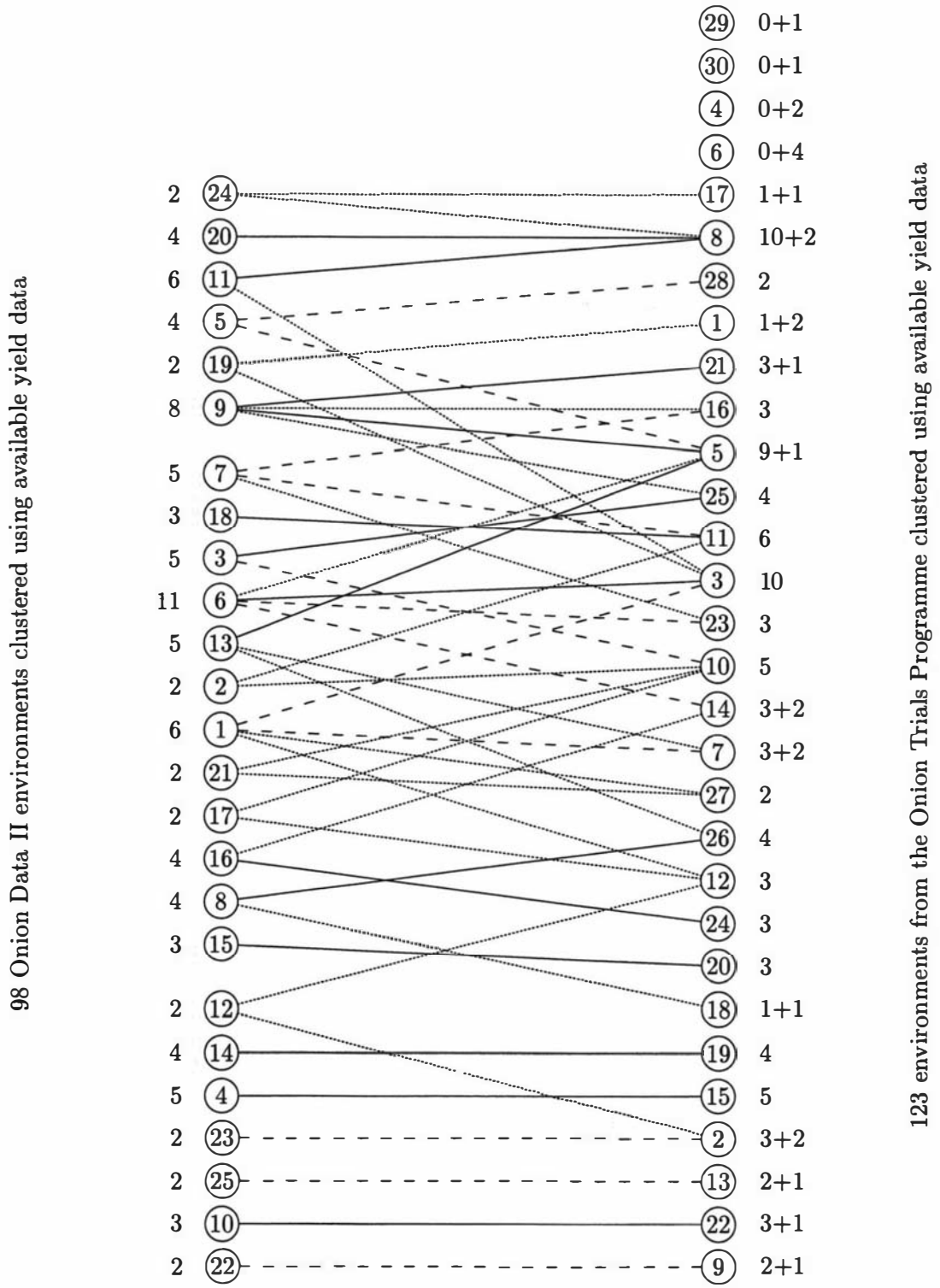


Figure 8.6: Cluster influence diagram for the dendrograms presented in Figures 7.3 (right) and 7.5 (left). Left-hand-side clusters were formed using available yield data from Onion Data II, while right-hand-side clusters were formed using the entire Onion Trials Programme data. Full details of these clusterings is presented in Section 7.2.

Cluster influence diagram	Dendrograms in Figures	Asymmetric λ	Asymmetric U	Number of distortions	No. of unchanged clusters	Rand's R_g
8.4	7.3 and 7.4	0.691(0.047)	0.834(0.017)	24	7	0.958
	109 Onion Data I environments and all 123 environments					
8.5	7.4 and 7.5	0.742 (0.046)	0.860 (0.016)	17	9	0.968
	98 Onion Data II environments and 109 Onion Data I environments					
8.6	7.3 and 7.5	0.636(0.053)	0.793(0.020)	28	6	0.949
	98 Onion Data II environments and all 123 environments					

Table 8.2: Numerical summary statistics corresponding to cluster influence diagrams presented in Figures 8.4, 8.5, and 8.6; which compare clustering based on available yield data for three sets of environments presented in Section 7.2. Formulae for asymmetric λ and U are presented in (8.2) and (8.4). Asymptotic standard errors are provided in brackets. The number of distortions, the number of unchanged clusters, and Rand's R_g are also presented.

available yield data is now discussed. The impact of adding data from more environments, and therefore genotypes, to that of Onion Data II, is presented in Figures 8.5 and 8.6. Comparing these cluster influence diagrams to Figure 8.4 shows that there was greater environment cluster consistency between results from Onion Data II to Onion Data I, than there was from Onion Data I to the entire set of data from the Onion Trials Programme; and that this was greater than the environment cluster consistency between Onion Data II and the entire set of data.

Of the nine clusters that stayed intact in Figure 8.5, only three of them stayed intact in Figure 8.6. A more complex path through the cluster analyses for Figure 8.5's LHS Cluster 10 can be followed. This cluster had three observations initially, which separated into RHS Clusters 22 and 6. Looking at Figure 8.4, these observations (now LHS Clusters 22 and 6) came back together to form RHS Cluster 22, or as can be seen in Figure 8.6 the original cluster stayed intact to form RHS Cluster 22.

When counting distortions, it is clear that Figure 8.5 (17 distortions) is less distorted than Figure 8.6 (28 distortions). The relevant asymmetric λ and U values from Table 8.2 concur with the above findings, as $\lambda\{\text{Figure 8.5}\} > \lambda\{\text{Figure 8.4}\} > \lambda\{\text{Figure 8.6}\}$ and $U\{\text{Figure 8.5}\} > U\{\text{Figure 8.4}\} > U\{\text{Figure 8.6}\}$. These findings should present no surprise as we know that Onion Data I and II are more closely related in terms of the data they contain than Onion Data II is to the entire set of data from the Onion Trials Programme. The consistency of mega-environments created using available yield data is greater than the consistency of clusterings based on the covariate information presented previously. The next section investigates the consistency of clustering based on imputed values, including their relationsuip to clustering based on the available yield data.

Choice of the data used in clustering

Now to answer the question, "Why not use all the genotype information available to cluster particular sets of environments?". Recall that data from genotypes that were not grown in the minimum number of environments were discarded as the subset data of Onion Data I and II were created. One answer is that there is insufficient genotype data to standardize their yields. This can be disregarded, however, as only two points are needed to do the standardization. In this instance this genotype's data would only affect the comparison of one pair of environments. Looking a little closer, however, may alter this finding slightly. Take for example, two environments that are effectively the same, and therefore have a distance near zero when found using the sparse data from well-represented genotypes. Inclusion of standardized data from a genotype that is only grown in those two environments will inflate the distance measure between them, and increase the likelihood that they will not cluster together so soon in the analysis.

The desire to use as much information as possible to make every comparison, hoping to improve accuracy, is now seen to be short-sighted when using standardized data. However, the benefits of having more data on hand cannot be disregarded so simply. How much data is needed from each genotype to ensure that observed yields reflect its potential? The decision about the threshold that will turn less valued data into an integral part of the data must be made with care. The presentation above (Figure 8.5) showed the effects of changing this threshold from seven to eight, and while consistency of outcomes was high, there is still a noticeable amount of distortion in results. The best option is probably to apply the adage, 'Prevention is better than cure'. Chapter 10 presents methods for reducing the impact of these issues.

8.5 The consistency of imputed data

In this section, three aspects of consistency are considered:

1. Does the clustering of environments alter once two-stage imputation has been used to complete the G×E matrix?

2. Is two-stage imputation of Onion Data I and II providing the same outcome in terms of environment clusters?

3. Is two-stage imputation of Onion Data I and II providing the same outcome in terms of genotype clusters?

In Section 6.5 this last comparison was given without the tools of this chapter, and can now be re-visited in a more thorough manner. The second and third questions are complements of one another; the complement of the first question is left to Section 8.7, where the effects of two-stage and nearest cluster imputation are compared.

Cluster influence diagram	Dendrograms in Figures	Asymmetric λ	Asymmetric U	Number of distortions	No. of unchanged clusters	Rand's R_g
8.7	7.4 and 7.8	0.313 (0.047)	0.597 (0.019)	68	0	0.942
109 Onion Data I environments clustered using sparse and fully imputed yields						
8.8	7.5 and 7.9	0.287 (0.050)	0.572 (0.020)	59	0	0.918
98 Onion Data II environments clustered using sparse and fully imputed yields						

Table 8.3: Numerical summary statistics corresponding to cluster influence diagrams presented in Figures 8.7 and 8.8 which compare clustering based on sparse and two-stage imputed yield data for environments of Onion Data I and II. Formulae for asymmetric λ and U are presented in (8.2) and (8.4). Asymptotic standard errors are provided in brackets. The number of distortions, the number of unchanged clusters, and Rand's R_g measure are also presented.

Figure 8.7 shows the impact of using the fully imputed yield data for Onion Data I, over the sparse data used to create Figure 7.4. This cluster influence diagram has 68 distortions, and asymmetric λ and U of 0.313 and 0.597 respectively (see Table 8.3 for details). No LHS clusters stayed intact in this cluster influence diagram. Clearly there is little consistency between the two sets of environmental clusters found using sparse and two-stage imputed data. Is this a problem? In the sparse data, there were environments that may be similar in terms of the conditions that can be found there, but that had little commonality of test genotype sets. These environments would be forced apart in the LHS clustering, while it is hoped that by successfully imputing the missing data that they will now be deemed similar. This theory is not easily tested; certainly some of the simulation testing of Section 6.4 could be repeated with the outcome of post imputation clustering compared to complete data based clustering. The best that such testing could show is that the two-stage imputation process may succeed, but there is no reason to suggest that it will happen every time two-stage imputation is applied. In the case of the data arising from the Onion Trials Programme it is uncertain which of the two cluster analyses is superior. There is therefore a need to compare clustering of environments and then genotypes between Onion Data I and II, presented later in this section.

The impact of two-stage imputation on the clustering of Onion Data II environments is presented in Figure 8.8. It shows similar results to Figure 8.7, having 59 distortions corresponding λ of 0.287, and a U value of 0.572. Once again no LHS clusters stayed intact in this cluster influence diagram. While these λ's and U's are low, their standard errors are not high enough to suggest that there is no relationship between the clustering of environments in either Onion Data I or II.

A crucial question to ask of two-stage imputation is, "Are imputation results determined by the subset of data used to find them?". This question was answered in several ways in Section 6.5, but the additional tools presented in this chapter are now available.

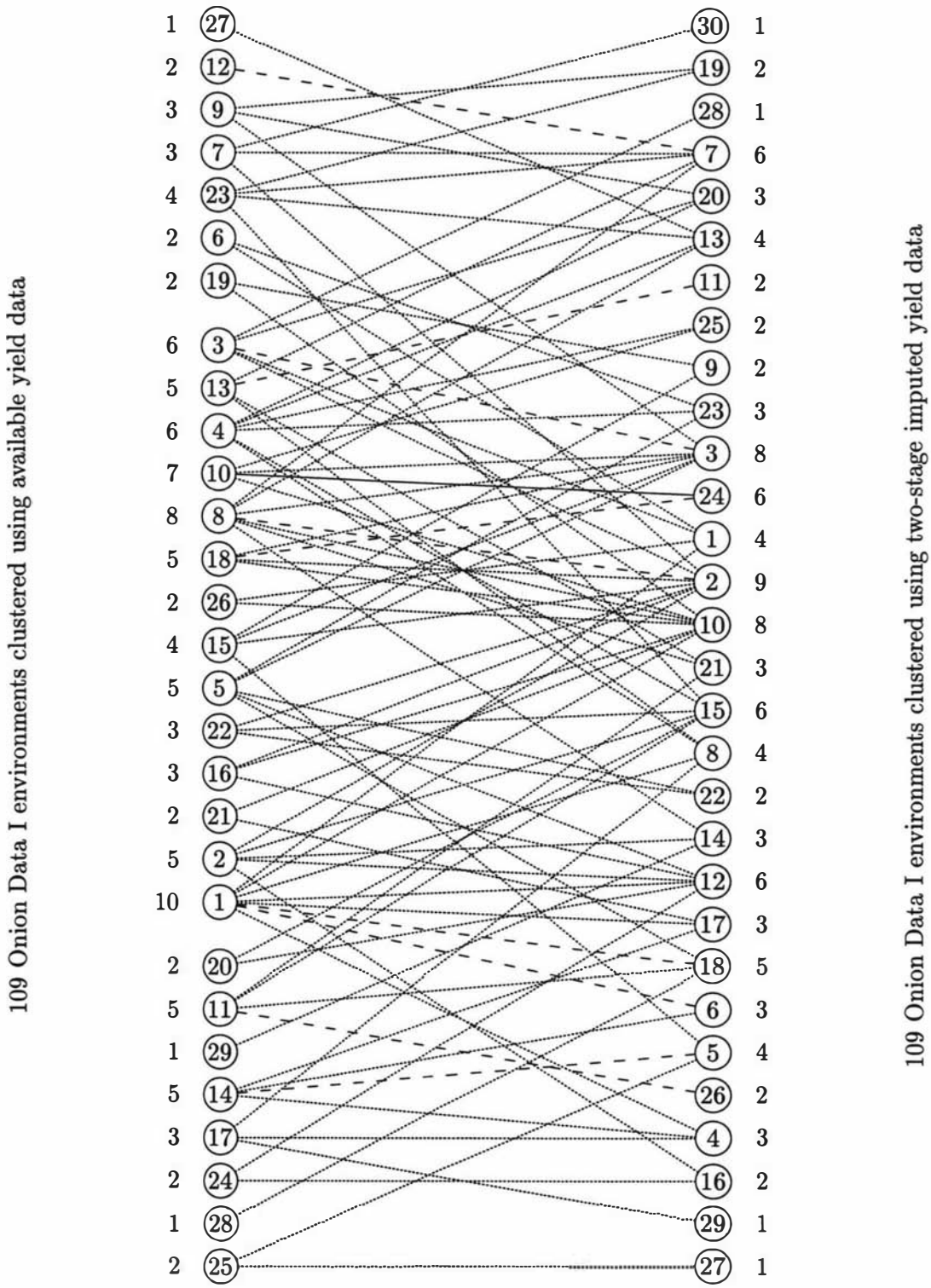


Figure 8.7: Cluster influence diagram for the dendrograms presented in Figures 7.4 (left) and 7.8 (right). First stage clustering (as described in Section 5.2) of Onion Data I environments was performed in both cases, with sparse data used to form left-hand-side clusters and two-stage imputed data used to form right-hand-side clusters.

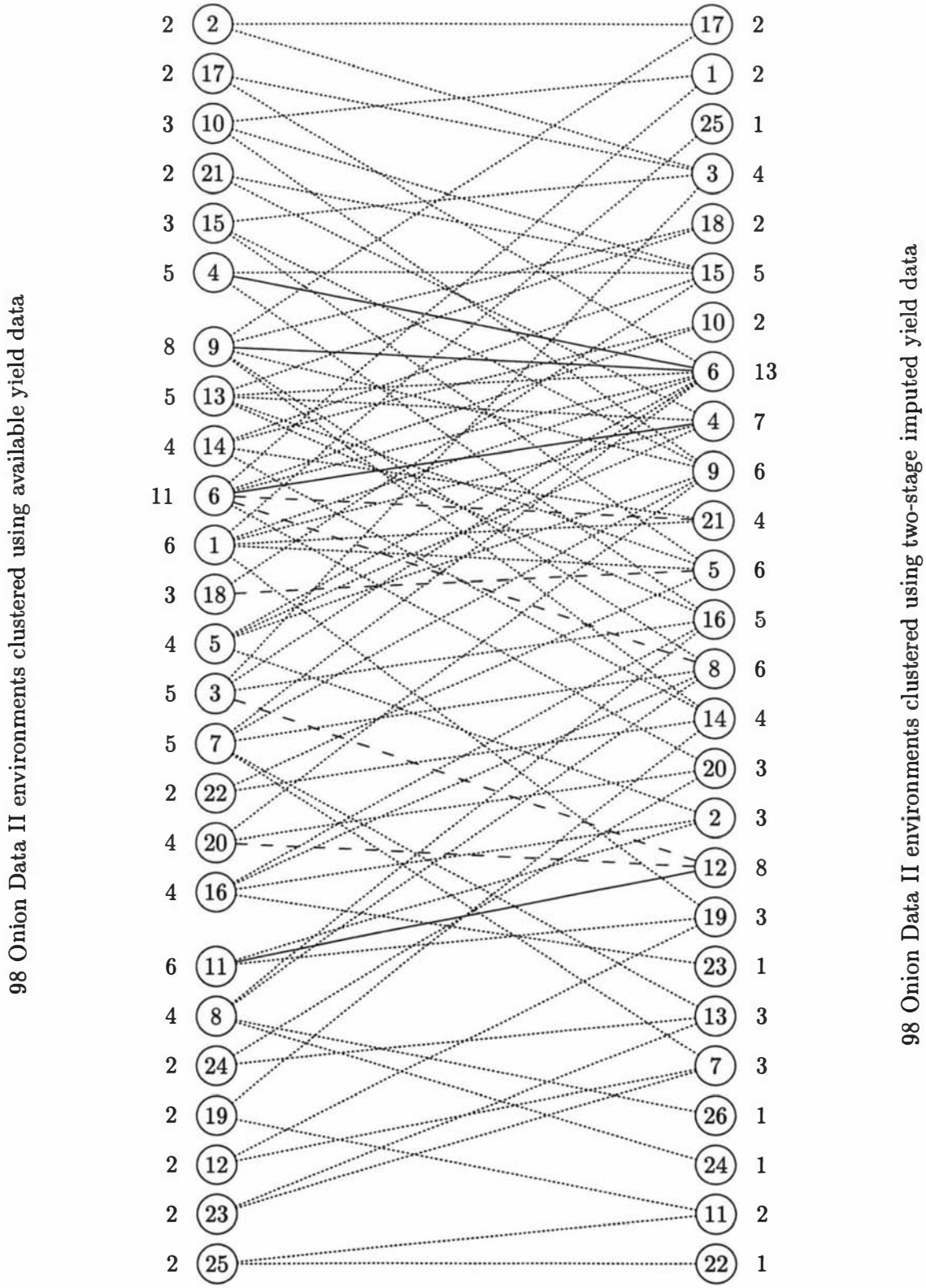


Figure 8.8: Cluster influence diagram for the dendrograms presented in Figures 7.5 (left) and 7.9 (right). First stage clustering (as described in Section 5.2) of Onion Data II environments has been performed in both cases, with sparse data used to form left-hand-side clusters and two-stage imputed data used to form right-hand-side clusters.

Cluster influence diagram	Dendrograms in Figures	Symmetric λ	Symmetric U	Number of distortions	No. of unchanged clusters	Rand's R_g
8.9	7.8 and 7.9	0.397 (0.044)	0.649 (0.019)	53	1	0.931
98 Onion Data II and 109 Onion Data II environments, based on two-stage imputed yields						
8.10	6.7 and 6.8	0.639 (0.050)	0.746 (0.022)	22	0	0.953
87 Onion Data II and 104 Onion Data II genotypes, based on two-stage imputed yields						

Table 8.4: Numerical summary statistics corresponding to cluster influence diagrams presented in Figures 8.9 and 8.10 which compare clustering based on imputed yields of Onion Data I and II in terms of environments and then genotypes. Formulae for symmetric λ and U are presented in (8.3) and (8.5). Asymptotic standard errors are provided in brackets. The number of distortions, the number of unchanged clusters, and Rand's R_g measure are also presented.

Figure 8.9 shows the alteration of environment cluster memberships found in Figures 7.8 and 7.9 which were based on two-stage yield data of Onion Data I and II respectively. This cluster influence diagram appears quite cluttered and has 53 distortions, although one LHS cluster remained intact.

In this instance the symmetric versions of λ and U were used because the focus is now on similarity of outcomes as against one set of outcomes being used to predict another. Table 8.4 shows these values to be 0.397 and 0.649 respectively, which are comparatively high. Table 8.8 lists all symmetric λ and U values for the comparisons made in this chapter. Only those for similarity of cluster memberships based on available yield data in Section 8.4 were higher. Rand's R_g of 0.931 is in the middle of the range of those scores presented in Table 8.8 .

As noted previously, this investigation of consistency can be done using clusters of genotypes instead of environments. Figure 8.10 presents the cluster influence diagram for genotype clustering using two-stage imputed data from Onion Data I and II. This cluster influence diagram appears relatively uncluttered, but has more distortions (22) than Figure 8.5. Symmetric λ and U in this example are 0.628 and 0.746 respectively, while it has a Rand's R_g of 0.953, indicating that the consistency of clustering of genotypes is much higher than that of the environment clustering presented above, even though no LHS clusters remained intact.

On the basis of the comparisons made in this chapter, the post-imputation cluster analyses provide consistent sets of mega-environments as well as consistent sets of genotypes, thus supporting the findings of Section 6.5. Further investigations into the effect of imputing G×E matrices, and those arising from the Onion Trials Programme in particular, will be presented in Sections 8.7 and 8.8. The next section investigates the relationships between clusterings based on covariate information and those based on available and imputed yield data.

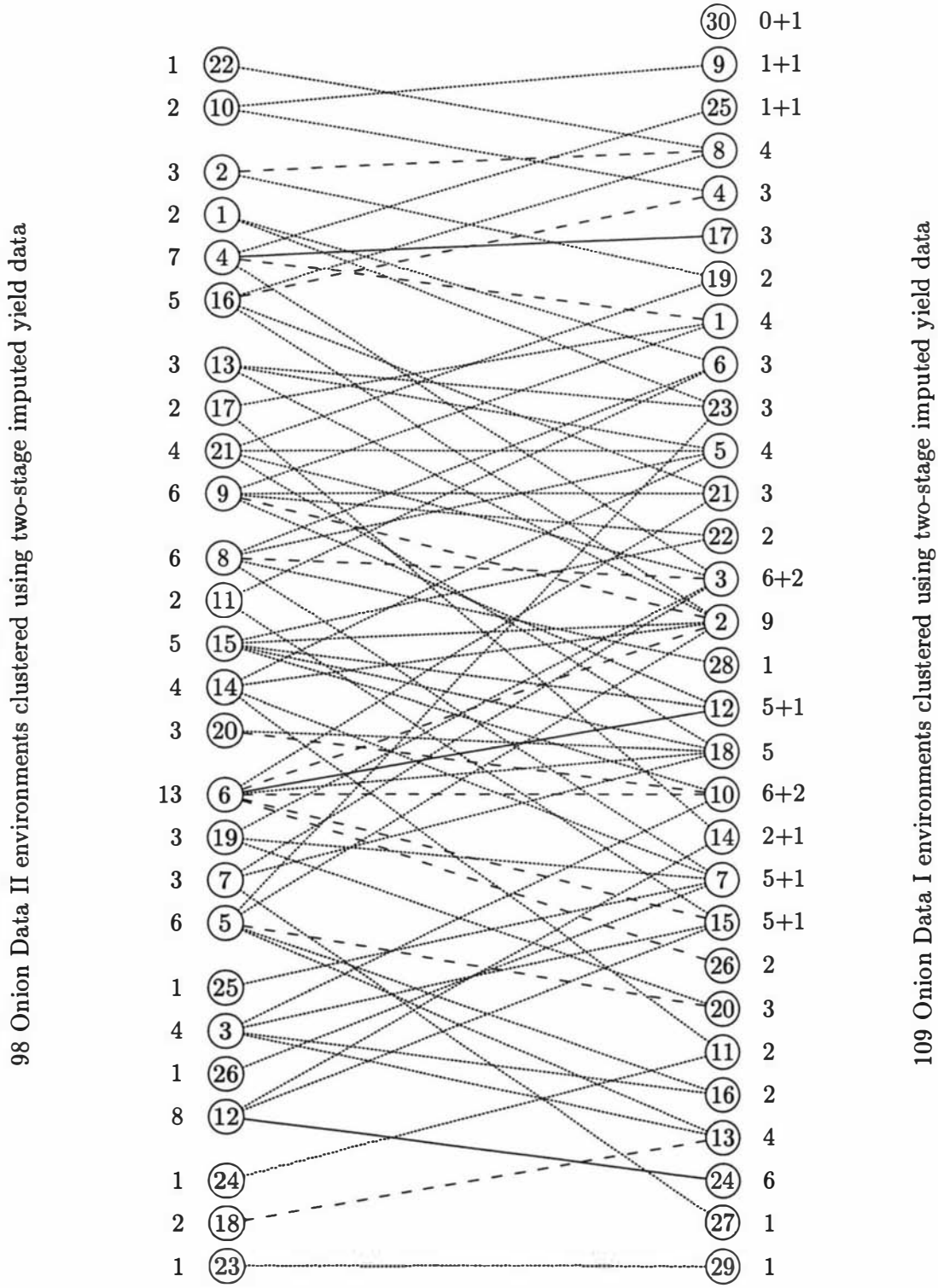


Figure 8.9: Cluster influence diagram for the dendrograms presented in Figures 7.8 (right) and 7.9 (left). Two-stage imputed data has been used to cluster environments, as described in Section 7.3, in both cases. Left-hand-side clusters are formed using the 98 environments of Onion Data II, while the 109 environments of Onion Data I have been used to form right-hand-side clusters.

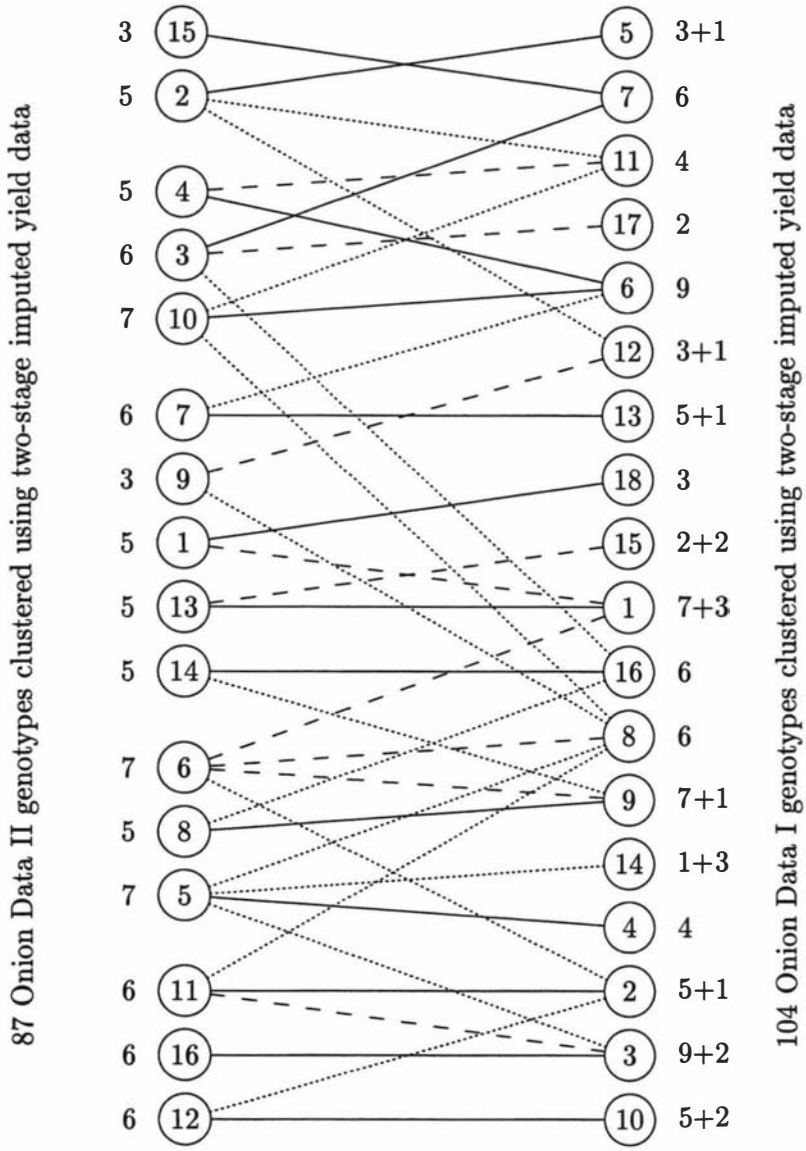


Figure 8.10: Cluster influence diagram for the dendrograms presented in Figures 6.7 (right) and 6.8 (left). Two-stage imputed data from Onion Data II (and Onion Data I) have been used to form left-hand-side (and right-hand-side) clusters of genotypes.

8.6 The ability of yield data to reflect covariate information

Many authors have used yield data from experiments to group environments (Abou-El-Fittouh *et al.*, 1969; Byth *et al.*, 1976; Ivory *et al.*, 1991; Lin and Morrison, 1992). This implies that genotype yields summarize the impact of environmental growing conditions, or more specifically, that G×E interaction patterns can be used to distinguish environments. If this is the case, and covariates used in Section 7.4 are the only influencing factors, dendrograms based on covariate information and on available yield data will be similar in appearance. Certainly a cluster influence diagram would show correspondence between resulting sets of mega-environments. Clustering environments based on covariate data (Section 7.4) was initially compared to clustering based on available yield data (Section 7.2) using all environments of the Onion Trials Programme, followed by the environments of Onion Data I and II; subsequently clustering of environments based on two-stage imputed data for Onion Data I and II (Section 7.3) was compared to the clustering based on covariates.

The methods employed to create Figures 7.3 and 7.13 differ in two respects. The clustering of 123 environments in Figure 7.3 uses a different set of variables (genotype yields) to that of Figure 7.13 which uses covariate information. Another 22 environments were also added to the data. The two influencing factors (variables and number of environments) can only be separated by undertaking a third cluster analysis based on yield data, of the subset of environments for which covariate information was available. The changes that occur between this cluster analysis and the two clusterings actually being compared would then need to be investigated. If however, there is no interest in identifying the exact cause of difference in cluster membership, but rather that differences exist, the cluster influence diagram is appropriate. This extends its application beyond the original intention.

Recognizing limitations of cluster influence diagrams in this instance, and therefore not attempting to attribute the alterations of the cluster groupings to any one factor, Figure 8.11 is presented. It shows the change of group membership from the clustering of 101 environments with full covariate information (Figures 7.3) to the clustering of all 123 environments based on available yield data (7.13). The large discrepancy in the number of mega-environments determined by the two methods is clearly evident in this figure. While the clustering of 101 environments which had full covariate information available determines that fifteen mega-environments exist, the clustering based on available yield data of all 123 environments shows thirty mega-environments. There are no LHS clusters that match RHS clusters and 72 distortions.

The figure shows that the marked differences in cluster memberships cannot be attributed to premature truncation of the dendrogram in Figure 7.3. Premature truncation would result in RHS clusters merging to form LHS clusters.

Table 8.5 shows the λ and U values for the comparisons presented in this section. λ and

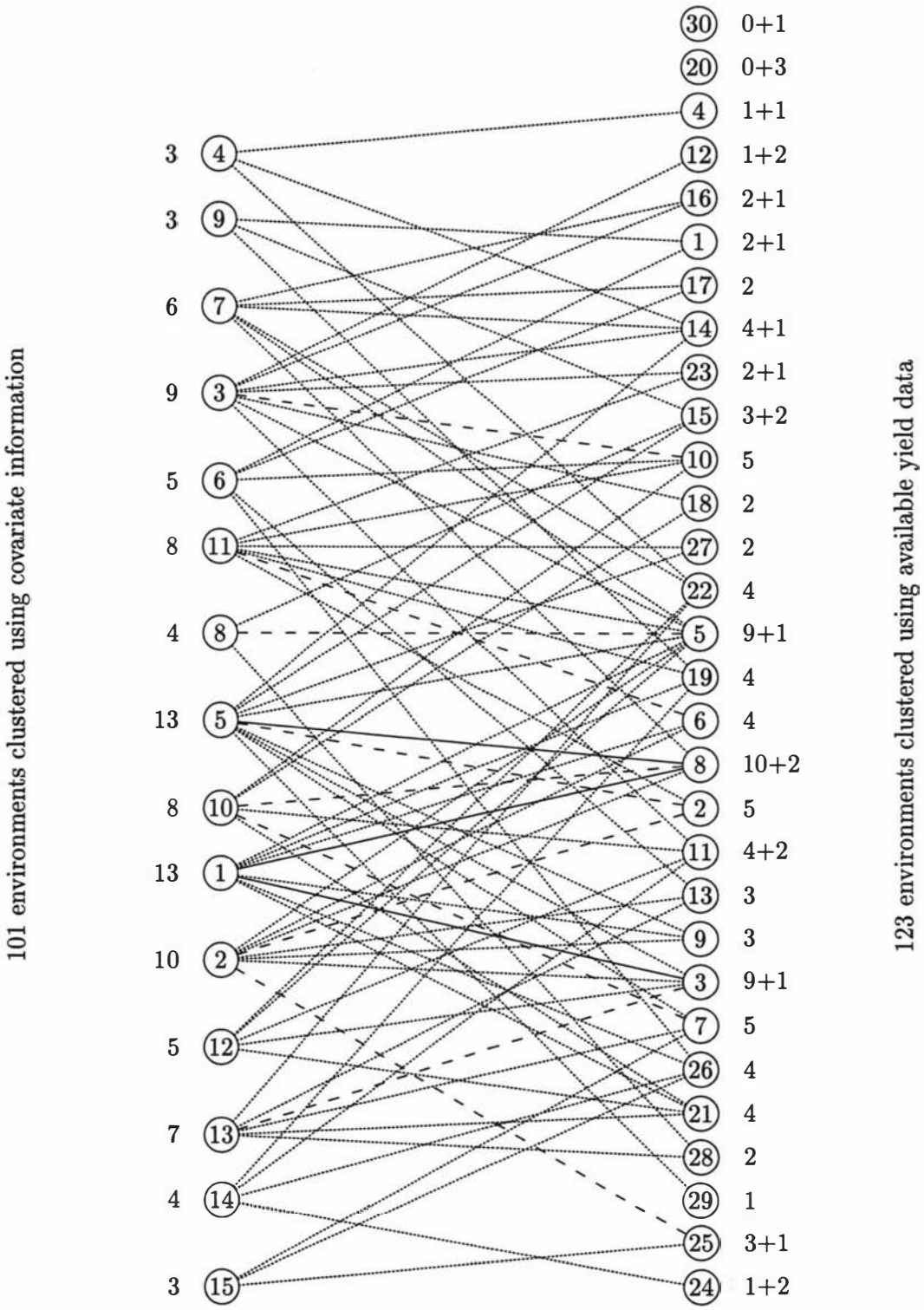


Figure 8.11: Cluster influence diagram for the dendrograms presented in Figures 7.3 (right) and 7.13 (left). Left-hand-side clusters were formed using data from the 101 environments for which full covariate information was available (Section 7.4), while right-hand-side clusters are based on available yield data from all 123 environments from the onion trials programme (Section 7.2).

Cluster influence diagram	Dendrograms in Figures	Asymmetric λ	Asymmetric U	Number of distortions	No. of unchanged clusters	Rand's R_g
8.11	7.13 and 7.3	0.295 (0.052)	0.527 (0.022)	72	0	0.894
101 environments clustered using available covariate information and 123 environments with sparse yield data						
8.12	7.4 and 7.16	0.289 (0.063)	0.526 (0.023)	65	0	0.881
89 Onion Data I environments clustered using available covariate information and 109 Onion Data I environments with sparse yield data						
8.13	7.5 and 7.17	0.364 (0.064)	0.514 (0.031)	54	0	0.875
78 Onion Data II environments clustered using available covariate information and 98 Onion Data II environments with sparse yield data						
8.14	7.8 and 7.16	0.276(0.062)	0.515(0.023)	70	0	0.882
89 Onion Data I environments clustered using available covariate information and 109 Onion Data II environments using two-stage imputed yield data						
8.15	7.9 and 7.17	0.364 (0.064)	0.522 (0.029)	51	0	0.869
78 Onion Data II environments clustered using available covariate information and 98 Onion Data II environments using two-stage imputed yield data						

Table 8.5: Numerical summary statistics corresponding to cluster influence diagrams presented in Figures 8.11 to 8.15; which compare clustering based on covariate information to that based on sparse and two-stage imputed yield data. Formulae for asymmetric λ and U are presented in (8.3) and (8.5). Asymptotic standard errors are provided in brackets. The number of distortions, the number of unchanged clusters, and Rand's R_g measure are also presented.

U in this instance are 0.295 and 0.527 respectively. Using other λ and U values found in Table 8.8 to place the current comparison in context, the correspondence between mega-environments determined using covariate information and available yield data appears tenuous. It would be difficult to determine which of the following causes may be responsible for this weak correspondence:

1. Insufficient covariate information is available.
2. The choice of G×E combinations used in the Trials Programme may have undue influence on the outcome of clustering based on available yield data.
3. The environments that do not have full covariate information and therefore were not included in Figure 7.13, may have impacted on this clustering by their absence.
4. The theory that yield data can provide the same information as collection of covariate data, does not hold in this instance.

Figures 8.12 and 8.13 which present cluster influence diagrams to compare clustering based on covariate information and those based on available yield data for Onion Data I and II respectively, support this argument.

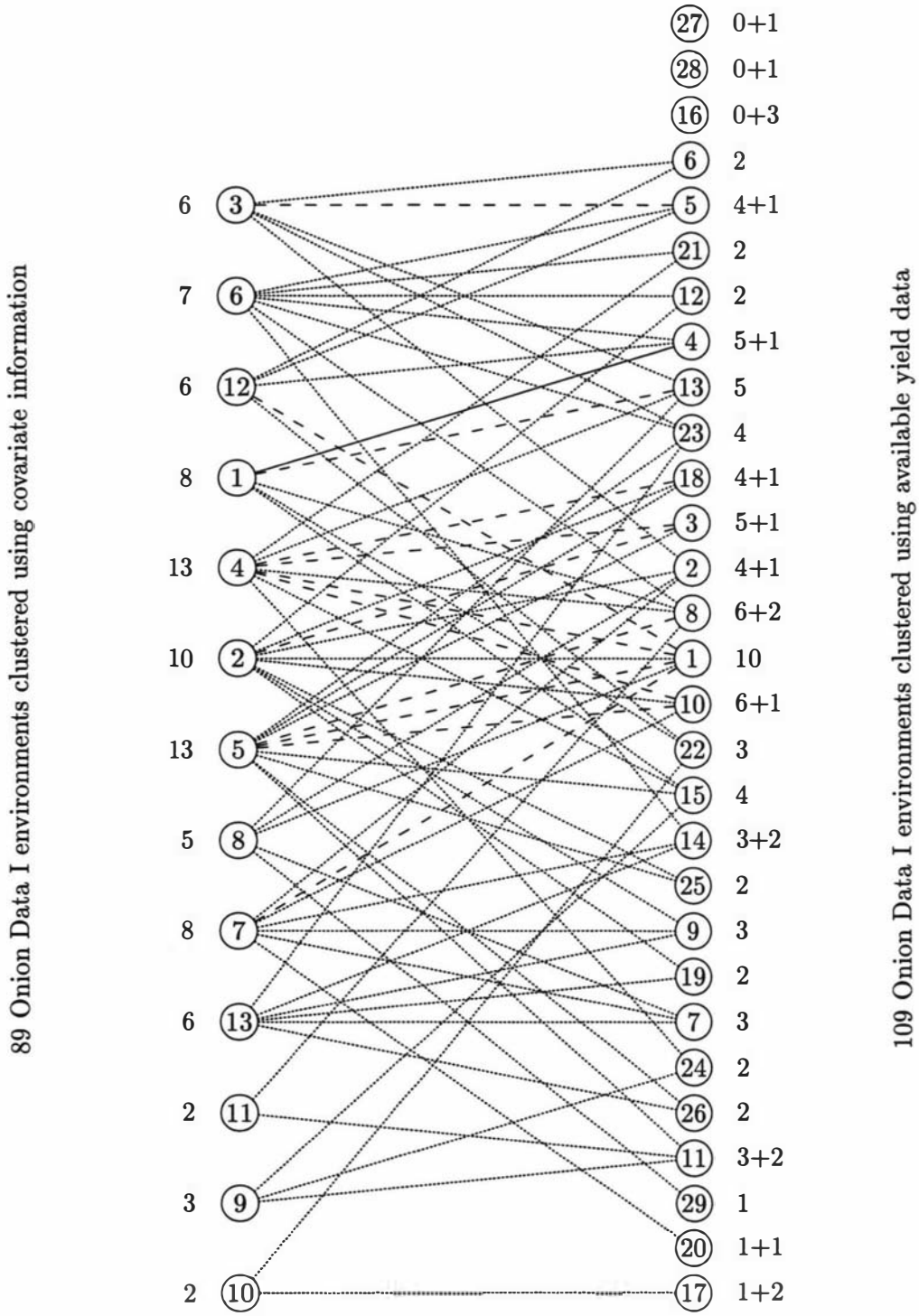


Figure 8.12: Cluster influence diagram for the dendrograms presented in Figures 7.4 (right) and 7.16 (left) which use the environments of Onion Data I. Left-hand-side clusters are those formed using the 89 environments with available covariate information (Section 7.4), while right-hand-side clusters were formed using available yield data from 109 environments (Section 7.2).

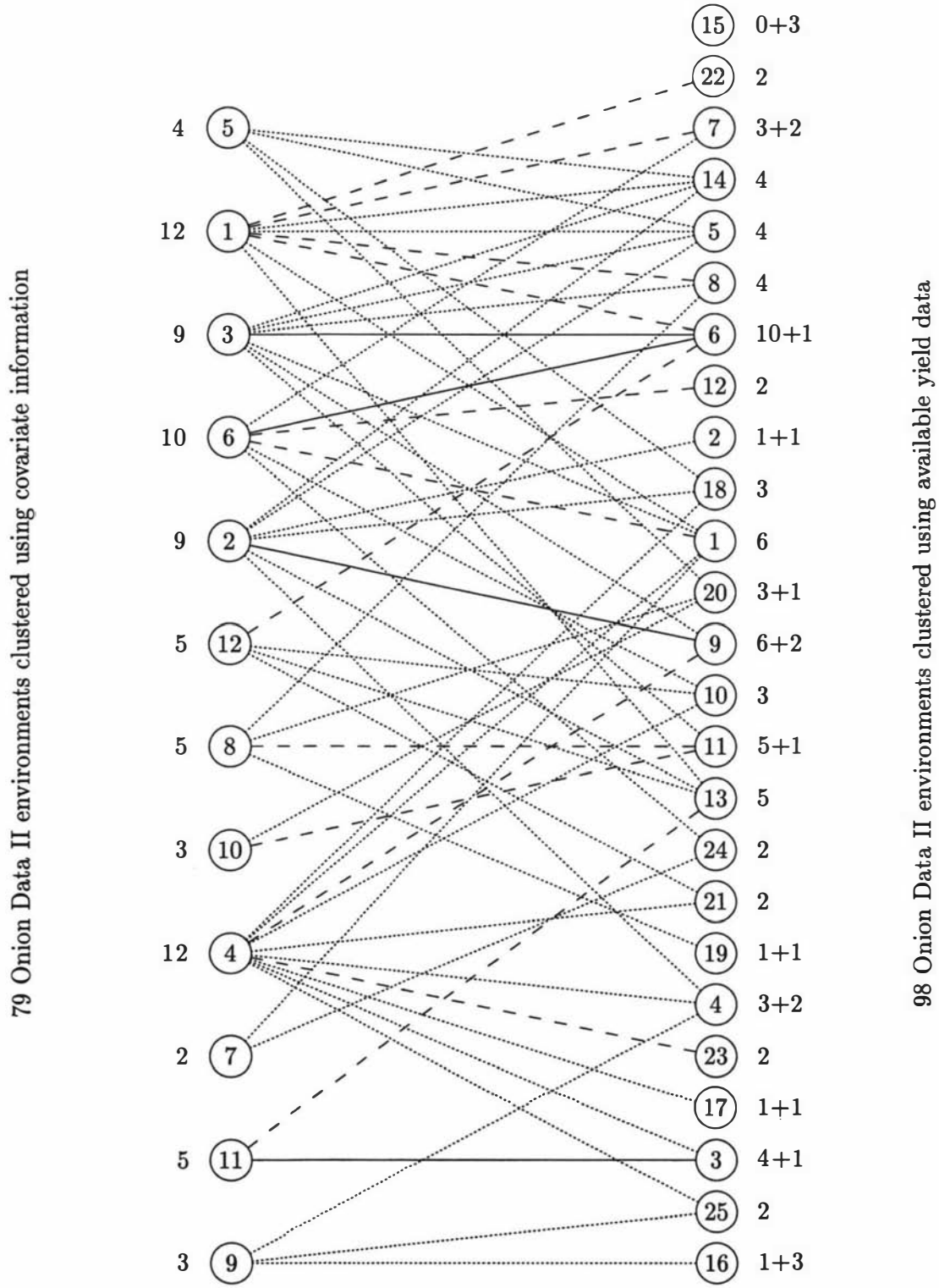


Figure 8.13: Cluster influence diagram for the dendrograms presented in Figures 7.5 (right) and 7.17. (left) which use the environments of Onion Data II. Left-hand-side clusters are those formed using the 79 environments with available covariate information (Section 7.4), while right-hand-side clusters were formed using available yield data from 98 environments (Section 7.2).

These cluster influence diagrams have 65 and 54 distortions respectively, and neither has any LHS to RHS cluster consistency. Again the asymmetric λ and U values for these comparisons were relatively low, as seen in Table 8.8. Rand's R_g measure for the comparisons presented in Figures 8.11 to 8.13 was also relatively low, being 0.894, 0.881, and 0.875 respectively. On the whole it appears that there was little consistency between clustering of environments based on covariate information and available yield data. The relationship between clustering based on covariate information and clustering based on two-stage imputed yield data is now considered.

Figure 8.14 shows the difference between the clustering of 89 Onion Data I environments based on available covariate information (Figure 7.16), and the clustering of all 109 Onion Data I environments based on two-stage imputed yield data (Figure 7.8). This cluster influence diagram has 70 distortions and no LHS to RHS cluster consistency. Figure 8.15 presents this comparison for the covariate information based clustering (Figure 7.17) and two-stage imputed yield data based clustering (Figure 7.9) for Onion Data II. It has 51 distortions and no LHS to RHS cluster consistency.

Asymmetric λ and U values corresponding to these cluster influence diagrams are lower than most others in Table 8.8. There is little evidence to support the notion that clustering based on two-stage imputed yield data will suffice in place of the outcome of clustering environments based on the covariate information collected. This poses a problem for programme organizers, in that the collected covariate information may be insufficient. This assertion may be refuted when the covariate information from the remaining 22 environments is made available and this analysis can be repeated. Until that time at least one of the methods for determining mega-environments presented in Chapter 7 must be considered as giving incorrect results. The mega-environments chosen will therefore determine the outcome so care will need to be taken in the selection of recommended genotypes.

8.7 The effect of imputation on the G×E structure

This section considers the effects of imputation on inter-genotype relationships within the data. As discussed throughout this work, exploitation of the G×E interaction structure is the mechanism by which gains will be made. Retention of the G×E structure, rather than creation of a new structure, must therefore be a goal of any imputation strategy used. The strengths and weaknesses of two-stage, nearest cluster, and closest observation methods, introduced in Chapter 6,will be highlighted.

One strength of the EM-AMMI (Gauch and Zobel, 1990) method for imputing missing yields is that results are consistent with the G×E interaction structure. The inherent weakness is that this G×E structure is determined by the model being used, and therefore leads to a self-consistent outcome. Selection of an inappropriate model will, therefore, re-

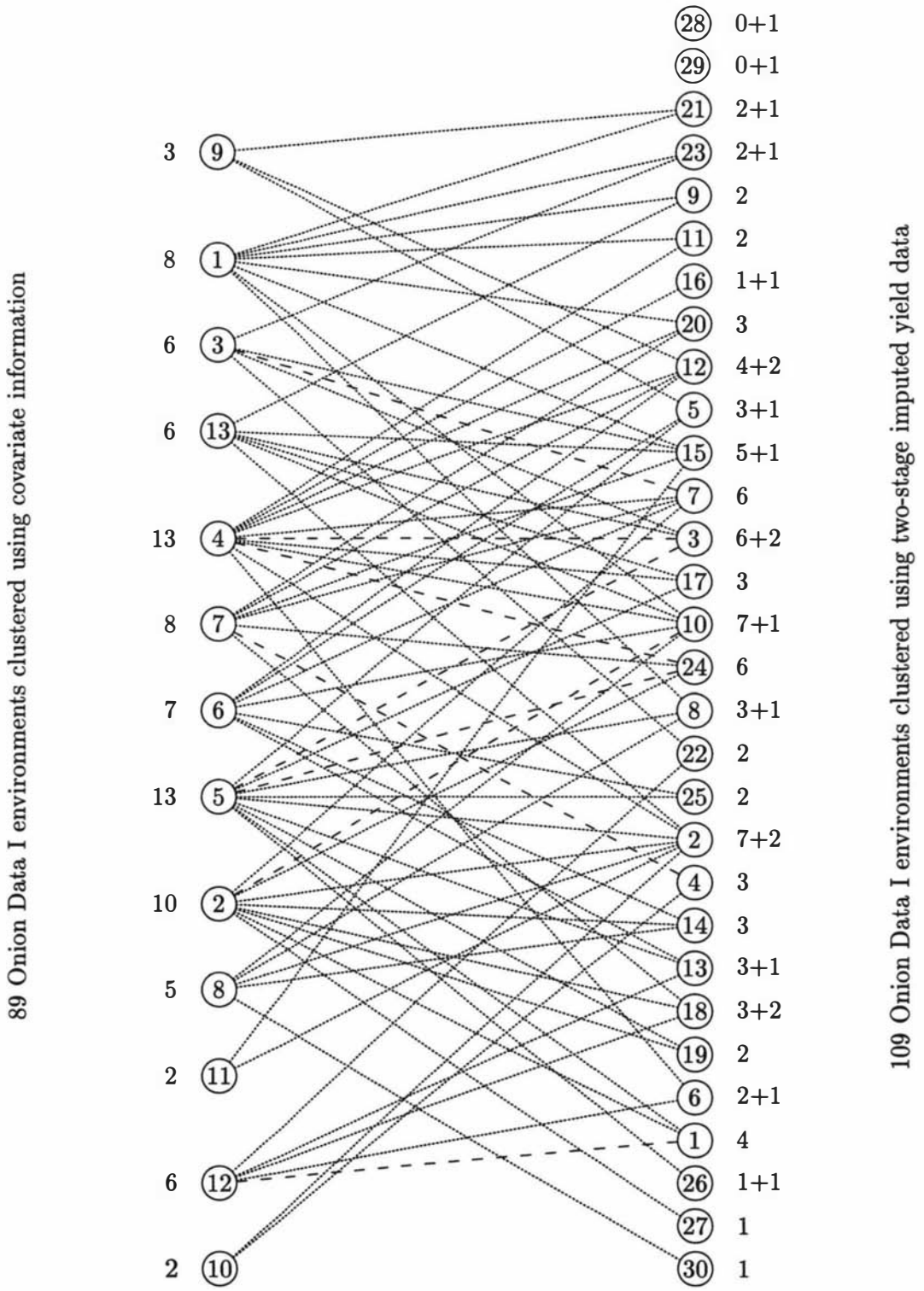


Figure 8.14: Cluster influence diagram for the dendrograms presented in Figures 7.8 (right) and 7.16 (left). Left-hand-side clusters are those formed using the 89 environments of Onion Data I with available covariate information (Section 7.4), while right-hand-side clusters were formed using two-stage imputed yield data for 109 environments (Section 7.3).

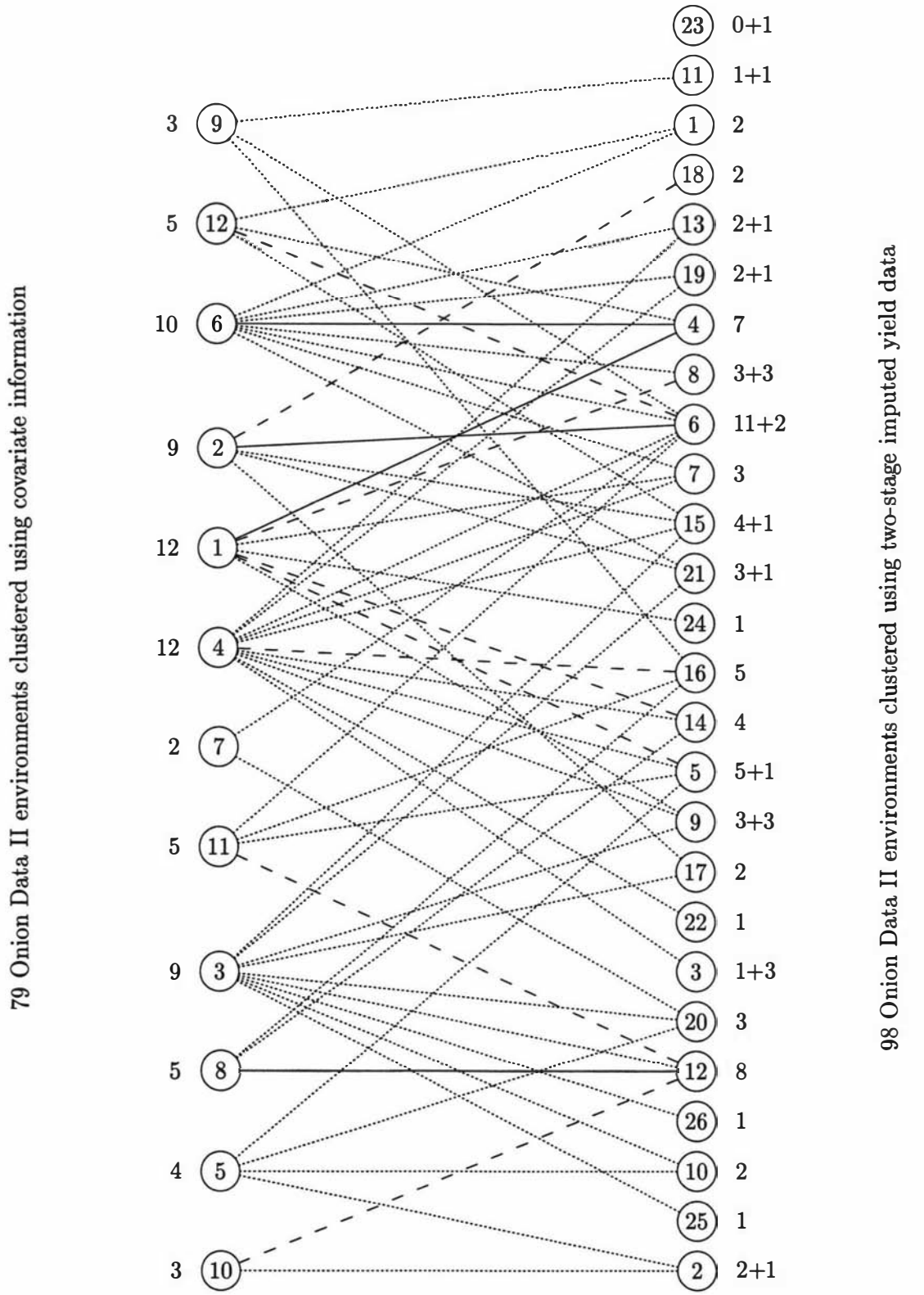


Figure 8.15: Cluster influence diagram for the dendrograms presented in Figures 7.9 (right) and 7.17 (left). Left-hand-side clusters are those formed using the 79 environments of Onion Data II with available covariate information (Section 7.4), while right-hand-side clusters were formed using two-stage imputed yield data for 98 environments (Section 7.3).

sult in an inferior set of imputed values. Two-stage, nearest cluster and closest observation imputation methods do not rely on a pre-determined G×E structure, but it is desirable that post-imputation results are consistent with pre-imputation results.

The performance of the nearest cluster (Drake, 1981) method of imputing G×E yields will be considered first. If this method of imputing results is consistent, genotypes that cluster when only available yield data is used will also be clustered when fully imputed data is used. Cluster influence diagrams for Onion Data I and II are presented in Figures 8.16 and 8.17 respectively. Dendrograms used in this comparison were formed using Euclidean distance, the incremental sums of squares method of forming clusters, and the stopping criterion discussed in Section 5.2.

In both these figures, post-imputation data fall into a smaller number of clusters than the sparse data. This can be attributed to the way the stopping criterion responds to the use of imputed distances in the clustering process. When sparse data was clustered, many distances were formed using (4.26) and the methodology outlined in Section 4.6. Imputed distances were greater than observed distances on average, so the total of all inter-genotype distances was greater for the sparse data, which had both imputed and observed distances, than for the fully imputed data which had only observed distances. This assertion is made on the grounds that all post-imputation clusters were formed by the merger of one or more clusters formed using sparse data. The nearest cluster method of imputing unobserved onion yields can be considered extremely consistent in its treatment of inter-genotype structure because:

1. In both cases there were no distortions (according to the notion developed earlier in this chapter).

2. Both λ and U were equal to one, for both Onion Data I and II.

3. There are four and six clusters that remained unchanged for Onion Data I and II respectively.

4. Rand's R_b for Onion Data I and II were 0.966 and 0.932 respectively.

Arguably, interaction distance should have been used to cluster genotypes in these figures to investigate how nearest cluster imputation has altered the G×E structure. Post-imputation clusters were, however, no different when interaction distance was used instead of Euclidean distance because many genotype pairs had zero difference in environments where data was imputed; pairs of genotypes used to impute one another's missing data therefore had observed distances that are in fact smaller than the distances observed using the incomplete data. These genotype pairs also had very similar interaction profiles as a direct result of nearest cluster imputation. On the other hand, pre-imputation clusters were affected by the distance measure used. Table 8.6 shows the impact of using Euclidean and interaction distances on the comparisons made in Figures 8.16 and 8.17. Decreases in

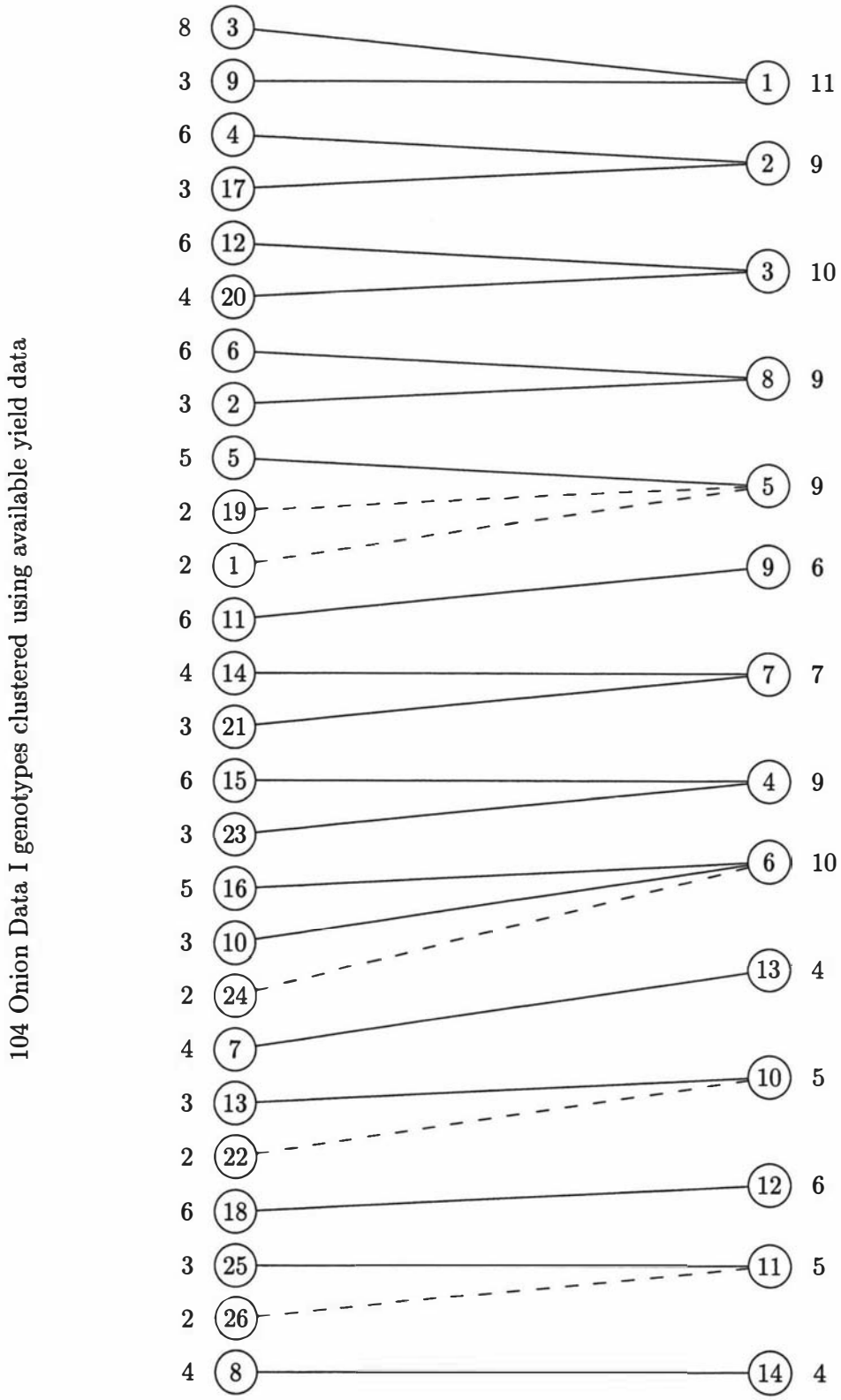


Figure 8.16: Effects of nearest cluster imputation on clustering of genotypes in Onion Data I, as seen in the dendrograms in Figures 6.9 (left) and 6.11 (right). The left-hand-side represents the clusters formed using the sparse data, while the right-hand-side shows the clusters formed after imputation.

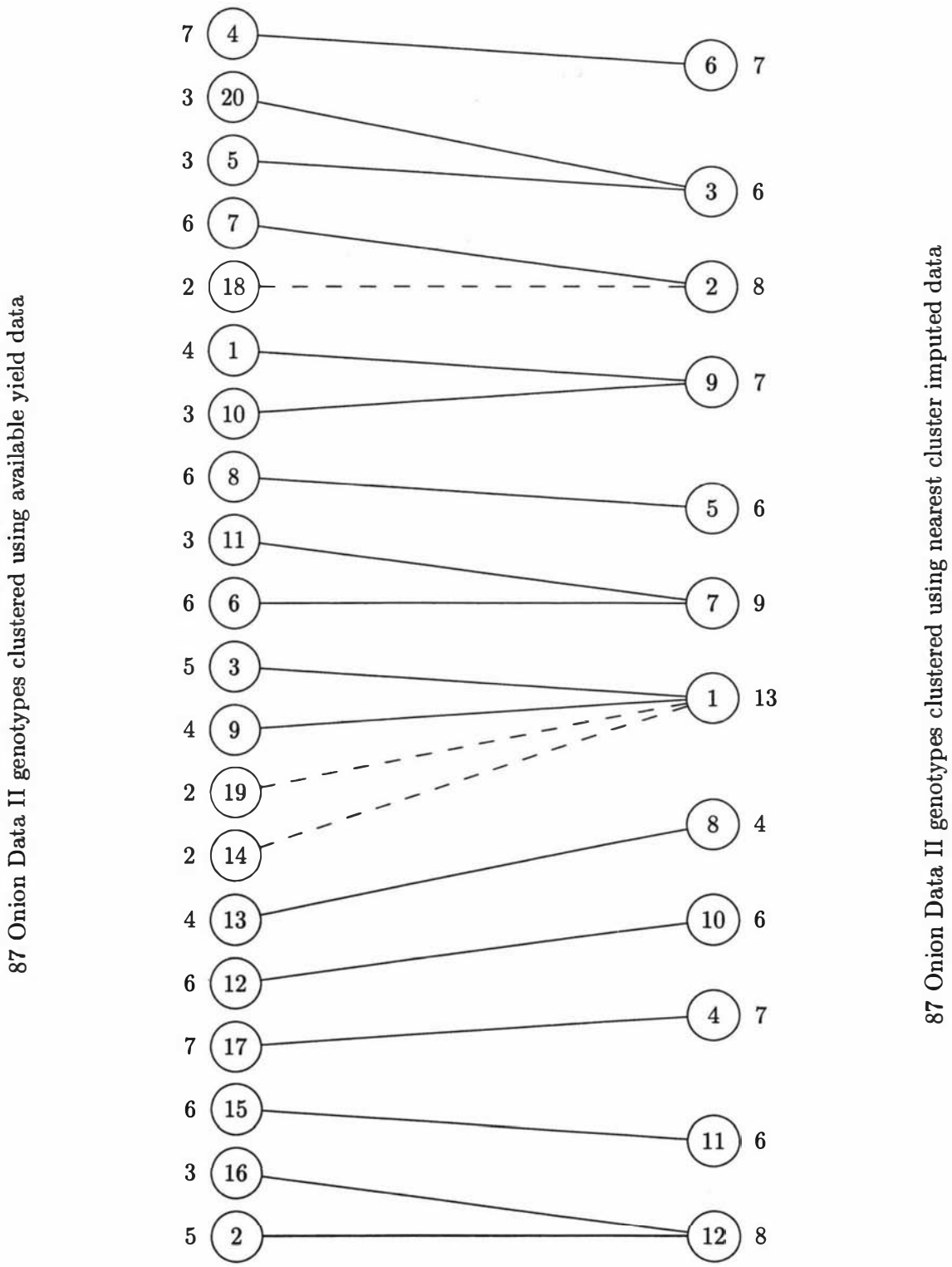


Figure 8.17: Effects of nearest cluster imputation on clustering of genotypes in Onion Data II, as seen in the dendrograms in Figures 6.10 (left) and 6.12 (right). The left-hand-side represents the clusters formed using the sparse data, while the right-hand-side shows the clusters formed after imputation.

asymmetric λ and U values along with an increase in the number of distortions, for both Onion Data I and II indicated that nearest cluster imputation definitely altered the G×E interaction structure.

Numerical summaries for relationships between genotype clusters before and after application of closest observation imputation are also presented in Table 8.6. These show that closest observation imputed values were less consistent than nearest cluster imputed values, as asymmetric λ's and U's were lower and the number of distortions rose for each distance measure applied to both Onion Data I and II.

Cluster influence diagrams presented in Figures 8.18 and 8.19, as well as numerical summaries found in Table 8.6 show that two-stage imputation is not as self-consistent as nearest cluster imputation. It did not however, alter the G×E structure of the data as much as nearest cluster or closest observation imputation. Asymmetric λ and U values presented in this section use the post-imputation cluster memberships as the dependent factor in each relationship described. This was done to investigate the extent to which sparse data predicts a complete set of data. Table 8.8 shows that genotype clustering based on two-stage imputed values are strongly related to those found using the sparse data.

The stopping criterion, described in Section 5.2, determined a different number of genotype clusters for each set of data, distance measure, and imputation method. Table 8.7 shows the number of genotype clusters that are identified when using the sparse data with Euclidean or interaction distance, and after each of the three imputation strategies have been applied. Most notable was the reduction in the number of genotype clusters resulting from nearest cluster imputation. Whether the clustering has now combined two or more

Data	Imputation method	Distance measure	Asymmetric λ	Asymmetric U	Number of distortions
Onion Data I	Nearest Cluster	Euclidean	1.000	1.000	0
		Interaction	0.806	0.840	16
	Closest Observation	Euclidean	0.479	0.701	42
		Interaction	0.447	0.659	47
	Two-stage	Interaction	0.774	0.862	13
Onion Data II	Nearest Cluster	Euclidean	1.000	1.000	0
		Interaction	0.811	0.878	11
	Closest Observation	Euclidean	0.450	0.656	37
		Interaction	0.438	0.0.667	37
	Two-stage	Interaction	0.875	0.933	6

Table 8.6: Summary statistics that quantify effects of imputation methods on clustering of genotypes. The distance measure was applied to sparse data and data imputed using each of two-stage, nearest cluster, and closest observation imputation methods for both Onion Data I and II. Asymmetric λ and U values use the after imputation clustering as the dependent factor in the relationship.

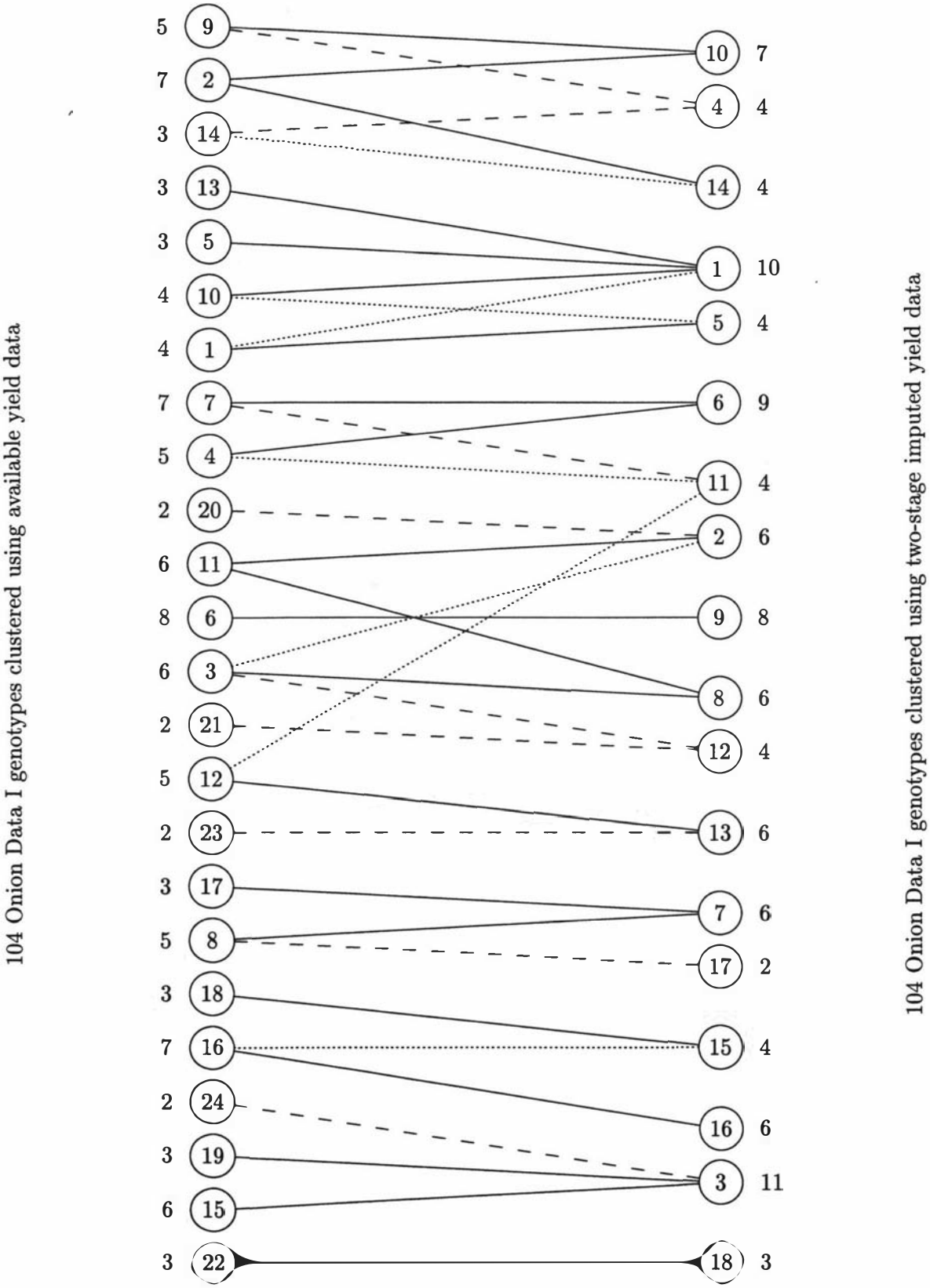


Figure 8.18: Effects of two-stage imputation on clustering of genotypes in Onion Data I, as seen in the dendrograms in Figures 5.8 (left) and 6.7 (right). Left-hand-side clusters were formed using sparse data, while right-hand-side clusters were formed after two-stage imputation.

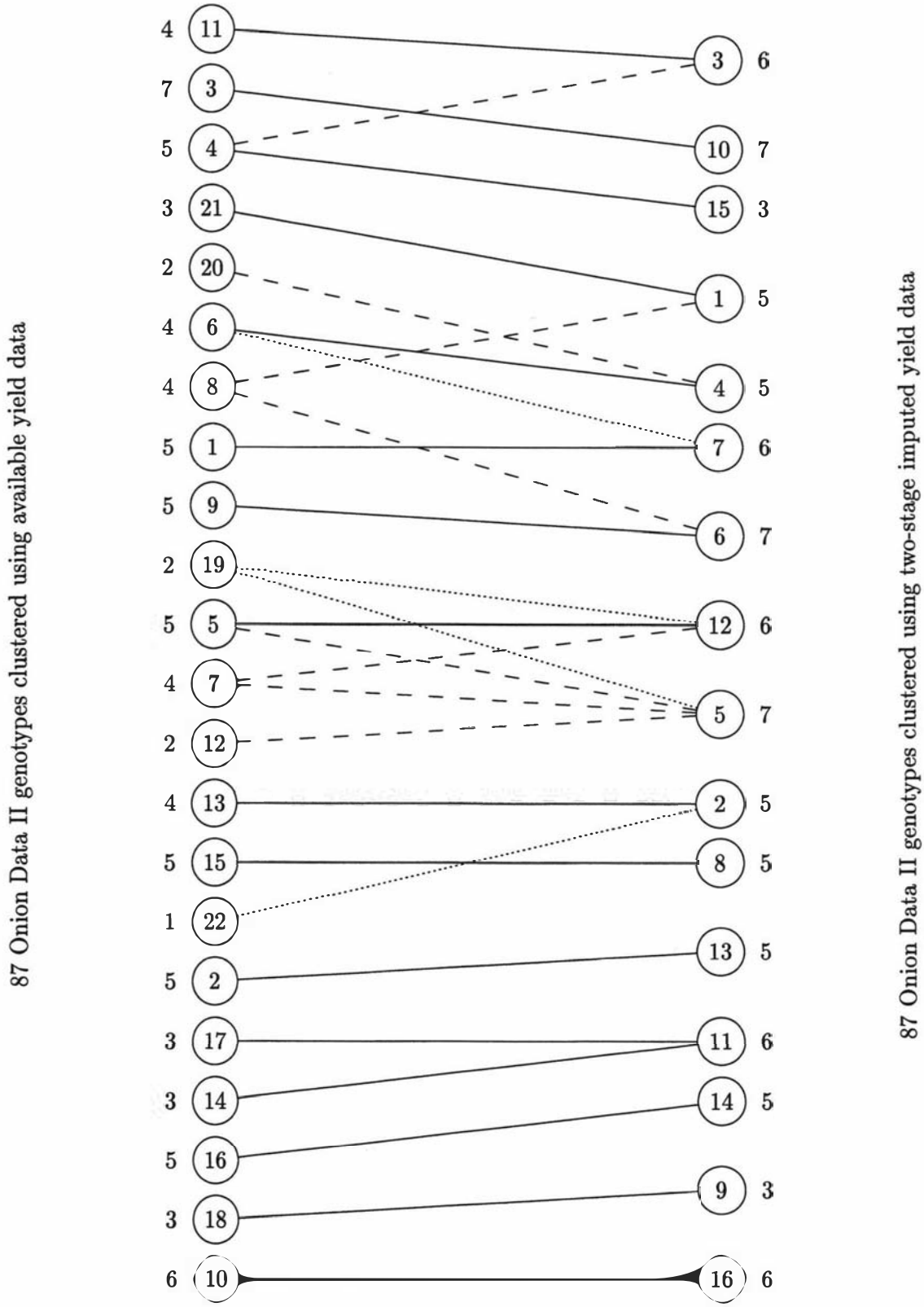


Figure 8.19: Effects of two-stage imputation on clustering of genotypes in Onion Data II, as seen in the dendrograms in Figures 5.9 (left) and 6.8 (right). Left-hand-side clusters were formed using sparse data, while right-hand-side clusters were formed after two-stage imputation.

groups that are different into one cluster, cannot be determined. On this basis, it cannot be said that nearest cluster imputation has altered the G×E structure, hence the need to use interaction distance on this modified set of data (given above).

When this same examination was extended to the data formed using closest observation imputation, it was easy to determine that the G×E structure was significantly altered. There was little need to create cluster influence diagrams in this instance as there were a large number of distortions, as seen in Table 8.6, and an increased number of clusters as seen in Table 8.7.

Use of two-stage imputed values led to a reduction in the number of clusters. Genotypes that should be deemed similar, but were grown in different sets of environments may have clustered sooner in the clustering based on imputed data. Such pairs of genotypes would have been given over-estimated distances for the clustering that used sparse data, instead of the observed distance between them that would have been recorded if they had been grown in a higher number of common environments.

Two-stage imputation is less self-consistent than nearest cluster imputation when applied to the sparse data of Onion Data I and II, but has greater self-consistency than closest observation imputation. While self-consistency may seem important, the most desirable imputation method will be the one that takes the sparse data and provides imputed values that are as close to the unobserved values as possible. Section 6.4 showed that none of the imputation methods are perfect in their ability to reconstruct a complete data set.

The findings of this section are reliant on the assumption that sparse data is capable of providing the same information as would complete data. In other words, these findings are true for data that is 'missing completely at random', but not necessarily true for data that is 'missing at random'. The investigation has not used a data set that can be compared to the complete data, so reliance has been placed on the notion that sparse data

Data	Condition	Distance measure	No. of clusters
Onion Data I	Sparse	Euclidean	26
		Interaction	24
	Nearest cluster imputed	Either	14
	Two-stage imputed	Either	18
	Closest observation imputed	Either	27
Onion Data II	Sparse	Euclidean	20
		Interaction	22
	Nearest cluster imputed	Either	12
	Two-stage imputed	Either	16
	Closest observation imputed	Either	22

Table 8.7: Number of genotype clusters found using sparse and imputed data from Onion Data I and II. The number of clusters obtained depends on the distance measure employed and the method of imputation. Clustering of fully imputed yields gives identical results when either Euclidean or interaction distance was used.

represents the unobserved data as well. If this is so, the G×E structure within the sparse data must then represent the G×E structure of the unobserved complete data. An aim of any imputation method must then be to alter the G×E structure of the data as little as possible; two-stage imputation is therefore superior to both nearest cluster and closest observation imputation, in that it alters the G×E structure of the data the least. In the next section, the notion that the sparse data of the Onion Trials Programme are in fact 'missing completely at random' is investigated.

8.8 The dependence of imputations on commonality of test environments

The relationship between post-imputation results and the sparsity of data is investigated in this section. Comparison of genotype clustering based on imputed data versus that based on commonality of environments provided an indication of two-stage imputation's relevance to answering the principal research question.

Simulated testing of the efficacy of two-stage imputation over competing methods was given in Section 6.4, but did not extend to the level of sparsity encountered in the data arising from the Onion Trials Programme. That testing was limited by computing capacity, but had this been available, such testing would not have uncovered any relationship between the imputed results and the commonality of test environments as described in this section.

A matrix representing use of G×E combinations in the Onion Trials Programme, was used to form clusters based on commonality of test environments for genotypes. Euclidean distance was applied to this complete matrix of zeros and ones, which indicated absence and presence respectively of data in the original G×E matrix. The same methods for forming clusters and determining the truncation level for clustering, described in Section 5.2, were applied to indicator matrices for both Onion Data I and II. Cluster influence diagrams have not been constructed for these cluster analyses, but as Table 8.9 shows, there was a significant relationship between the clustering of genotypes based on imputed results and on commonality of tested environments. This problem is inherent to the sparsity of these two data sets, rather than the imputation method applied, as the λ's and U's presented show consistently strong relationships for all imputed results.

These findings indicate that imputed values are dependent on the existence of observable inter-genotype relationships, rather than the strength of these observable relationships. The upper bound approach used to estimate unobserved distances provided distances that were, on average, much greater than observed distances. First stage clusters of Onion Data I and II genotypes are, therefore, based on the commonality of environments more than the similarity of interaction profiles as intended. Because of this, there is great potential for imputations to be based on the wrong set of genotypes.

Cluster influence diagram in Figure	Dendrograms in Figures	Asymmetric λ	Asymmetric U	Symmetric λ	Symmetric U	Number of distortions	Number of unchanged clusters	Rand's R_g
8.1	7.16 and 7.17	0.224 (0.073)	0.366 (0.028)	0.239 (0.056)	0.372 (0.027)	33	1	0.840
8.2	7.13 and 7.16	0.390 (0.063)	0.534 (0.033)	0.412 (0.059)	0.549 (0.032)	32	2	0.881
8.3	7.13 and 7.17	0.250 (0.069)	0.391 (0.025)	0.281 (0.056)	0.409 (0.025)	33	0	0.852
8.4	7.3 and 7.4	0.691 (0.047)	0.834 (0.017)	0.679 (0.039)	0.826 (0.015)	24	7	0.958
8.5	7.4 and 7.5	0.742 (0.046)	0.860 (0.016)	0.761 (0.038)	0.864 (0.016)	17	9	0.968
8.6	7.3 and 7.5	0.636 (0.053)	0.793 (0.020)	0.640 (0.041)	0.792 (0.018)	28	6	0.949
8.7	7.4 and 7.8	0.313 (0.047)	0.597 (0.019)	0.312 (0.037)	0.594 (0.017)	68	0	0.932
8.8	7.5 and 7.9	0.287 (0.050)	0.572 (0.020)	0.285 (0.042)	0.576 (0.018)	59	0	0.918
8.9	7.8 and 7.9	0.371 (0.057)	0.628 (0.020)	0.397 (0.044)	0.649 (0.019)	53	1	0.931
8.10	6.7 and 6.8	0.628 (0.058)	0.742 (0.022)	0.639 (0.050)	0.746 (0.022)	22	0	0.953
8.11	7.13 and 7.3	0.295 (0.052)	0.527 (0.022)	0.235 (0.043)	0.477 (0.018)	72	0	0.894
8.12	7.4 and 7.16	0.289 (0.063)	0.526 (0.023)	0.213 (0.043)	0.462 (0.020)	65	0	0.881
8.13	7.5 and 7.17	0.388 (0.068)	0.580 (0.027)	0.301 (0.045)	0.509 (0.023)	54	0	0.875
8.14	7.8 and 7.16	0.276 (0.062)	0.515 (0.023)	0.209 (0.045)	0.446 (0.019)	70	0	0.882
8.15	7.9 and 7.17	0.373 (0.059)	0.550 (0.026)	0.289 (0.044)	0.484 (0.020)	51	0	0.869
8.16	6.9 and 6.11	1.000 (0.000)	1.000 (0.000)	0.831 (0.025)	0.898 (0.007)	0	4	0.966
8.17		1.000 (0.000)	1.000 (0.000)	0.857 (0.027)	0.911 (0.010)	0	6	0.932
8.18	5.8 and 6.7	0.774 (0.043)	0.862 (0.015)	0.683 (0.035)	0.821 (0.013)	13	2	0.955
8.19	5.9 and 6.8	0.875 (0.038)	0.933 (0.013)	0.806 (0.038)	0.889 (0.014)	6	6	0.973

Table 8.8: Complete listing of all summary statistics corresponding to cluster influence diagrams presented in Figures 8.1 to 8.19. Formulae for asymmetric and symmetric cases of λ and U are presented in (8.2) to (8.5). Asymptotic standard errors are provided in brackets. Although both asymmetric and symmetric forms were not used in the preceding sections, they have all been presented to allow comparison. The number of distortions, the number of unchanged clusters, and Rand's R_g measure are also presetned.

The principal research question, "Given a certain (possibly new) environment, which onion varieties are most likely to succeed in terms of their edible yield?" now seems rather elusive, and may need to be set aside in favour of another question that can be answered. This is not to say that the aim of answering the original question should be ignored, but that a pragmatic use of the information currently available would be to answer an auxiliary question instead. Given the inability of current data to answer the principal research question, what should be done to allow the question to be answered in the future? Development of strategies to minimize the negative impacts of sparsity seems at first to be a logical initiative. To understand these negative impacts, they must be known. Some methods for limiting the impact of sparsity in future will be discussed in Chapter 10.

Extremely sparse data has been shown to provide results that are determined by what has already been tried. This is not to say that the results are biased towards what has been tried before. Bias towards already tested G×E combinations would be shown by imputed data confirming already tested G×E combinations as the 'best' selections. This is not the case with either Onion Data I or II, because Section 6.5, highlighted a large number of imputed values that were greater than the observed maximum environmental yield. These imputed values lead to the set of best selections including both tested and untested G×E combinations.

Results from imputation are likely to include some values that seem anomalous to agronomists. Sets of best genotypes for environments show which varieties should be tested in similar environments. Only actual testing of anomalous results will separate them from good predictions. When more trials are added to the programme, current theories based on current imputed values can be tested. Their results can then be added into the G×E matrix allowing recalculation of imputed values. Further testing and progression of the trials programme should therefore improve imputations.

To close this discussion, an auxiliary question needs to be offered. Use of imputation methodology gives results based on current information. These results come in the form of a suggested list of varieties for testing in each new environment. The auxiliary research question must therefore become,

Imputation method	Data set used	Asymmetric	
		λ	U
Two-stage	Onion Data I	0.516	0.666
	Onion Data II	0.513	0.682
Nearest cluster	Onion Data I	0.581	0.707
	Onion Data II	0.568	0.694
Closest observation	Onion Data I	0.383	0.621
	Onion Data II	0.375	0.605

Table 8.9: λ and U coefficients for clustering of genotypes based on imputed data and commonality of test environments.

"Given a certain (possibly new) environment and using our current knowledge, which onion varieties should we test in order to find out which succeed in terms of their edible yield?"

8.9 Summary

In this chapter, tools for comparing pairs of cluster analyses were introduced, and applied to cluster analyses presented in Chapters 6 and 7. Cluster influence diagrams were presented for many of the comparisons, while the Goodman and Kruskal λ, Theil's uncertainty coefficient U, Rand's R_g measure, and the number of distortions were measured to give quantitative tools for the comparisons examined.

The Goodman and Kruskal λ and Theil's uncertainty coefficient U values provided the same conclusions, but doubt exists about the efficacy of the standard error estimates, which are based on asymptotic theory. Their use did, however, show that all cluster analyses for environments, as well as those for genotypes, were related to one another.

Simulation testing showed that observed R_g scores for pairs of clusterings are more dependent on the number of observations being clustered and the number of clusters into which they are clustered, than the Rand (1971) definition suggests. This measure could be used as a rough guide for comparing results in this chapter, but should not be relied upon to provide a truly meaningful measure of the similarity of a pair of clusterings.

Cluster influence diagrams show the transfer of cluster membership from one clustering to another by:

1. Listing all clusters in a pair of cluster analyses.
2. Detailing the number of members in each cluster, including the addition of observations not included in the other set of observations being clustered.
3. Using different lines to represent the number of observations being transferred.

Use of cluster influence diagrams allowed the mega-environment clusterings presented in the previous chapter to be compared. Mega-environments created using yield data were generally more consistent than those created using the covariate information that was collected for 101 of the 123 environments included in the Onion Trials Programme. The fact that clusterings based on covariate information were different to those based on yield data brought the usefulness of the mega-environments formed in Chapter 7 into doubt. Further investigation is required to attribute agronomic factors to environments so that sense can be made of the clusterings created using yield data from the Onion Trials Programme.

Two-stage imputed values were shown to be consistent between Onion Data I and II in terms of both environment and genotype clusterings, in Section 8.5. This corroborated the findings of Section 6.5. Consistency of inter-genotype clustering was investigated in

Section 8.7 to gauge the impact different imputation methods have on the G×E structure of the data sets. Nearest cluster imputation was seen to be extremely consistent, while two-stage imputation was shown to be superior to the closest observation method. The fact that there is no way of gauging how well these imputation methods are estimating unobserved yields means that the consistency of genotype clusters is not necessarily a good yardstick. Ideally, the consistency of genotype clusters would need to be investigated through use of simulations similar to those presented in Section 6.4.

The strongest relationships among the sets of cluster influence diagrams were within the mega-environments based on available yield data. Although not investigated thoroughly, the similarity of results found using the imputed values based on the closely related sparse data came as no surprise. Confirmation that imputed values are only as good as the data used to create them, and more particularly, finding that imputed results were dependent on the actual G×E combinations tested in the Onion Trials Programme meant that the principal research question behind this investigation needed to be qualified in the previous section. The final part of this investigation shows how the amended research question can be answered.

Part III

The Solution: Results and Implications

Chapter 9

Genotype selections for a new environment

9.1 Introduction

This part of the investigation brings together the findings of Chapters 4 to 8 in order to answer the auxiliary research question, defined in Section 8.8:

> "Given a certain (possibly new) environment and using our current knowledge, which onion varieties should we test in order to find out which succeed in terms of their edible yield?"

In this chapter results are presented from the application of some common G×E methodologies (discussed in Chapter 2) to the complete G×E matrices created by two-stage imputation in Section 6.5. Some reflections upon the investigations presented in this and preceding chapters are then covered in Chapter 10.

The graphical approaches presented in Section 9.2 highlight the difficulty of working with data sets of the size of Onion Data I and II. Selection of varieties by their wide adaptability to the environments covered by the Onion Trials Programme will be discussed in Section 9.3 using two-stage imputed matrices. This revisits the analysis given in Section 3.7 which used the sparse G×E matrices.

The environment clusterings given in Chapter 7 can be used to take subsets of the fully imputed G×E matrices in order to identify the relevant data to consider when selecting varieties for new environments. In Chapter 8, however, the suitability of this method was brought into question by demonstrating the inconsistency of these clusterings, which were based on three different sets of variables. Some pragmatic approaches to environment selection were then used to illustrate alternatives for selecting genotypes for new environments.

Twelve environments in the Onion Trials Programme, that had latitude, altitude, temperature, and photoperiod data available, were not included in Onion Data I or II

Tricode	Country	Year	No. of Genotypes
A01603	Yemen	1993	12
A01606	Yemen	1994	15
A05901	Fiji	1991	7
C04401	PNG	1993	11
C04402	PNG	1993	8
F02703	Tanzania	1996	6
L01103	Kenya	1998	16
O02707	Sri Lanka	1994	12
W17505	Greece	1995	6
W17506	Greece	1995	6
Z12201	Malawi	1996	12
Z13501	Nigeria	1997	9

Table 9.1: Twelve environments whose yield data was not part of Onion Data I but had covariate information collected for latitude, altitude, temperature, and photoperiod. Columns provide information on the trial's code (Tricode), the name of the country, the year in which the trial commenced, and the number of genotypes grown in the trial.

because they did not have yield data for enough genotypes. As can be seen in Table 9.1, which gives details of these environments, there are two situations where the definition used to create environments has caused some of the sparsity in the data. The Greek and Papua New Guinean pairs of trials were considered to be different environments because they were held in different parts of the same seasons. Combining these pairs of trials would have led to their inclusion in the analyses presented in Sections 5.4 and 6.5, but would have been in error. This is because these pairs of environments differed in the range of temperature and photoperiod, both of which are well known to affect the growth of onions (Brewster (1997). The fact that these environments were not included in those analyses allowed them to be used, in lieu of new environments, to show how imputed values and environment clusterings can be used to illustrate variety selections for a new environment.

In remaining sections of this chapter, a fictitious Yemeni trial, based on the trial not included in either of Onion Data I or II, will be allocated varieties for testing using:

1. Similarity of geographic locations (Section 9.4).

2. The clustering of environments using the covariate information presented in Chapter 7 (Section 9.5).

3. Using partially imputed (rather than fully imputed) G×E matrices (Section 9.6).

The third method used imputed values derived from only first stage clustering of the available data, and is therefore presented for comparison.

9.2 Starting with graphical approaches

This section aims to find simple graphical solutions for answering the principal research question in an efficacious manner. Graphical summaries are commonplace in G×E analyses, and can aid the researcher by giving some relativity to the range of genotype performances under examination.

Two methods are presented. First, the aim of the Onion Trials Programme to find genotypes that suit the lower latitude environments of the tropics and subtropics led the investigation towards a description of the relationship between latitude and yield. Second, the commonly used technique of biplotting, based on additive main effects and multiplicative interaction (AMMI) modelling, was investigated. The findings of these investigations are presented in this section.

Response to lower latitude environments

The primary interest of Dr Currah, which led to the organization of the Onion Trials Programme, was to determine which onion varieties suit the shorter, hotter days of the tropics and subtropics. A simple graphical summary of the ability of varieties to succeed in terms of yield in lower latitudes is now presented.

In Section 3.5 a parametric model was developed for the sparse data arising from the Onion Trials Programme. It was determined that a quadratic function of latitude could be used to explain the performance of each genotype. Each genotype therefore required three parameters to be estimated using this model, and therefore were not able to be plotted on a simple graph.

Rank correlation between imputed genotype performances and environment latitude was chosen to compare genotype responses to latitudes because it is robust to nonlinear relationships. The Spearman's rank correlations are plotted against genotype mean yields in Figure 9.1 for both Onion Data I and II. High positive correlation indicates specific adaptation to high latitudes, while the negative correlations indicate specific adaptation to low latitude environments. Genotypes that have significant negative correlations are presented in Table 9.2 with their average yield from the imputed G×E matrices.

The majority of genotype rank correlations were not significantly different to zero. An explanation for this could be that a genotype may be specifically adapted to a small range of environments within the latitude range of the Onion Trials Programme. There is a need to look for the optimal latitude for each genotype. If the quadratic model found in Section 3.5 for the sparse data was found to be suitable for the imputed data, the optimal latitude could be found for each genotype as the genotype's fitted curve maximum value. This was not included in the current investigation, because it was superseded by the investigation that follows in Section 9.4, which uses only information from a small range of latitudes to make genotype selections for a new trial.

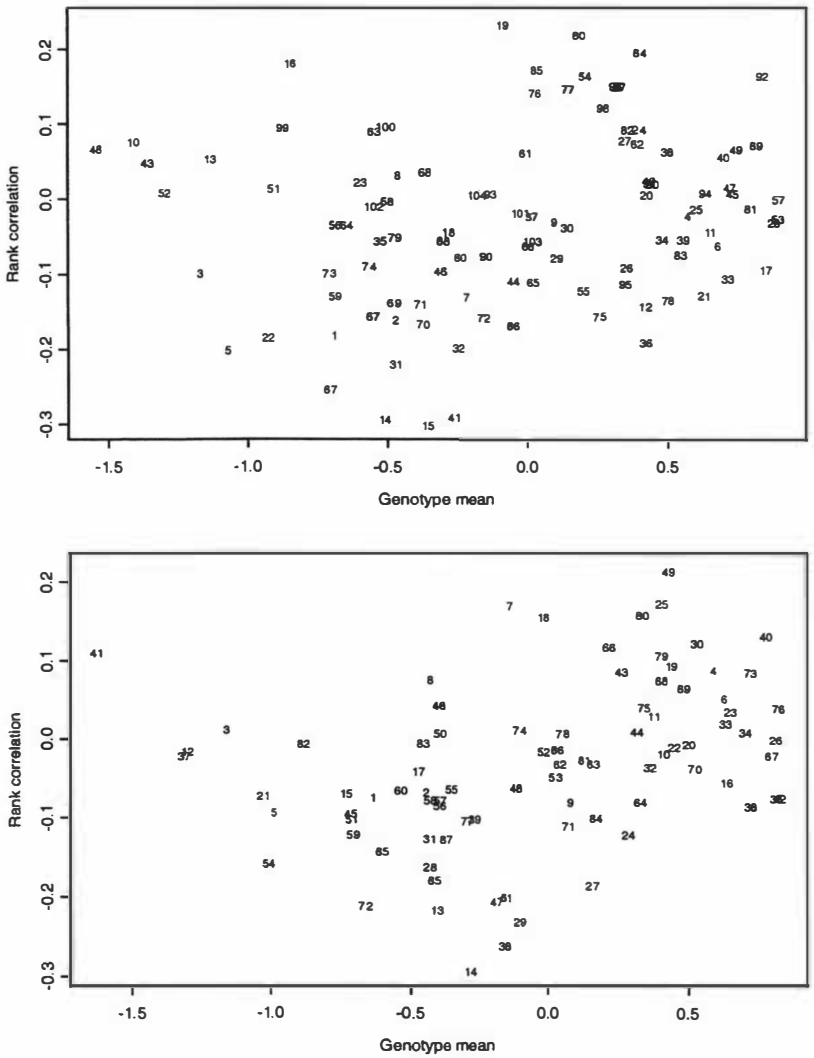


Figure 9.1: Correlations of expected yield and latitude plotted against mean yields.

	Onion Data I		Onion Data II	
Genotype	Correlation	Mean		
Creole Red PRR PS	−0.291	−0.508	−0.216	−0.406
Dehydrator No 3 SS	−0.299	−0.361	−0.292	−0.285
HA-222 HZ	−0.216	−0.470		
HA-226 HZ			−0.230	−0.102
IRAT-69 MA	−0.289	−0.265	−0.261	−0.157
Pusa Red AF			−0.205	−0.190
Rouge de Tana TS	−0.251	−0.701	−0.210	−0.655

Table 9.2: Genotypes that have significant negative rank correlation with latitude and therefore are determined to have good specific adaptation to low latitude environments. These genotypes all had rank correlations less than −0.200 in Figure 9.1. Mean performance across all environments is given also.

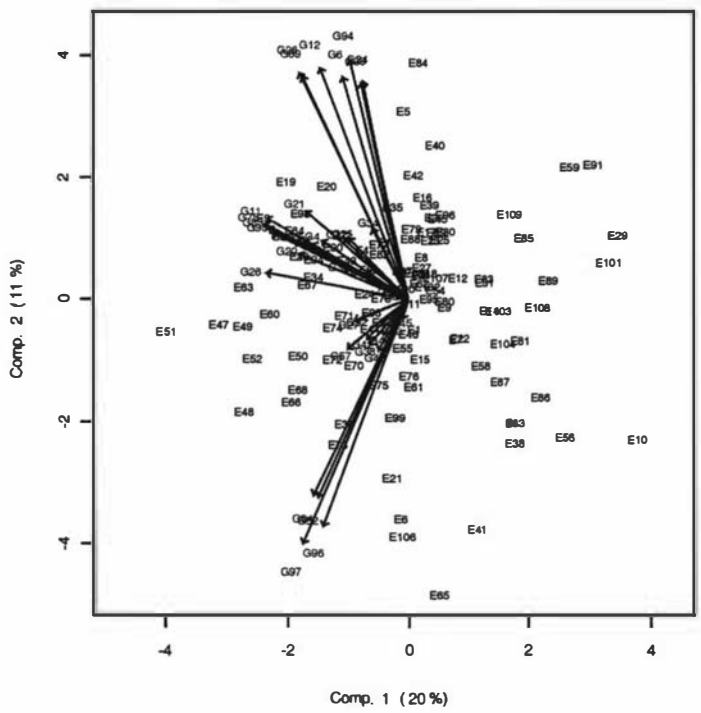


Figure 9.2: Biplot showing first and second G×E interaction principal component axes for the fully imputed yields of the top quarter of genotypes (in terms of mean yield across environments) and all 109 environments of Onion Data I. The percentage of total G×E interaction variation is given for each axis.

Use of the latitude of locations puts the environments on a predetermined ordinal scale. AMMI modelling will not pre-determine the ordering of environments in terms of any single covariate, so may provide a simpler method for determining the environments to which each genotype is suited.

AMMI analysis of Onion Data I and II

Biplots are commonly used to represent the relationships among genotypes and environments. As biplots for AMMI analyses are based purely on G×E interaction, specific adaptations are reflected, rather than superiority of performance. For example, the fact that a genotype performed above all others in every environment would not be shown by biplots, which would instead show where this genotype was best suited in comparative terms.

The high number of genotypes and environments in Onion Data I and II meant that the biplots would be unreadable if all genotypes and environments had been plotted. In Figures 9.2 and 9.3, the biplots plot the top quarter of genotypes, in terms of their mean yield across all environments for each data set. In these biplots, genotype and environment names have been replaced by short codes to improve clarity. These codes are given in Table 9.3 for genotypes and can be found in Appendix A for environments.

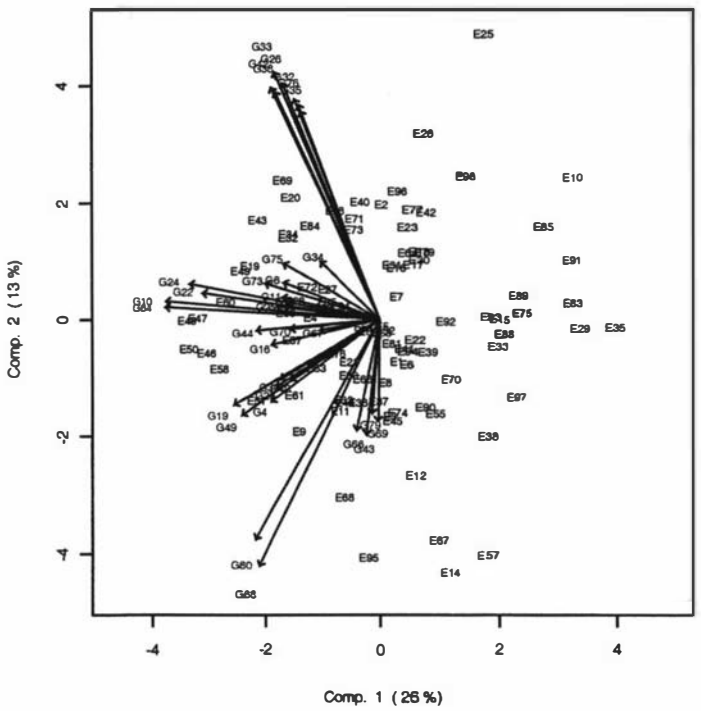


Figure 9.3: Biplot showing first and second G×E interaction principal component axes for the fully imputed yields of the top quarter of genotypes (in terms of mean yield across environments) and all 98 environments of Onion Data II. The percentage of total G×E interaction variation is given for each axis.

The biplots show that there are groups of genotypes with similar interaction profiles. It is hard to know whether these groupings are consistent from one biplot to another by just looking at the biplots themselves. For both Onion Data I and II, the first interaction principal component accounted for a large portion of the G×E interaction (20% and 26% respectively). As was expected, subsequent PC axes accounted for smaller portions of the existing G×E interaction.

It is difficult to examine the consistency between the biplot results for Onion Data I and II. If results were consistent, genotypes would have similar coordinates on the biplots for the first and second principal component axes. Some conflicting results are easily observed. The genotype 'Tropic Ace TK' can be seen at the bottom of both biplots, using the codes 'G97' and 'G80' in Figures 9.2 and 9.3 respectively, but results for the genotype 'Belem IPA-9 IP' differ as its code 'G6' is to be found at the top of Figure 9.2 and the left of Figure 9.3. If all genotypes and environments had similar coordinates on both biplots, the principal component axes would have a common interpretation and factors that determine them could be investigated.

Predictions are possible without an understanding of the causes of specific adaptation. In Figure 9.2, the genotype 'Gladalan White YA' with code 'G26' points towards environments 'E34', 'E63', and 'E67'. This biplot suggests that 'Gladalan White YA' would have yielded comparatively well in these three trials if it had been grown. The three tri-

Codes		Genotype name	Codes		Genotype name
4	4	Arad HZ	49	42	Mercedes PS
6	6	Belem IPA-9 IP	50		Mr Max RC
11	10	Colossal PVP SS	53		Nikita RC
12	11	Composto IPA-6 IP		44	NuMex BR-1 RC
17	16	Dessex SS	59	49	PS 8392 PS
20	19	El Ad HZ	62	63	RAM 710 HZ
21	20	Equanex PS	74		Redbone AS
24	22	Galil HZ	77		Ringer Grano SS
25	23	Gladalan Brown YA	80	66	Rio Bravo RC
26		Gladalan White YA	81	67	Rio Hondo RC
27	25	Granex 33 AS	82	68	Rio Raji Red RC
28	26	Granex 429 AS	83	69	Rio Ringo RC
33		HA-230 HZ	89	73	Savannah Sweet PS
34	30	HA-489 HZ	91	75	Sivan HZ
36	32	HA-817 HZ	92	76	Superex TK
38	33	HA-950 HZ	94		Texas Grano 438 AS
39	34	Houston AS	95		Texas Grano 502 PRR AS
40	35	Hurricane RS	96	79	Texas Grano LO
42		Jaguar PS	97	80	Tropic Ace TK
45	38	Linda Vista PS	98		Tropic Gold NW
47	40	Marathon HZ			

Table 9.3: Genotype codes used in the biplots of Figures 9.2 and 9.3. These genotypes were selected because their mean performance was in the top quarter of the range of results over the entire data set. The two columns of numbers indicate the genotype's number in each of Onion Data I and II; missing entries in these columns indicate that the genotype's mean yield was not in the top quarter of the range of genotype mean performances, or was not within Onion Data II.

als were 'L01101', 'W16508', and 'W17504' and were run in Kenya (1997), Brazil (1995), and Greece (1994) respectively. Unfortunately, the only time any of these three trials were clustered into the same mega-environment in Chapter 7 was in Figure 7.16, so this prediction seems unreliable.

Biplots should not be analysed individually when specific adaptation uses more than two significant PC axes. Scree plots are often used to gauge how many PC axes are significant, as they show how much of the interaction variation is explained by successive principal component axes. Figure 9.4 shows the contributions of the first ten principal components of Onion Data I and II. It shows that ten principal component axes could be used to explain over 75% of the G×E interaction for both Onion Data I and II. This number of axes is higher than is normally desired, as biplots and AMMI analyses are usually aimed at summarizing data in a simple form.

The ANOVA tables for the AMMI analyses of Onion Data I and II are presented in Tables 9.4 and 9.5 respectively. For both Onion Data I and II, the AMMI analysis showed that the first ten (at least) principal component axes were significant using the results

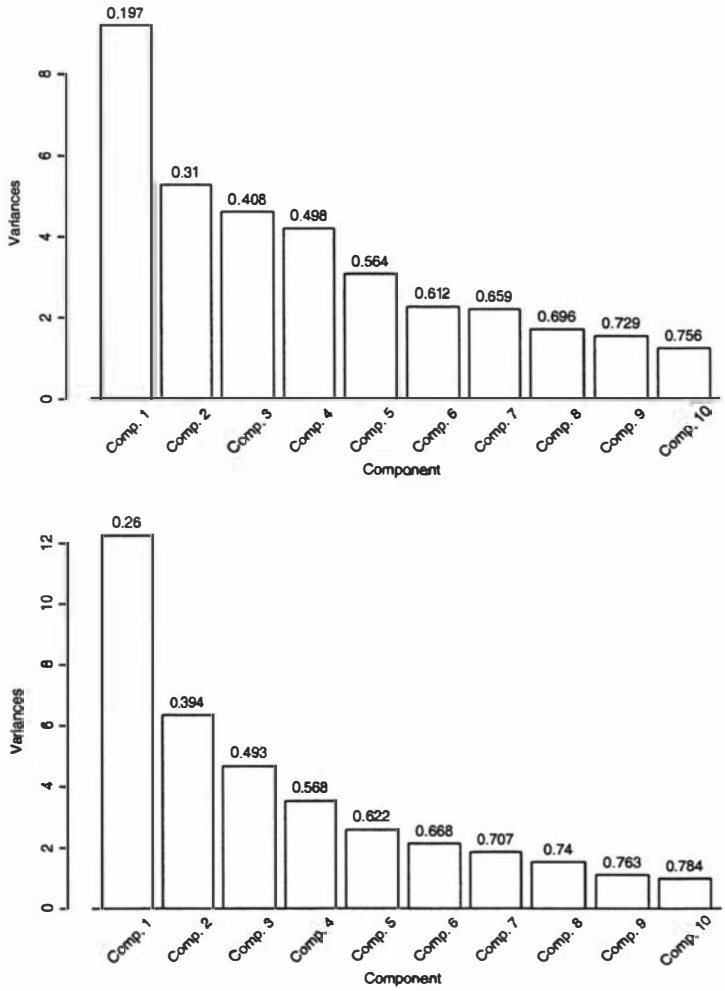


Figure 9.4: Scree plots showing the contributions of the first ten principal component axes for Onion Data I(top) and Onion Data II(bottom). Cumulative proportions of explained G×E interaction are shown as data labels.

of F-tests presented in the tables. Inclusion of all ten axes would result in a model that explained 87.6% and 88.0% of the total variation in Onion Data I and II respectively. The significance of the additive (no interaction terms) model showed that much of the total variation was due to genotypic main effects, because the environmental main effects were removed before imputation of Onion Data I and II. More interaction terms could have been added, but explaining the axes in real world terms would become even more difficult. Certainly, the task was not possible for the data arising from the Onion Trials Programme, as insufficient covariate information was available to explain so many significant principal components.

Without the ability to relate principal components axes to real world phenomena, the usefulness of the AMMI model and biplots as a tool for description of observed results is suspect. Selecting varieties for a new trial would require extrapolation of current results, and furthermore, is not possible if the new trial cannot be related to results for existing

Source	SS Explained	Model df	F ratio	Pr > F	R^2
AMMI(0)	4667	211	50.66	0.000	49.0
AMMI(1)	5623	421	37.38	0.000	59.0
AMMI(2)	6171	629	31.34	0.000	64.8
AMMI(3)	6650	835	29.11	0.000	69.8
AMMI(4)	7086	1039	28.81	0.000	74.4
AMMI(5)	7405	1241	28.44	0.000	77.8
AMMI(6)	7640	1441	27.86	0.000	80.2
AMMI(7)	7869	1639	28.14	0.000	82.6
AMMI(8)	8047	1835	28.22	0.000	84.5
AMMI(9)	8208	2029	28.63	0.000	86.2
AMMI(10)	8339	2221	28.89	0.000	87.6

Source	Seq SS	df	Resid SS	Resid df	F ratio	Pr > F
additive	4667	211	4856	11124	50.66	0.000
IPCA(1)	956	210	3900	10914	12.74	0.000
IPCA(2)	548	208	3352	10706	8.42	0.000
IPCA(3)	479	206	2873	10500	8.50	0.000
IPCA(4)	436	204	2437	10296	9.03	0.000
IPCA(5)	319	202	2118	10094	7.53	0.000
IPCA(6)	235	200	1883	9894	6.17	0.000
IPCA(7)	229	198	1654	9696	6.77	0.000
IPCA(8)	178	196	1476	9500	5.84	0.000
IPCA(9)	161	194	1315	9306	5.88	0.000
IPCA(10)	131	192	1185	9114	5.23	0.000

Table 9.4: Partial ANOVA for the first 11 AMMI models, applied to Onion Data I after imputation using the two-stage method. The second half of the table shows the contributions of successive interaction principal component axes (IPCA) after the AMMI(0) (or additive model)has been fitted.

environments.

This section showed results from applying graphical solution approaches to the data arising from the Onion Trials Programme. The first approach, aimed directly at the main focus of the research, was too simplistic and could only show which genotypes were suited to equatorial locations. The second and more common approach of biplotting had several problems:

1. There were too many genotypes and environments to include in each biplot to allow it to be read clearly.

2. There were too many significant interaction principal components to allow comparison of results using all biplots in combination.

3. The genotypic main effects counted for much of the total variation, but are not reflected in the biplots.

Source	SS Explained	Model df	F ratio	Pr > F	R^2
AMMI(0)	3281	183	36.46	0.000	44.4
AMMI(1)	4346	365	32.00	0.000	58.9
AMMI(2)	4897	545	28.85	0.000	66.3
AMMI(3)	5302	723	27.51	0.000	71.8
AMMI(4)	5609	899	26.83	0.000	76.0
AMMI(5)	5833	1073	26.15	0.000	79.0
AMMI(6)	6019	1245	25.82	0.000	81.5
AMMI(7)	6181	1415	25.86	0.000	83.7
AMMI(8)	6315	1583	25.93	0.000	85.5
AMMI(9)	6411	1749	25.58	0.000	86.8
AMMI(10)	6497	1913	25.37	0.000	88.0

Source	Seq SS	df	Resid SS	Resid df	F ratio	Pr > F
additive	3281	183	4102	8342	36.46	0.000
IPCA(1)	1065	182	3037	8160	15.73	0.000
IPCA(2)	551	180	2485	7980	9.83	0.000
IPCA(3)	405	178	2080	7802	8.54	0.000
IPCA(4)	307	176	1773	7626	7.50	0.000
IPCA(5)	224	174	1549	7452	6.20	0.000
IPCA(6)	186	172	1363	7280	5.77	0.000
IPCA(7)	162	170	1201	7110	5.64	0.000
IPCA(8)	133	168	1068	6942	5.16	0.000
IPCA(9)	97	166	971	6776	4.06	0.000
IPCA(10)	86	164	885	6612	3.92	0.000

Table 9.5: Partial ANOVA for the first 11 AMMI models, applied to Onion Data II after imputation using the two-stage method. The second half of the table shows the contributions of successive interaction principal component axes (IPCA) after the AMMI(0) (or additive model) has been fitted.

4. It was not possible to give real world meaning to all principal component axes.

Graphical solution approaches were abandoned in favour of numerical approaches, which are used in the remaining sections of this chapter to find variety selection for a new trial. The next section looks at selection of check varieties using wide adaptation stability measures, before the search for genotypes specifically adapted to the new location are found.

9.3 Genotype selections using stability measures

Check varieties have been used for many reasons in G×E analyses. Generally they are used to act as a set of varieties that have known, or relatively well understood, responses to environmental conditions. Many standard stability measures discussed in Section 2.7 can be applied to the two-stage imputed data found in Section 6.5. Measures for determining which genotypes have the best wide adaptability will give trial programme organizers a

means of selecting genotypes that should act as 'check varieties' in future trials. The methods employed to determine these varieties in this section are:

1. A comparison of genotype mean and standard deviation, as achieved by the use of the adjusted coefficient of variation $CV_i^\star$ presented in Section 3.7.
2. The Adjusted Wricke's ecovalence $\hat{W}_i^\star$ as presented in Section 3.7.
3. The adjusted Lin and Binns superiority score, $p_i^\star$ as presented in Section 3.7.
4. Various nonparametric stability measures introduced in Section 2.7, two of which were cited by Becker and Leon (1988) and two given by Hühn and Nassar (1989).

Selection of check varieties using the adjusted coefficient of variation

The adjusted coefficient of variation was used in Section 3.7 to identify genotypes that contributed heavily to the G×E interaction of Onion Data I and II. In this section, the aim is to identify varieties that have good wide adaptation through various stability measures, so the results are not directly comparable to those presented in Section 3.7.

In Section 7.3, the two-stage imputed yields for Onion Data I and II were used for forming mega-environments. In Figure 7.6 genotype means and standard deviations for the imputed data were plotted against one another. Check varieties would be chosen from among the genotypes that had low standard deviations, and would preferably also have average to above average mean yields.

Table 9.6 shows the adjusted coefficients of variation $CV_i^\star$ for the best fifteen genotypes of Onion Data I and II. Negative values arose as a consequence of negative imputed yields which were relative to environmental means. The results between Onion Data I and II show little consistency, as seen by the changes in rankings of the genotypes in Table 9.6.

Trials Programme organizers would be advised to select varieties 'Marix ZU', 'Jenin HZ', and 'Agrifound Rose AF' as check varieties using the adjusted coefficient of variation. These three varieties had negative mean imputed yields. If varieties that were above average in terms of their mean yield were required, the varieties 'Dessex SS', 'Superex TK', and 'Rio Bravo RC' would be recommended instead. The varieties 'Nasik Red LO' and 'Cadix ZU' look promising but were not used in enough environments to have been included in Onion Data II. If they are selected for use in future trials, they may become candidates for use as check varieties using this criterion.

Selection of check varieties using adjusted Wricke's ecovalence

Wricke's ecovalence has been used seldom in modern G×E analyses. It does, however, have a simple construction, and because it was used in Section 3.7, has been used to select check varieties. Table 9.7 shows the adjusted Wricke's ecovalence scores $\hat{W}_i^\star$ for the best fifteen genotypes of Onion Data I and II. The consistency between data sets appears low

Genotype	Onion Data I Score (Rank)		Onion Data II Score (Rank)	
Nasik Red LO	–40.89	(1)		
Marix ZU	–41.40	(2)	–35.55	(1)
Cadix ZU	–47.17	(3)		
Dessex SS	52.78	(4)	78.10	(13)
Jenin HZ	–54.74	(5)	–59.14	(2)
Agrifound Rose AF	–56.07	(6)	–68.20	(7)
N-53 LO	–57.40	(7)		
PS 8392 PS	59.34	(8)		
Superex TK	59.91	(9)	64.48	(4)
Rio Bravo RC	60.15	(10)	67.25	(6)
Nikita RC	60.34	(11)		
Granex 429 AS	61.62	(12)	75.42	(11)
Australian Brown ST	–65.33	(13)	–78.93	(15)
Deko HZ	–68.12	(14)	–97.81	(20)
Savannah Sweet PS	69.02	(15)	76.34	(12)
Hurricane RS	70.09	(16)	66.74	(5)
Creamgold YA	–71.19	(17)	–60.58	(3)
Tropicana RS	–72.09	(18)	–69.17	(9)
Marathon HZ	73.22	(19)	69.13	(8)
Mercedes PS	75.10	(21)	78.76	(14)
Red Creole AS	–118.69	(42)	–69.51	(10)

Table 9.6: Adjusted coefficients of variation for the top fifteen genotypes from each of Onion Data I and II. The two-stage imputed data found in Section 6.5 was used in the same manner as the analysis in Section 3.7 which used the sparse data. The ranks given in brackets show how the scores vary from one data set to the other.

as a total of 24 genotypes are listed. Many of the top genotypes from one data set have less than desirable scores when looking at the other data set. Figure 9.5 shows that the distributions of the ecovalences for Onion Data I and II are noticeably different.

Considering both sets of results in conjunction to deal with the inconsistency, the recommendations would probably include 'Dessex SS', 'Rio Bravo RC', and 'Equanex PS'. If such varieties as 'Nasik Red LO' and 'N-53 LO' were used in future trials, their wide adaptability could be reinvestigated to see if they should be used as 'check varieties'.

Selection of check varieties using the adjusted Lin and Binns superiority score

The adjusted Lin and Binns superiority scores p_i^* for the two-stage imputed G×E matrices of Onion Data I and II are presented in Table 9.8. Again, results for the top fifteen genotypes for each data set are shown, and there is noticeable variation between the rankings found for Onion Data I and II. Figure 9.6 shows some similarity in terms of the superiority scores' distributions.

The three best genotypes using the adjusted Lin and Binns superiority score were

Genotype	Onion Data I		Onion Data II	
	Score	(Rank)	Score	(Rank)
Nasik Red LO	0.25	(1)		
Dessex SS	0.25	(2)	0.29	(1)
N-53 LO	0.25	(3)		
Rio Bravo RC	0.28	(4)	0.31	(4)
Equanex PS	0.28	(5)	0.33	(6)
Hurricane RS	0.28	(6)	0.36	(13)
Red Bombay RS	0.29	(7)	0.45	(43)
Jaguar PS	0.30	(8)		
Superex TK	0.31	(9)	0.34	(8)
HA-891 HZ	0.31	(10)		
Marathon HZ	0.32	(11)	0.31	(5)
Deko HZ	0.32	(12)	0.47	(46)
Bon Accord HT	0.32	(13)	0.45	(42)
PS 8392 PS	0.33	(14)		
Primero SS	0.33	(15)		
Pera IPA-4 IP	0.34	(18)	0.34	(9)
HA-817 HZ	0.34	(19)	0.34	(7)
Marix ZU	0.37	(29)	0.30	(3)
Granex 33 AS	0.37	(30)	0.36	(15)
Tropicana RS	0.38	(31)	0.36	(12)
Savannah Sweet PS	0.38	(36)	0.35	(11)
Rio Ringo RC	0.41	(50)	0.30	(2)
Texas Grano LO	0.48	(70)	0.35	(10)
Rio Blanco Grande RC	0.50	(79)	0.36	(14)

Table 9.7: Wricke's ecovalence scores for the top fifteen genotypes from each of Onion Data I and II. The two-stage imputed data found in Section 6.5 was used in the same manner as the analysis in Section 3.7 which used the sparse data. The ranks given in brackets show how the scores vary from one data set to the other.

'Dessex SS', 'Granex 429 AS', and 'Superex TK', although the varieties 'Nikita RC' and 'PS 8392 PS' show promise if only results from Onion Data I are considered. In general, adjusted superiority scores changed as data from more environments were added, therefore use of this criterion for selection of check varieties is likely to mean that the set will change from year to year.

The correlation between the adjusted Lin and Binns superiority scores for the sparse and two-stage imputed yields for Onion Data I and II were 0.744 and 0.735 respectively. There were ten genotypes in common between the lists of the top fifteen genotypes found using the sparse data (Table 3.8) and the two-stage imputed data (Table 9.8) of both Onion Data I and II. This would suggest that the sparsity of data and its subsequent imputation had little impact on the adjusted Lin and Binns superiority score.

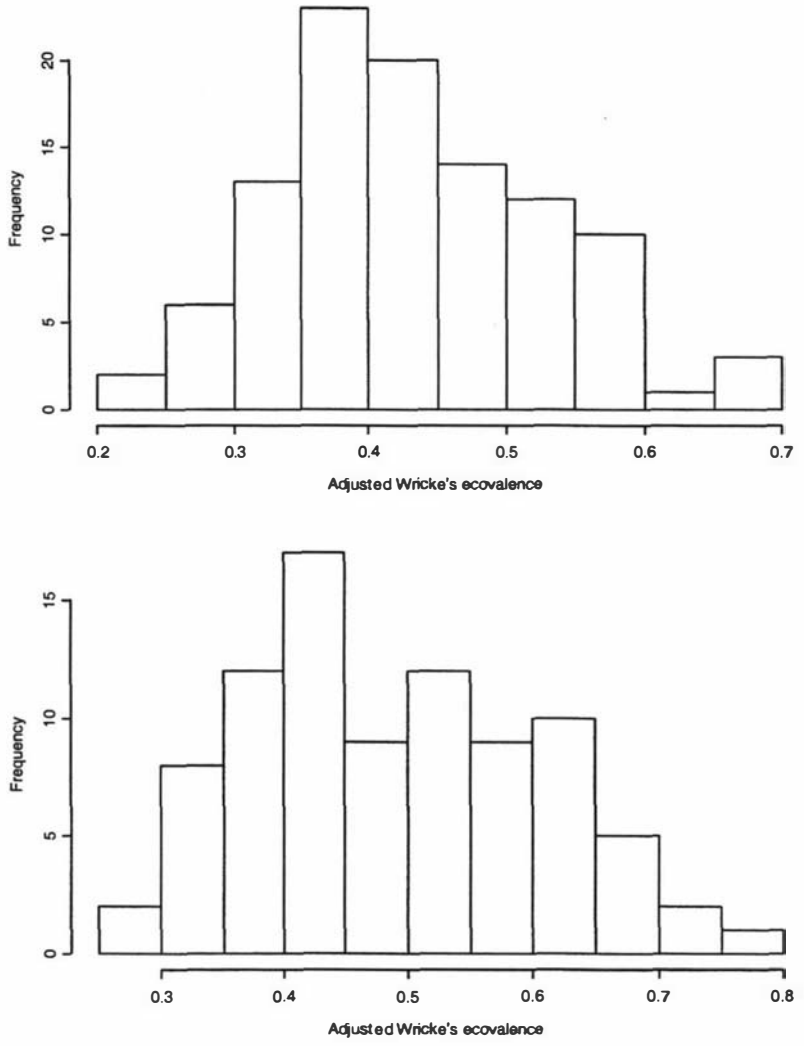


Figure 9.5: Histograms of Wricke's ecovalence scores for the 104 genotypes of Onion Data I (top) and 87 genotypes of Onion Data II (bottom).

Selection of check varieties using various nonparametric stability measures

The nonparametric stability measures introduced in Section 2.7 were not used on the analysis of the sparse data in Section 3.7 because there was no simple modification of the parameters' formulation that was employable. Instead of taking ranks within environments, as needed for the measures $s_i^{(1)}$, $s_i^{(2)}$, $s_i^{(3)}$, and $s_i^{(6)}$ introduced on page 45, range standardization could have been used. All scores would then be on the scale zero (for the worst performing genotype) to one (for the best performing genotype). This was not presented as it was not necessary in that analysis to show how every stability parameter could be adapted for use in incomplete data scenarios, but rather to show that their adaptation raised issues when working with sparse data.

The complete data resulting from two-stage imputation was used to select check varieties by applying four nonparametric stability measures. These were $s_i^{(1)}$ and $s_i^{(2)}$ cited by Becker and Leon (1988) and $s_i^{(3)}$ and $s_i^{(6)}$ proposed by Hühn and Nassar (1989).

Genotype	Onion Data I		Onion Data II	
	Score	(Rank)	Score	(Rank)
Nikita RC	0.76	(1)		
PS 8392 PS	0.84	(2)		
Dessex SS	0.86	(3)	1.18	(7)
Granex 429 AS	0.88	(4)	1.04	(5)
Superex TK	0.91	(5)	0.93	(1)
Linda Vista PS	1.01	(6)	1.32	(12)
Rio Bravo RC	1.03	(7)	0.97	(3)
Hurricane RS	1.04	(8)	0.94	(2)
Savannah Sweet PS	1.05	(9)	1.20	(8)
Mercedes PS	1.10	(10)	1.08	(6)
HA-230 HZ	1.12	(11)		
Marathon HZ	1.14	(12)	0.99	(4)
Belem IPA-9 IP	1.23	(13)	1.28	(10)
Equanex PS	1.25	(14)	1.48	(16)
Colossal PVP SS	1.34	(15)	1.89	(26)
Gladalan Brown YA	1.47	(17)	1.29	(11)
Arad HZ	1.50	(18)	1.43	(14)
Houston AS	1.53	(19)	1.20	(9)
HA-950 HZ	1.55	(21)	1.38	(13)
Rio Ringo RC	1.88	(31)	1.44	(15)

Table 9.8: Adjusted Lin and Binns superiority scores for the two-stage imputed G×E matrices of Onion Data I and II. The top fifteen results for each data set are shown, with the ranks for scores given in parentheses. Three of the top fifteen genotypes of Onion Data I were not part of Onion Data II. Some genotypes were listed in the top fifteen for one, but not both data sets. In these cases the scores found using the other data set are provided for comparison.

Kang and Pham (1991) observed in their study that $s_i^{(3)}$ and $s_i^{(6)}$ were correlated, but would be useful for "simultaneously selecting for yield and yield stability". These scores were also highly correlated for the data arising from the Onion Trials Programme, as were those for $s_i^{(1)}$ and $s_i^{(2)}$. The $s_i^{(1)}$ $s_i^{(2)}$ scores had correlations of 0.970 and 0.952 for Onion Data I and II respectively. Likewise, $s_i^{(3)}$ and $s_i^{(6)}$ were also highly correlated (0.951 and 0.964 respectively). Tables 9.9 to 9.12 therefore present the top fifteen results for the correlated pairs of nonparametric stability measures for Onion Data I and II.

Using $s_i^{(1)}$ and $s_i^{(2)}$ scores, the best variety from both Onion Data I and II is 'Marix ZU', and the next best in terms of consistency between data sets was 'Jenin HZ'. Many of the other top varieties for Onion Data I were either not included in Onion Data II, or had much poorer scores and were not included in Table 9.12. Results from Tables 9.11 and 9.12 it can be seen that $s_i^{(3)}$ and $s_i^{(6)}$ scores were more consistent. On the grounds of these scores, the four varieties 'Dessex SS', 'Hurricane RS', 'Rio Bravo RC', and 'Superex TK' were in the top five varieties of both Onion Data I and II.

Genotype	$s_i^{(1)}$	(Rank)	$s_i^{(2)}$	(Rank)
Marix ZU	9.52	(1)	153.45	(1)
Nasik Red LO	11.64	(2)	176.24	(2)
Cadix ZU	12.22	(3)	201.49	(3)
Dessex SS	15.99	(7)	239.16	(4)
N-53 LO	15.47	(6)	259.17	(5)
Jenin HZ	14.54	(4)	283.02	(7)
Deko HZ	17.55	(9)	275.01	(6)
Nikita RC	17.46	(8)	306.61	(10)
Hurricane RS	17.65	(10)	293.25	(8)
Rio Bravo RC	18.22	(13)	302.45	(9)
Tropicana RS	17.82	(11)	326.74	(13)
Agrifound Rose AF	15.40	(5)	365.69	(21)
Superex TK	17.88	(12)	345.42	(15)
Red Bombay RS	18.80	(16)	350.74	(16)
Australian Brown ST	18.49	(14)	356.30	(19)

Table 9.9: Fifteen genotypes from Onion Data I, most suitable for use as check varieties using the nonparametric measures $s_i^{(1)}$ and $s_i^{(2)}$ cited by Becker and Leon (1988) .

Genotype	$s_i^{(1)}$	(Rank)	$s_i^{(2)}$	(Rank)
Marix ZU	6.72	(1)	88.61	(1)
Tropicana RS	14.71	(3)	198.35	(2)
Creamgold YA	14.20	(2)	241.28	(4)
Hurricane RS	15.32	(5)	241.64	(5)
Marathon HZ	16.45	(9)	235.99	(3)
Jenin HZ	14.84	(4)	265.25	(10)
Superex TK	16.04	(6)	249.57	(9)
Red Creole AS	16.08	(7)	245.52	(8)
Rio Bravo RC	16.82	(10)	243.93	(6)
Dessex SS	17.58	(11)	244.54	(7)
Red Synthetic HZ	18.17	(16)	266.59	(11)
Savannah Sweet PS	17.86	(13)	291.55	(15)
Red Bombay RS	18.12	(14)	294.81	(16)
Yellow Creole SS	18.41	(18)	282.04	(13)
Red Star PS	18.16	(15)	296.05	(17)

Table 9.10: Fifteen genotypes from Onion Data II, most suitable for use as check varieties using the nonparametric measures $s_i^{(1)}$ and $s_i^{(2)}$ cited by Becker and Leon (1988) .

Genotype	$s_i^{(3)}$	(Rank)	$s_i^{(6)}$	(Rank)
Dessex SS	2.79	(1)	14.25	(1)
Nikita RC	3.58	(2)	15.48	(2)
Hurricane RS	3.69	(4)	16.52	(4)
Rio Bravo RC	3.59	(3)	17.33	(5)
Superex TK	4.07	(5)	16.29	(3)
Linda Vista PS	4.40	(8)	17.52	(6)
Granex 429 AS	4.11	(6)	18.93	(9)
PS 8392 PS	4.45	(9)	17.63	(7)
Equanex PS	4.17	(7)	20.59	(10)
Mercedes PS	4.91	(11)	18.81	(8)
Marathon HZ	4.53	(10)	20.85	(11)
Savannah Sweet PS	5.02	(12)	21.26	(13)
HA-230 HZ	5.31	(13)	21.25	(12)
Colossal PVP SS	5.33	(14)	21.95	(14)
Jaguar PS	5.57	(15)	22.88	(15)

Table 9.11: Fifteen genotypes from Onion Data I, most suitable for use as check varieties using the nonparametric measures $s_i^{(3)}$ and $s_i^{(6)}$ given by Hühn and Nassar (1989).

Genotype	$s_i^{(3)}$	(Rank)	$s_i^{(6)}$	(Rank)
Hurricane RS	3.42	(2)	15.06	(1)
Marathon HZ	3.40	(1)	17.07	(3)
Superex TK	3.54	(4)	16.43	(2)
Rio Bravo RC	3.53	(3)	17.38	(4)
Dessex SS	3.78	(5)	19.25	(6)
Savannah Sweet PS	4.31	(6)	18.69	(5)
Granex 429 AS	4.81	(8)	19.32	(7)
Mercedes PS	4.82	(9)	19.64	(8)
Gladalan Brown YA	4.83	(10)	20.65	(9)
Rio Ringo RC	4.72	(7)	22.88	(13)
Belem IPA-9 IP	5.01	(12)	21.89	(12)
Houston AS	5.30	(13)	21.81	(11)
Equanex PS	4.99	(11)	23.31	(15)
Linda Vista PS	5.75	(18)	20.95	(10)
Texas Grano LO	5.39	(14)	24.43	(17)

Table 9.12: Fifteen genotypes from Onion Data II, most suitable for use as check varieties using the nonparametric measures $s_i^{(3)}$ and $s_i^{(6)}$ given by Hühn and Nassar (1989).

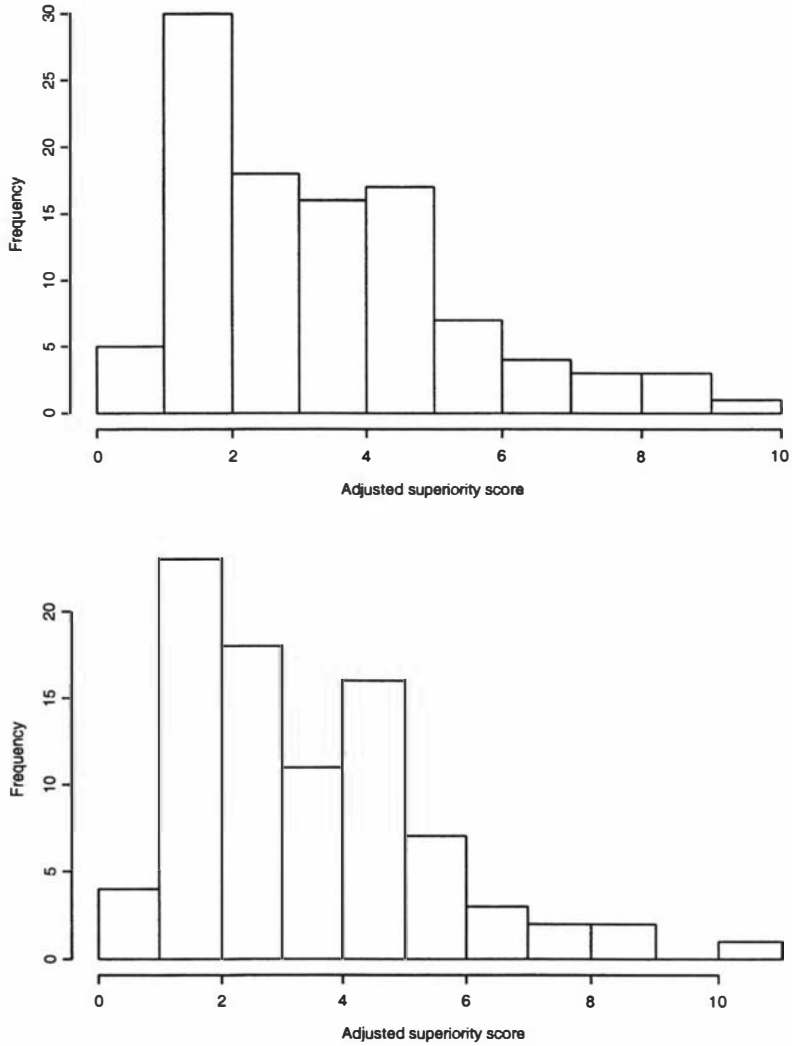


Figure 9.6: Histograms of adjusted Lin and Binns superiority scores for the 104 genotypes of Onion Data I (top) and 87 genotypes of Onion Data II (bottom).

Comparison of check variety selections

The methods for selecting check varieties presented in this section gave mixed results. Some varieties, notably 'Dessex SS', 'Rio Bravo RC', and 'Superex TK' were recommended as check varieties by more than one stability measure, but the measures used in this section were often highly correlated. Table 9.13 lists the correlations between the seven stability measures employed for the selection of check varieties. Absolute values of the $CV_i^\star$ scores were used for calculating their correlations with other scores to remove the problem caused by negative scores.

Notable correlations in this table include the fact that the adjusted superiority score is highly positively correlated with $s_i^{(3)}$ and $s_i^{(6)}$, whereas it was negatively correlated with $s_i^{(1)}$ and $s_i^{(2)}$. The nonparametric measures were correlated within pairs, and slightly negatively or uncorrelated between pairs. There should therefore be two sets of check variety recommendations resulting from these measures. If $p_i^\star$, $s_i^{(3)}$, and $s_i^{(6)}$ are used, 'Dessex SS' and 'Superex TK' are recommended. If on the other hand, $s_i^{(1)}$ and $s_i^{(2)}$ are

	$CV_i^\star$	$\hat{W}_i^\star$	$p_i^\star$	$s_i^{(1)}$	$s_i^{(2)}$	$s_i^{(3)}$	$s_i^{(6)}$
						Onion Data I	
$CV_i^\star$		0.050	−0.032	0.165	0.150	0.004	−0.032
$\hat{W}_i^\star$	0.226		0.231	0.625	0.747	0.591	0.380
$p_i^\star$	−0.052	0.376		−0.371	−0.267	0.883	0.973
$s_i^{(1)}$	0.416	0.524	−0.355		0.970	0.025	−0.214
$s_i^{(2)}$	0.388	0.695	−0.223	0.952		0.161	−0.100
$s_i^{(3)}$	0.045	0.679	0.893	0.011	0.175		0.951
$s_i^{(6)}$	−0.016	0.513	0.972	−0.196	−0.053	0.964	
	Onion Data II						

Table 9.13: Correlations for the wide adaptability stability measures used in Section 9.3 for each of Onion Data I (upper triangle) and Onion Data II (lower triangle). Note that the correlations for $CV_i^\star$ were calculated using the absolute value of these scores to remove the problem caused by negative scores.

used, the varieties 'Marix ZU' and 'Jenin HZ' would be recommended.

The recommendations for check varieties are dependent on the stability measures employed. They are, however, also dependent on the environments over which the measures were calculated. In Chapters 7 and 8, mega-environment memberships were shown to be inconsistent as new trials were added to the programme. A consequence of adding more environment data is that the weightings certain types of environments have on results will be altered, and the stability measures are likely to change. The check variety selections are therefore made on the grounds of current knowledge, and may be altered as new knowledge is gained.

On purely statistical grounds the set of check varieties can be found in any number of ways, some of which were presented in this section. There is a need to consider other issues that may or may not mean that the above selections are actually recommended. Collaborator preferences need to be considered, as well as the availability of sufficient seed to send to all new trials. Such considerations are discussed further in Chapter 10. The following sections investigate the selection of varieties for testing in specific environments in order to eventually determine which varieties succeed in which environments.

9.4 Genotype selections using geographically similar environments

A simple method for identifying the best genotypes for a new trial is to use only the data arising from similar geographic locations. For example, trial programme organizers might want to select genotypes for testing in a new trial located in Yemen which mirrors the trial A01606, at 600 metres above sea level and 15.9° north of the equator. This new location has plenty of space and can grow fifteen genotypes. Which genotypes should be sent to

Onion Data I		Onion Data II	
Genotype	Performance	Genotype	Performance
Granex 429 AS	1.05	Gladalan Brown YA	1.29
PS 8392 PS	1.05	Marathon HZ	1.20
Houston AS	1.02	HA-489 HZ	1.11
Gladalan Brown YA	0.99	Mercedes PS	1.09
Dessex SS	0.97	Hurricane RS	1.04
Marathon HZ	0.97	Arad HZ	1.02
HA-230 HZ	0.93	El Ad HZ	1.01
Texas Grano 438 AS	0.93	Superex TK	0.93
HA-489 HZ	0.91	Houston AS	0.89
Savannah Sweet PS	0.91	RAM 710 HZ	0.86
Nikita RC	0.90	Savannah Sweet PS	0.83
Mr Max RC	0.88	Rio Hondo RC	0.82
Mercedes PS	0.88	Granex 33 AS	0.80
Superex TK	0.87	Granex 429 AS	0.79
Colossal PVP SS	0.84	Colossal PVP SS	0.76

Table 9.14: Names of the fifteen genotypes selected for testing in the new Yemeni environment based on their imputed results across geographically similar environments.

the collaborator?

Data from only those environments that are within three degrees of latitude, including southern hemisphere latitudes of the same magnitude, of this new Yemeni location were used. Of these environments, only those within 250 metres of altitude were considered. These arbitrary limits for consideration left nine and seven environments from Onion Data I and II respectively. Taking the average of the imputed scores for genotypes from Onion Data I and II across these geographically similar environments gives 'average relative performance' results, the top fifteen of which are presented in Table 9.14.

The average relative performances for Onion Data I and II are presented in histograms in Figure 9.7. They show that the best fifteen genotypes had average relative performances well above the majority of genotypes. Of the fifteen varieties listed as recommended using Onion Data I, two were not included in Onion Data II, and four more do not appear in the list of recommendations based on Onion Data II. It would be advisable to recommend that the varieties 'Nikita RC' and 'Texas Grano 438 AS' are tested in the new environment as a means of reducing the number of varieties included in Onion Data I that are not in Onion Data II.

The remaining nine varieties that were included in the recommendations based on both Onion Data I and II should be included in the final recommendations for the new Yemeni trial using this simple approach. The next section uses the information from Chapter 7 to show how selections could be based on the data from the right mega-environment.

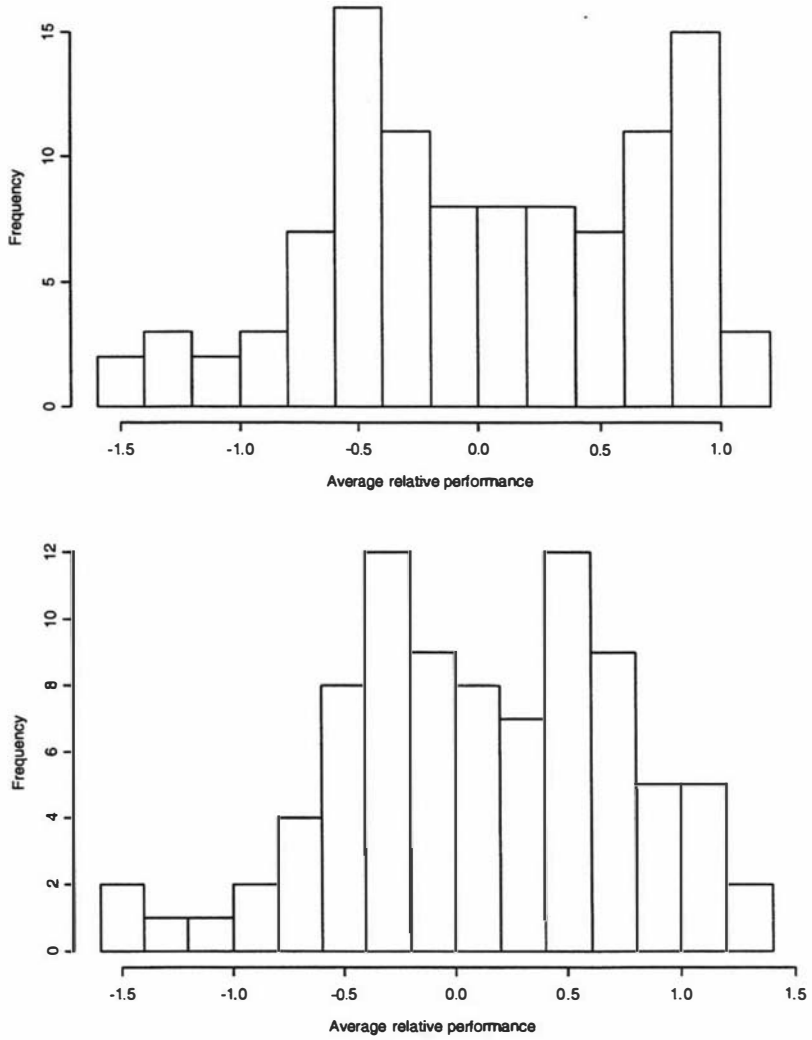


Figure 9.7: Average relative performances of 104 and 87 genotypes from Onion Data I and II, respectively, measured over environments within 250 metres of altitude and three degrees of latitude to the new Yemeni trial.

9.5 Genotype selections using mega-environments

The clustering of 101 environments for which proxy covariates were calculable was presented in Section 7.4. The fifteen mega-environments can now be used to select subsets of environments to relate to a new environment. In the previous section a fictitious Yemeni environment was used as an example and will be used in this section for comparability. This new trial uses the covariate information from the trial A01606, which was not included in Onion Data I or II.

When mega-environments were formed using the 101 environments for which covariate information was available, this trial was placed in mega-environment 13, along with environments from Nigeria, Malaysia, Papua New Guinea, Belize, Bangladesh, and Mexico. Average relative performances were calculated for genotypes in Onion Data I and II over only five environments, as another of the environments listed in Table 9.1 is included in mega-environment 13. It would therefore be given the same recommendations for variety

Onion Data I		Onion Data II	
Genotype	Performance	Genotype	Performance
Rio Bravo RC	0.96	Rio Bravo RC	1.07
Dessex SS	0.80	Savannah Sweet PS	0.75
Equanex PS	0.71	Equanex PS	0.72
Ringer Grano SS	0.70	Rio Ringo RC	0.66
Colossal PVP SS	0.67	Dessex SS	0.62
Granex 429 AS	0.67	Houston AS	0.59
Savannah Sweet PS	0.65	Granoble PS	0.59
Galil HZ	0.59	Marathon HZ	0.57
Rio Raji Red RC	0.56	Gladalan Brown YA	0.56
Texas Grano 438 AS	0.56	Belem IPA-9 IP	0.55
Gladalan Brown YA	0.54	Linda Vista PS	0.53
PS 8392 PS	0.53	Galil HZ	0.45
Marathon HZ	0.53	Superex TK	0.45
Houston AS	0.52	HA-489 HZ	0.43
Texas Grano 502 PRR AS	0.48	Arad HZ	0.42

Table 9.15: Genotypes selected for their suitability in the new trial given their high average relative performance in mega-environment 13, which included the trials A00501, A03101, A03401, A04103, and D08401.

selections as the new Yemeni trial if this method was used. Results for the top fifteen genotypes are presented in Table 9.15.

The average relative performances for Onion Data I and II are presented in histograms in Figure 9.8.

Of the fifteen varieties recommended using the results from Onion Data I, two genotypes were not included in Onion Data II and six more were not in the top fifteen genotypes from Onion Data II. It would be safest to use the seven genotypes common to the results from Onion Data I and II in order to limit the potential of poor imputations to impact on recommendations. In this example, only the imputed values for Onion Data I and II can alter the average relative performances because the same set of five environments was used.

Comparing the lists of varieties from Tables 9.14 and 9.15 showed that seven Onion Data I genotypes were recommended by both methods, while eight Onion Data II genotypes were recommended by both methods. The variety 'Nikita RC' was also recommended by both methods, but was not tested in sufficient environments to have been included in Onion Data II. As noted previously, testing this genotype in the new trial would mean that its results would be included in a data set constructed using the same criteria as Onion Data II.

In this and the preceding section the fully imputed G×E matrices were used. The next section investigates the impact of limiting the imputations to only information gained from first stage clustering.

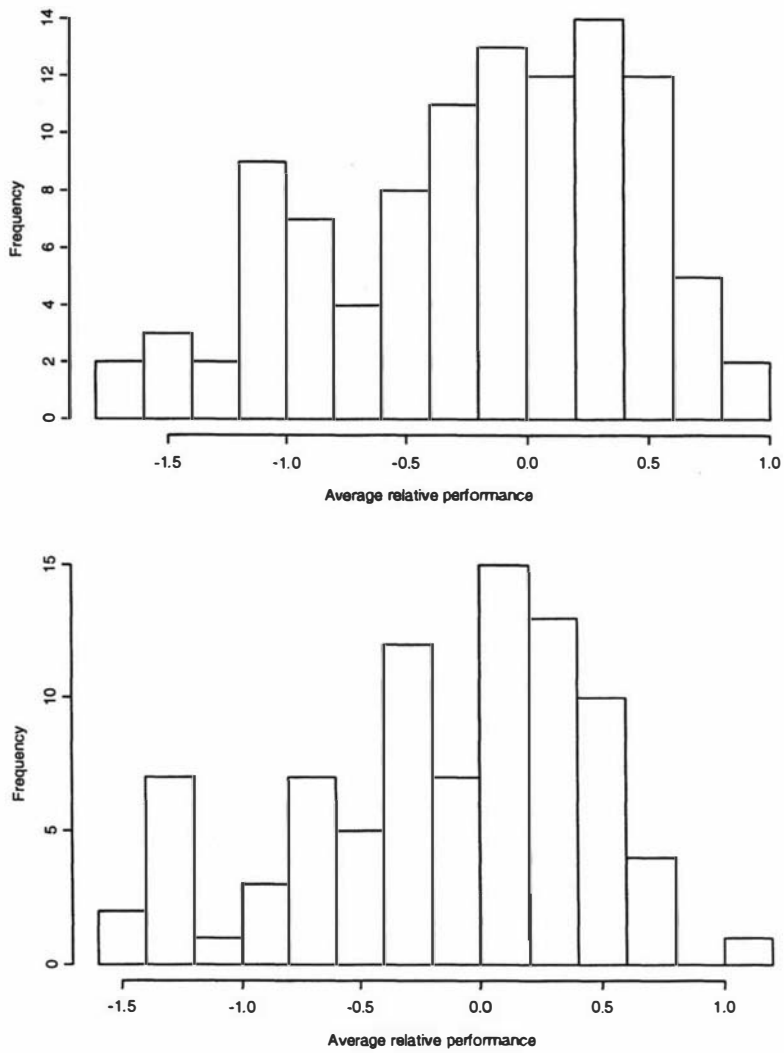


Figure 9.8: Average relative performances of 104 and 87 genotypes from Onion Data I and II, respectively, measured over environments from mega-environment 13, which would include the new Yemeni trial.

9.6 Results using partial imputation

The portions of G×E matrices used in the previous sections were complete, because the first step of the two-stage imputation process, described on page 147, searches for inter-genotype relationships until an imputation is possible. The selections made for the new environment treated the imputed and observed values as equally valid. It could be argued that many of the values imputed using data from varieties in a different first stage cluster , are less reliable than either those imputed using yields of varieties that are in the same first stage cluster, or those that were actually observed. This section investigates the impact on the selections given in the previous section when averaging was performed over only those values found by imputations based on yields of genotypes in the same first stage cluster.

For comparability, the example used in Section 9.5 was repeated, using imputations based on the relationships developed by first stage clustering. The trimming process of the

Data set	Genotypes
Onion Data I	'Belem IPA-9 IP', 'Granex 33 AS', 'Granex 429 AS', 'Savages Flat White YA'*, 'Tropic Gold NW', and 'Violet de Galmi TS'.
Both Onion Data I and II	'Gladalan White YA', 'IRAT-69 MA', 'Mercedes PS', and 'RAM 710 HZ'.
Onion Data II	'Colossal PVP SS', 'Creamgold YA', 'Houston AS', 'Lockyer Gold NW', 'Red Creole PRR PVP SS', 'RS 209 RS'.

Table 9.16: List of the ten genotypes for each of Onion Data I and II that were not given any imputed results in mega-environment 13 using partial imputation, so could not have average relative performances calculated. Note that all Onion Data II genotypes are in Onion Data I, but that of the six Onion Data I genotypes listed, only 'Savages Flat White YA' was not included in Onion Data II.

final step of two-stage imputation was performed, and affected a relatively small number of imputed values.

This 'partial imputation' strategy failed to give results for some genotypes in some mega-environments due to the sparsity of Onion Data I and II. In particular, ten varieties were not imputed in any of the five environments of mega-environment 13 for either Onion Data I or II. The names of these varieties appear in Table 9.16. Of these varieties, four were not given an imputed value in any mega-environment 13 environment, by imputation based on Onion Data I or II. There were five and six varieties, for Onion Data I and II respectively, that were given an average relative performance when using the other data set. Sixteen of the seventeen Onion Data I genotypes not included in Onion Data II were given an average relative performance when considering the Onion Data I results.

On occasion, a genotype chosen using fully imputed data has not been selected using partially imputed data. The variety 'Galil HZ', for example, has a poor ranking for partially imputed results (49 and 40 for Onion Data I and II respectively), but does have a good ranking using fully imputed results (8 and 12 for Onion Data I and II respectively). It, therefore, does not appear in the results for the best fifteen varieties selected for the new Yemeni trial based on the partially imputed Onion Data I and II matrices, presented in Table 9.17.

The relevant performances for genotypes from results based on the fully imputed Onion Data I and II G×E matrices have also been included in Table 9.17 for comparability. The consistency between data sets was low, and the partially imputed and fully imputed results also differed markedly.

The differences caused by choosing to partially or fully impute the G×E matrices created a dilemma. One of the following options must be chosen:

1. The results using fully imputed data matrices are trusted, regardless of the tenuous relationships among genotypes used to obtain the imputations. In this case an aver-

	Onion Data I				Onion Data II			
	Partially imputed		Fully imputed		Partially imputed		Fully imputed	
Genotype	ARP	Rank	ARP	Rank	ARP	Rank	ARP	Rank
Agrifound Rose AF	1.32	(1)	−1.41	(100)	1.21	(3)	−1.32	(83)
Brownsville AS	1.32	(2)	−0.60	(77)	1.21	(4)	0.00	(44)
Redbone AS	1.24	(3)	0.34	(24)	−1.42	(71)	0.18	(32)
Serrana AS	1.24	(4)	0.06	(41)	0.69	(14)	−0.35	(61)
Rojo SS	1.20	(5)	−0.10	(52)	−0.28	(49)	0.13	(33)
Nikita RC	1.12	(6)	0.32	(26)		()		()
PS 8392 PS	1.10	(7)	0.53	(12)		()		()
Yellow Granex Imp PRR SS	1.10	(8)	−0.02	(46)	0.65	(17)	0.25	(27)
Dessex SS	0.98	(9)	0.80	(2)	1.21	(5)	0.62	(5)
Red Burgundy Imp NE	0.90	(10)	0.09	(39)	−0.01	(42)	0.19	(31)
Texas Grano LO	0.88	(11)	0.05	(43)	1.45	(1)	−0.02	(45)
Rio Bravo RC	0.86	(12)	0.96	(1)	0.85	(10)	1.07	(1)
Equanex PS	0.84	(13)	0.71	(3)	0.84	(11)	0.72	(3)
Jaguar PS	0.78	(14)	0.14	(37)		()		()
Utopia AS	0.72	(15)	−0.25	(62)	−0.30	(50)	0.01	(42)
Marathon HZ	0.68	(17)	0.53	(13)	1.09	(8)	0.57	(8)
Tropic Ace TK	0.54	(25)	−0.03	(47)	0.79	(12)	0.05	(38)
Arad HZ	0.50	(26)	0.40	(21)	0.74	(13)	0.42	(15)
Savannah Sweet PS	0.45	(28)	0.65	(7)	0.92	(9)	0.75	(2)
Linda Vista PS	0.12	(42)	0.37	(22)	0.67	(15)	0.53	(11)
Regal PVP SS	0.11	(44)	0.01	(45)	1.28	(2)	0.32	(20)
Belem IPA-9 IP		()	0.47	(16)	1.10	(7)	0.55	(10)
Granex 429 AS		()	0.67	(6)	1.21	(6)	0.37	(16)

Table 9.17: Average relative performances (ARP) for the partially and fully imputed G×E matrices of Onion Data I and II. Ranks within results have been provided in parentheses to allow comparison.

age relative performance can be estimated for all genotypes, allowing all genotypes an opportunity to be selected for a new trial using any criterion.

2. Only imputed values based on first stage clustering are used. This would result in many G×E combinations not having imputed values, and therefore, potentially leaving the analysis without an estimate for these genotypes.

The risks involved in this decision are simple. Either risk overlooking a genotype by not having an estimate at all, or recommend the wrong set of genotypes because the estimated average relative performances are based on poor imputed values.

If all genotypes had been tested in at least one environment within every mega-environment, there would be a better case for only using the imputations based on first stage clustering. Given that this has not occurred in the data sets under consideration, recommendations must be made using the fully imputed data. Some of these recommendations may be based on what may turn out to be poor imputed values when more data are available. Testing these genotypes in a new trial will show trials programme organizers and collaborators whether imputed values were good estimates of the unobserved G×E combinations. Once these recommendations have been tested, the new data can be included in updated data sets, allowing new imputations to be made for the already tested environments.

9.7 Summary

This chapter developed selections for a new (fictitious) trial, using various approaches, namely:

1. Check varieties were found in Section 9.3 by investigating genotype responses across all environments, using:

   (a) The adjusted coefficient of variation $CV_i^\star$.

   (b) Adjusted Wricke's ecovalence $\hat{W}_i^\star$.

   (c) The adjusted Lin and Binns superiority score $p_i^\star$.

   (d) Four nonparametric stability measures; $s_i^{(1)}$, $s_i^{(2)}$, $s_i^{(3)}$, and $s_i^{(6)}$.

   There were mixed results from these selections as some stability measures were highly correlated, while being uncorrelated with others.

2. Variety recommendations based on imputed data for geographically similar environments were found in Section 9.4.

3. Variety recommendations based on imputed data for environments within the same mega-environment (found in Chapter 7) were found in Section 9.5.

These methods were used, as the graphical approaches investigated in Section 9.2 were uninformative for making variety selections because too many principal component axes were significant.

The various methods could be used separately or in combination to give variety recommendations for new trials. They have been provided for illustrative purposes, and are likely to be improved upon when full covariate data is available for both genotypes and environments. The power of the recommendation process will be greatly enhanced when the fully imputed data is combined with a web interface currently under development.

The investigation presented in Section 9.6 showed the problem of using only imputations based on first stage clustering, rather than the full two-stage imputation process. Some genotypes were not given imputed values in any environment of the subset under consideration, and could not therefore, be given an estimated performance to allow their possible selection for the new trial. Consequently, this method will not be made available through the planned web interface.

The next chapter provides advice for trials programme organizers based on findings gained from two sources. First from the analysis and working with the data arising from the Onion Trial Programme, and second from the impact two-stage imputation can have on the future development of a trials programme.

Chapter 10

Advice for future trials programme designers

10.1 Introduction

The scale of the Onion Trials Programme is seemingly unparalleled. Its uniqueness stems from the combination of the international focus and the sparsity of the data. Attempts to provide meaningful information to the organizers, collaborators, and interested growers based on such sparse data have uncovered many difficulties during the project's lifetime. Lessons from experience, made over the data collection phase, were presented by Dr Currah at the Third International Symposium on Edible Allaceae, Currah *et al.* (2001). Currah *et al.* (2001) summarized the state of competition between organizers and collaborators of the Onion Trials Programme, stating: "In a voluntary effort such as ours, there is clearly a trade-off between the ideal recommendations for an integrated trial system (reducing the 'sparsity' in subsequent combined analyses) and the needs of researchers who run the trials primarily to find varieties which they can recommend to farmers in their particular locality.". This chapter builds on those findings, and presents some new developments since that time which may assist trials programme organizers in the future. The main points of interest are:

1. Consistency of results from individual trials.
2. Improvement in the linkage between individual trials in the programme to build a cohesive trials programme.
3. Minimization of wasted resources.
4. Selection of varieties and locations to include in future years of the trials programme.

These points need to be worked through to facilitate the application of statistical models to the data that are collected. In turn the statistical modelling will assist in the agronomic development of the trials programme over time.

Development of a modelling approach for sparse G×E data provides an opportunity for trials programme organizers to save resources, or expand their horizons given the same resources. Judicious use of available resources will assist the statistical modelling process, as data collection is expensive and its wastage due to excessive sparsity is undesirable.

Practical experience gained from theoretical development of the two-stage imputation method has prompted presentation of guidelines that may assist trials programme organizers in the future. Trials programme organizers play a crucial role in ensuring that the results of all trials can be brought together for analysis. As a first step, the criteria used to develop Onion Data I and II is revisited.

Recall that in Section 3.6 the minimum representation of genotypes and environments was considered. Results from many trials were discarded when they did not meet the standards chosen at that time. If a similar standard is to be employed in future, an obvious precaution is to ensure that each environment added to a trials programme meets, or preferably exceeds, this standard. Other issues relevant to the selection of genotypes and environments are now considered in the remaining sections of this chapter which focus on:

1. Selection of environments to include in a trials programme.
2. Covariate information to be collected.
3. Experimental design of individual trials.
4. Modelling of planting density.
5. Selection of genotypes for new trials, notably including the use of check varieties.
6. Enhancing the connectedness of the G×E matrix.

This last section discusses the impact imputation of unobserved G×E combinations can have on the ability to use resources efficiently.

10.2 Selection of environments

The selection of environments seems at first to be a simple activity. Recall that the definition of environment used throughout this investigation was a time-location combination. The selection of environments is therefore a two-part problem, which adds to the complexity of the trials organization process. Introduction of new environments creates both advantages and disadvantages that need to be considered, so this section is aimed more at provoking thought, than at providing strict rules for application.

The environments used in a trials programme should reflect the population of climatic, geographic, and edaphic conditions to which results will be applied. The aim is to form

a broad base from which to make predictions, but differs from standard ideas of sampling in the following way.

A researcher would normally include environments with different attributes in the same proportions as are likely to occur in the target population of environments. When results from a large experiment, such as the Onion Trials Programme, are available predictions can be based on a subset of the environments tested using analyses similar to those presented in Chapter 9. It follows that each type of environment needs to be given adequate exposure in the programme so that predictions are available for a new environment of any type. The analyses presented in Chapter 7 however, showed that the right number of mega-environments can be difficult to determine, let alone that there is adequate representation within each mega-environment.

On the other hand, if the results from the entire trials programme are to be used in a standard G×E analysis, there is a risk that factors associated with over-represented environment types will incorrectly emerge as important. In these circumstances, attempts should be made to have the set of environments included in the programme more closely represent the target population. This will assist researchers who wish to know why particular varieties suit particular environments.

In the early years of a trials programme the goal of ensuring that the set of environments matches the target population, and that of ensuring that each mega-environment is sufficiently represented, may be in competition. Over time, however, it should be possible to include more environments that lie within the bounds of under-represented mega-environments. At the same time, the range of environments is likely to develop until the set of environments does in fact represent the target population. Judicious selection of environments should ensure that agronomically significant factors will also be statistically significant, and show the correct hierarchy of agronomic importance. Take a simple example with ten environments, nine of which differ in their maximum temperatures, while the tenth stands alone due to its heavy rainfall. The performance of genotypes and the interactions may be attributed to the maximum temperatures rather than the rainfall, which could in fact play a much greater agronomic role in the performance of the genotypes.

Selection of environments poses a much more interesting problem as the trials programme progresses over seasons, especially when the decision to use a certain location is beyond the control of organizers. In a particular season, availability of resources at a location is determined by the collaborative researcher. Programme organizers may find themselves with offers of assistance from collaborators at a limited range of environments. Meeting the goal of ensuring adequate representation of certain mega-environmental conditions, while matching the target population of environments, will therefore become more difficult.

The co-operation of collaborators can ameliorate this situation, and contribute to their own knowledge by trying something different at the organizers' request. As an example,

consider a collaborator who wishes to run a trial this season, but would use a location that is well represented in the programme thus far. This collaborator could be encouraged to plant seed earlier than would normally have been considered, thus altering the environment in terms of its climatic conditions.

Local varieties may respond better than new varieties because they suit the agronomic practices of a location. Changing the nature of the environment may show the efficacy of local practices as well as the suitability of local varieties, but could also show that varieties new to the collaborator suit conditions that are not commonly experienced at that location.

Continual investigation of environmental inter-relationships is clearly important for international trials programmes with incomplete data, as it is for smaller regional programmes. The need to use resources wisely is paramount for both scenarios. In regional contexts, such as presented by Lin and Morrison (1992), the aim is to avoid duplication of results from one environment to another. In an international trials programme, where only a small portion of varieties will be tested at any given location, there is an aim to make the best predictions with comparatively scant resources.

It is possible to counter the problems caused by a set of environments not matching its target population. The analysis could be weighted to favour results from under-represented mega-environments, but this would involve subjective interference in the analysis. In situations where the data set is incomplete, this weighting cannot be applied without extremely careful consideration of the impact that tested G×E combinations will have on results. Re-weighting environments would be counter-intuitive to the need to standardize within environments discussed in Section 5.2. The specific aim of that transformation was to ensure that all environments contributed evenly to clustering, and therefore, to the imputation process (Section 6.2).

If a weighted analysis is to come from incomplete data, genotypes need to be balanced across mega-environments, so that individual environments do not unduly influence results. This in itself adds complexity to the selection of genotypes for new environments, as organizers cannot be completely sure that this new environment will form part of the expected mega-environment. On the other hand, when the data set is made complete by imputation, the weighting can be introduced when selections are made, as in Chapter 9 or via factor-based models.

This 'band-aid' type approach risks becoming cumbersome and disputable. A more logical approach would be to ensure that the right combination of environments is added to the programme each year. The breadth and depth of environment diversity (mega-environments) will be determined as the total set of environments changes from season to season, so decisions may actually have negligible impact. As long as collaborators can meet the restrictions placed on them by organizers with respect to the number of varieties tested and data collection requirements, their contribution can be welcomed with open

arms.

These minimum standards need to be chosen for the entire trials programme, and adhered to by all collaborators. The slightest diversion from these standards impacts on the entire analysis. Recall from Chapter 3 that many regression models were found using subsets of the data, as complete covariate information was unavailable for some environments. Section 10.3 covers possible covariate information that can be collected on environments and genotypes. A collaborator's contribution will be less valuable if they cannot meet the minimum standards. For example, their environment will be left out of analyses similar to those presented in Section 7.4, although their yield results could be incorporated into the imputation process, so their effort is not wasted.

Four criteria were presented by Williams *et al.* (1992) to assist the selection of environments for future use. These were:

1. The mean yield of the environment should be at an acceptable level.

2. The year-to-year variability for the environment should be low.

3. It should be easy to discriminate between genotypes' performance, especially the best performing ones.

4. The environment should represent some environment subset of the total population of environments, including the correlation of test results with actual performances.

This set of criteria may be useful in a regional programme, but the Onion Trials Programme was not focused on the better performing environments. Rather, the focus was towards the response genotypes had when grown in conditions found in tropical and subtropical environments. The need to find environments that have good stability from year to year is also not of great concern as there was benefit in covering a wide range of environmental conditions. The use of location-year combinations as environments meant that the differences between years has enriched the analyses. Thus the first and second criteria are less relevant in an international setting. The second criterion could be replaced with the need to ensure that observed yields would result if the same (or very similar) growing conditions were to be found elsewhere.

The third and fourth criteria however, are definitely important in an international setting. Individual trials do need to be planned so that adequate distinction between genotype performances can be found. The ability to have test results mirror off-trial results is a concern in all experimentation. The last two criteria and the suggested replacement for the second criterion can be catered for by effective use of experimental design methodology to be discussed in Section 10.4.

Aspect	Variables
Phenology	Final leaf number, and times of flower initiation, flowering, and maturity.
Crop canopy development	Plant height, tiller number, and maximum ground cover.
Pest resistance	Severity of infection by various diseases, of attack by insects, and of infestations by weeds.
Yield components	All morphological components of yield.

Table 10.1: Covariate information for genotypes that could be collected when trials are planned. Adapted from Yan and Hunt (1998).

10.3 Covariate information

As indicated in the previous section, there is a need for trials programme organizers to determine what covariate information is required from every collaborator. Some general suggestions, and some specific to onions, are offered in this section.

Successful application of factorial regression models, such as those presented in Section 2.5, rely on collection of sufficient covariate information to explain the variation among genotypes and environments. The principal data collected from the Onion Trials Programme included environmental covariates (e.g. latitude and altitude), as well as covariates that were specific to particular G×E combinations (growing periods for example). In some circumstances, researchers collect information on the particular covariates they are interested in understanding, while ignoring the impact of related (but unobserved) covariates.

There is a danger that variation can be attributed to a source that is related to, rather than actually being, the factor that influences the yield of crops. Trials programme organizers may avoid such pitfalls and allow fitting of factorial regression models, by establishing the covariate information that will be collected for each genotype and from each environment used in the programme. Clustering of environments to form mega-environments (Section 7.4) could have been improved if better covariate information had been available.

Yan and Hunt (1998) provided lists of covariates that should be collected for genotypes and environments where possible. These have been reproduced in Tables 10.1 and 10.2 respectively. These lists are in many ways quite comprehensive, but there has been concern that sociological and socio-economic factors have been overlooked in favour of physical aspects of environments (Wade *et al.*, 1996).

Table 10.2 lists 'irrigation' as a 'management' factor. In respect to the Onion Trials Programme, irrigation would need to be measured in several ways. The actual type of irrigation used, as well as the quantity of water, is important because some environments use 'furrow' irrigation which in turn has an impact on the layout of the trial. Irrigation is an example of a cultural practice that needs to be recorded so that genotypes most suited

Aspect	Variables
Weather	Temperature, day-lights, solar radiation, precipitation, wind speed.
Soil properties	pH, organic matter content, texture, contents of nitrogen, phosphorus, potassium, and micro-nutrient.
Management	Planting date, density, fertilization, irrigation, herbicides, pesticides, tillage, and other interventions.
Pests	Maximum severity of infection by various diseases, insects, and of infestation with weeds (using susceptible check cultivars).

Table 10.2: Covariate information for environments that could be collected when trials are planned. Adapted from Yan and Hunt (1998).

Aspect	Variables
	Genotype covariate information
Appearance	Particular factors of interest relevant to onions include colour, size, and shape.
Non-visual	Taste and smell (pungency for example), percentage dry matter.
	Environment covariate information
Experimental design	Details of the design used, and the mean squared error that resulted in any design factors having being removed.
Irrigation	Type of irrigation and the quantity administered.
Local preferences	If the crop is to be targeted at local consumption, the preferences of consumers must be known. This is meant to include such factors as pungency, colour, and size of the harvested crop.
Plot layout	The size of the plot, whether guard rows were used, etc.

Table 10.3: Covariate information for genotypes and environments that could be collected when trials are planned for addition to the Onion Trials Programme. These build on those given by Yan and Hunt (1998) which appeared in Tables 10.1 and 10.2. Discussion of some of these factors can be found in Currah *et al.* (2001).

to this kind of plot layout can be identified.

Other experimental design factors also need to be included in the data recorded by collaborators. The actual design used, and any resulting replicate error could be useful information. A list of additional genotypic and environmental factors that could have been collected by all collaborators for the benefit of the Onion Trials Programme is presented in Table 10.3.

The environments whose results were used for developing recommendations for a new trial in Section 9.4 were limited to latitude and altitude. The ability to select certain types of environments from the data base will be enhanced by inclusion of many of the covariates listed in Tables 10.2 and 10.3.

Brewster (1990) noted that there was potential for different maturity dates for onions

within sites caused by different climatic and cultural practice. This paper showed that seasonal differences caused supposedly by differences in temperature led to difference in maturity date. He used 'day-degrees' greater than five degrees Celsius, although Brewster (1997) improved the formula to accumulate 'day-degrees' between six and twenty degrees Celsius. Plant density is also known to affect the maturity date (Brewster, 1990) and the individual bulb weight of onions (Brewster, 1994). The relationship between planting density and bulb weight is covered in more detail in Section 10.5, but the issue of maturity date differences on the growing periods of varieties and their interaction with environments remains unclear.

The covariates recorded with the data arising from the Onion Trials Programme were for the most part collected for G×E combinations. Such covariates are easily included in a linear model if they are consistently collected for all G×E combinations. The growing periods of G×E combinations were available for many G×E combinations, but the temperature and photoperiod data for these combinations depended on the length of the growing period.

10.4 Experimental design for individual trials

This section aims to bring out various discussion points that must be considered by trials programme organizers, and employed by collaborators to ensure consistency of results. It does not include the design of the trials programme itself which is left to subsequent sections. The points to be considered include:

1. Options for reducing the influence of within-environment variation.
2. The use of replicates.
3. The layout of experimental units.

These points are encompassed by the term 'experimental design'. Effective design can overcome difficulties such as lost units or variation caused by within-site trends.

Choice of a particular design for application at all sites suggests that all data will be analysed together, perhaps using multi-level models. Another approach, used in the Onion Trials Data, is to use a single result for each G×E combination. Each method has its advantages and disadvantages, as discussed in this section.

The main advantage of the 'one design for all sites' approach is that the statistician can bring everything together and fit a model as complex as the design allows. The modelling is then a product of the original design. The disadvantage is that achieving this consistency across all sites and years may prove to be a task that is beyond programme organizers. Collaborators may need extremely elaborate instructions, which in turn need to be cognizant of the abilities of all collaborators. It is unrealistic to assume that all collaborators in the future will have this level of competence given that it is not known

who will be contributing from one year to the next. Pre-determination of the minimum requirements would impose restrictions that may impact on the willingness of collaborators to participate. It may also mean that organizers are then making a commitment to educate any collaborators who are willing to assist, but do not currently have the skills to contribute in the desired fashion. This approach is used by many G×E researchers worldwide, but in general these programmes are less diverse and geographically less spread than the sites used in the Onion Trials Programme.

This pragmatic approach would then encourage collaborators to choose the design that best suits their local conditions and abilities. There would then be a need for organizers to establish constraints so that data returned to them is consistent. As discussed in the previous section organizers need to decide what data is to be gathered by all collaborators. Allowing collaborators to decide how they achieve the desired outcome may improve participation and consistency.

The major drawback to this flexible, and potentially chaotic, approach is that there will be an impact on the analysis. All results would need to be amalgamated after individual trial results have been determined. That is, any experimental design factors used in an individual trial would be removed from the data that is brought into the main G×E data set. Data analysis is then performed on two levels, with the possibility that the collaborator may perform the local analysis. This is effectively the method that has been applied in the Onion Trials Programme.

The first option where every collaborator operates under the same constraints and leaves analysis to the organizers, is ideal. It is, however, an option that requires the greatest amount of management and communication. The second option leaves much of the decision-making to the collaborators and relies on finding pragmatic solutions after data has been collated. It also allows flexibility over time, as the demands of organizers and collaborators change. This is not to say that the questions of experimental design can be ignored. The quality of results must be ensured to minimize wastage of effort. The particular designs which could be used at each location are now considered.

Reducing the influence of within-environment variation

The field of experimental design has focused on removing the influence that non-treatment effects have on results. A complete review of historical developments from randomized complete block designs onwards, is unnecessary. Some noteworthy contributions are considered, however, that may be directly applicable to trials like those that comprise the Onion Trials Programme.

Differences in sites will need to be considered when designs for trials are chosen. Whichever design is employed, it must be capable of handling missing data. Randomized complete block (RCB) designs can, for example, find a mean value for each variety even if some units are lost to unforeseen circumstances. For this reason, use of RCB de-

signs seems a logical minimum standard for organizers to set collaborators. If blocking is needed in one direction, however, it is likely that blocking in two directions needs to be considered. Latin square and lattice designs do this, but in comparatively recent times spatial and neighbourhood approaches have become prominent.

There is a risk in overcomplicating a trial's design. John and Williams (1995) note "Although the results of neighbour analyses appear impressive relative to incomplete block analyses, the extra complexity of the models and estimation procedures mean that careful screening of results is usually required which perhaps limits the general use of these methods." Kempton *et al.* (1994) compare various one and two dimensional experimental designs for their effectiveness. They found that a well designed row and column experiment is almost as effective as the two dimensional nearest neighbour method. As collaborators provide their time and physical resources voluntarily, use of unnecessarily complex designs would appear inadvisable.

A collaborator knows their site better than the statistician or the programme organizer. Each collaborator should choose the design that best suits their site and skill level. Programme organizers must then trust that the collaborator is planning their trials in such a way that minimizes the impact of all within-site factors including fertility trends etc. occurring at their site.

Variety mean yields can be found regardless of the particular design applied at each location. Williams *et al.* (1992) note "Once the analyses at each location are completed, the estimated cultivar means can then be combined into a cultivar (C)× location (L) (or genotype×environment) table." They note that combining the results from trials held in the same year is likely to lead to a complete G×E table as the set of test varieties is commonly held constant within years and allowed to change over years. This is not the case for the Onion Trials Programme, because there was not a simple series of G×E tables to review as occurs in smaller regionally focused programmes. Keeping issues arising in specific trials separate from issues that arise when these trials are combined will simplify management and analysis of programme results.

The use of replicates

Increasing the number of replicates for each G×E combination limits the number of combinations that can be tested. Gauch and Zobel (1996) present tables to assist with decisions over the number of genotypes and replicates to test at a given site. This is particularly important as the trade-off between the number of genotypes and the number of replicates is made more difficult when the incomplete nature of the trials programme is also considered. There will be further pressure to get more results from each environment, but the results gained from all trials should be of the same standard. Decisions over how many replicates and genotypes to be used is dependent on the amount of physical resources a local collaborator is able to dedicate to the trial.

Organizers should aim to get the same quality of results from each trial to ensure consistency across the entire programme. Unfortunately this can be measured two ways; all trials could aim to have the same error variance as measured by the replicate error, or to have a common target ratio of between treatment to within treatment variance. Gauch and Zobel (1996) promote this ratio in terms of signal to noise. A collaborator should be able to determine what magnitude difference in yields is considered significant in their local environment based on past experience. The tables presented in Gauch and Zobel (1996) can then be used to determine the number of replicates that are required.

In either case, there is no guarantee that data supplied to organizers from all trials will have homogeneous variance. If two-stage imputation is to be applied as described in Chapter 6, these trial results will be standardized so that each trial contributes equally to distance measures and therefore imputation results. Two-stage imputation takes no account of the significance of results at particular locations as would be determined by the standard F test based on between variety variances and within variety variances.

The design and analysis stages of an experiment are not completely separable. The use of effective analyses can, according to Gauch and Zobel (1996), substitute for replicates. Gauch (1992) presents the relative efficiency of the AMMI model for example, but this assertion is based on the regional focus concentrated on by the author. Gauch and Zobel (1996) claim that an AMMI analysis on a data set with three replicates per G×E combination is generally as efficient as simple averaging of a data set with six replicates per G×E combination. It would be prudent to not rely on unpredictable efficiency of particular modelling strategies; rather advantage should be taken of any such gains made, counting them as serendipitous.

The layout of experimental units

Agronomic issues pertinent to collaborators include:

1. Minimum plot sizes within trials should be recommended. For trials of onions these should be not less than two square metres (Currah *et al.*, 2001).

2. Use of guard rows is advisable in many circumstances such as wide planting plots, but when furrow irrigation is used as the local practice this should be followed, making guard rows impossible (Currah *et al.*, 2001).

3. Plant spacing should be kept stable but may depend on local bulb size preference (Currah *et al.*, 2001). Variation of planting density may assist agronomic comparisons. Currah *et al.* (2001) identified this as most suitable for trials when the best varieties have already been chosen. Planting density is considered further in Section 10.5.

Given these comments, a universal standard for all trials in a programme may be inappropriate; a set of guidelines, including attainable minimum standards for all collaborators

and examples of best practice, may be a pragmatic solution.

10.5 Modelling of planting density

It is well-established that planting density has an effect on onion bulb weights (Bleasdale, 1966; Frappell, 1973; Brewster and Salter, 1980). In the Onion Trials Programme, choice of planting density was left to collaborators, and it has therefore been impossible to determine if planting densities were competitive, and what effect any competition may have had on each genotype, environment, and G×E combination. Removal of the effects of differing planting density from the results was not possible, because density effects were confounded with environment effects, and were therefore subsumed into the environment factor. This section shows the results from the investigation into methodology that would have been applied if planting densities within environments had varied more than planting densities across environments.

The theory that individual onion bulb weight will decrease as the density of plants increases has led to the use of a reciprocal yield model of the form

$$\frac{1}{w} = A\rho + B \tag{10.1}$$

where A and B are the parameters found when the reciprocals of onion buib weights w are regressed on the planting density ρ, (Nichols, 1970). The asymptote of the reciprocal function fitted for the per unit weight has the following properties:

1. When ρ increases to its maximum, w approaches $1/A$, where A is a measure of the yield potential of the environment.

2. As ρ tends to 0, w approaches $1/B$, where the reciprocal of B is a measure of the plant's genetic potential.

Nichols (1970) presented evidence that this model should only be used when competition exists between plants for resources, and introduced the yield density model that allows for both competitive and non-competitive planting densities. The model in (10.1) is modified to give

$$\begin{aligned} \frac{1}{w} &= AC + B, \quad \rho < C \\ &= A\rho + B, \quad \rho \geqslant C \end{aligned} \tag{10.2}$$

This model is fitted by varying the level C at which competition is deemed to start. All points where $\rho < C$ are fitted as if they were grown at density C. The level C at which competition actually starts is therefore determined to be the point where the Residual Sum of Squares (RSS) of the corresponding model is minimized.

The model presented in (10.2) is a simple piecewise linear function with one break point. The left part of this line is horizontal, while the right has an upward slope.

Boyd *et al.* (1976) used a complete search method to find the break point of two intersecting regression lines. Their approach was to order the independent variable so that $x_1 \leqslant x_2 \leqslant \ldots \leqslant x_{n-1} \leqslant x_n$, then split the data into two parts with $y_1, \ldots, y_i$ and $y_{i+1}, \ldots, y_n$ for $i = 2, 3, \ldots, n-2$. Regression lines were then fitted to each, so that two models

$$\begin{aligned} y &= b_0 + b_1 x \\ y' &= b'_0 + b'_1 x \end{aligned} \tag{10.3}$$

were found. The points at which these pairs of lines intersected were considered. All models that had the intersection of regression lines outside the interval between the two subsets of the data were discarded. Then the best of the remaining models was that which minimized the residual sum of squares in total.

This approach would allow for any change in slope between the two subsets of the data, and will therefore be capable of finding the same solution to that of Nichols (1970), given that $b_1 = 0$. In this case i can take the values $1, 2, \ldots, (n-2), n$, as the left hand part of the function will be allowed to be fitted by a single point, but the right-hand part needs a second point to fit the slope. Setting $i = n$ allows the competition level C to be higher than the observed range of planting densities. The Boyd *et al.* (1976) method provides an alternative strategy for fitting the same model as Nichols (1970), and is likely to be computationally faster as it will fit fewer models in its search for the optimal value of C. The number of models fitted by the Boyd *et al.* (1976) model is the number of observations in the data. The Nichols (1970) model performs a search along the range of planting densities, which will be as precise as the user determines. It is not possible to specify how many models will be fitted using this approach as this number will be dependent on the optimization method employed.

A search of the S-PLUS mailing list archive found discussion of fitting what was referred to as the 'hockey stick' model. Although solutions offered are effective for a piecewise linear function of the form in (10.3), they are over-complicated for fitting the piecewise yield-density model. If, however, the Nichols (1970) model in (10.2) is preferred, the optimization capability of S-PLUS could be used to fit the hockey stick model.

10.6 Selection of genotypes to include in each environment

The selection of genotypes for new trials should consider a number of factors, including:

1. Admission of new varieties to the trials programme.

2. Use of check cultivars.

3. Use of local varieties.

4. G×E combinations that fail.

5. Selection of genotypes for testing in a new environment.

Each of these concerns is addressed in this section, and is, where appropriate, linked to the findings of the previous chapter. The effects that these decisions will have on resulting G×E matrices will be discussed in Section 10.7, and will include ways to improve these matrices to get the most of past and future data collection efforts.

Admission of new varieties to the trials programme

A concern when adding a new variety to those being tested is whether it will be available for commercial use in future seasons. This is reliant on several factors.

First, the variety should be the same from year to year in terms of its genetic make-up, meaning that the genetic drift of the variety should be small. Seed purchased under the same varietal name must be the same from one year to the next if imputations and predictions are to have any future relevance. There would be little benefit in recommending a variety based on its performance in one season if it cannot be relied upon to perform as well in the same environmental conditions in another season. This is not the same as stability, as discussed in Chapter 2, which assumes that any change in performance is due to random events and differences between the environments, rather than any difference between the genotype in different seasons.

Second, the variety must be physically available. It should go without saying that there is little benefit growing a variety knowing that it is about to become obsolete. The resources of collaborators are limited, so each variety planted must benefit an ongoing trials programme.

Third, it is important that enough seed of a variety is available for testing in a sufficient number of trials to allow its inclusion in the G×E matrix that will be analysed. The obvious exceptions to this rule are local varieties that collaborators may include at their own discretion to test their particular theories, or to use as their own check cultivars.

Ideally a genotype added to the trials programme should be grown in at least one environment of each mega-environment. Given that the number of trials in which most genotypes have been tested thus far is smaller than the number of mega-environments established in Chapter 7, this objective is unlikely to be achieved. Obviously, risks must be taken to further develop the trials programme. Continually adding genotypes to the testing pool year after year cannot be sustained, because this would only increase the sparsity of the overall G×E matrix. Severe limitations may in fact be necessary if the connectivity of the G×E matrix is to be ensured.

Use of check cultivars

Check cultivars have been used extensively in G×E analyses as they provide a foundation on which to build (Lin and Binns, 1985; Eskridge *et al.*, 1993). In some circumstances they are varieties that are chosen because they have good predictability of performance, possibly in terms of low G×E interaction, as shown in Section 9.3. Another definition for check cultivars would be varieties that are in common circulation and well known, but not necessarily in terms of performance in every environment.

A set of check cultivars can be used to provide an understanding of the potential or variability of an environment. They should help cluster similar environments to one another when the G×E matrix is incomplete, or help define the nature of environments when insufficient covariate information is available.

Within each trial, there is likely to be some amount of physical variation among the plots allocated to varieties. Check cultivars can be used to gain an understanding of the land used to run the trial. For example, the check cultivars could be tested twice as frequently within each trial, allowing more varieties to be tested at each location, while providing a clear understanding of the spatial variation of the experimental site.

Eventually, trials programme organizers must decide how long a set of check cultivars is to be used, and whether they will all be used in every trial. There are reasons why it is useful to have all trials in a single season use the same set of check cultivars. Chief among these is that this helps with the connectivity of the G×E matrix, because all trials in a season are able to be linked to one another. This will therefore reduce the potential for a trial or set of trials to be left out of future analyses. The addition of data from more trials should provide the analysis with more information on inter-genotype relationships, and therefore, should also improve the accuracy of imputations.

The needs of collaborators to have varieties allocated that match local preferences cannot be overlooked. The set of check cultivars must, therefore, include onion varieties that are of differing styles, whether these be chosen by colour, taste, or some other preference. Currah *et al.* (2001) suggested using at least two cultivars of different types, for example 'Red Creole' and 'Texas Early Grano'. An inherent problem with using a single check cultivar was raised by Currah *et al.* (2001): The wishes of collaborators to grow onion varieties of one colour due to local market forces, limits the potential for this type of check cultivar selection.

Lin and Binns (1988a) advocate use of the maximum observed yield, in lieu of check cultivars, to estimate an environment's potential. The ability to provide a means of gauging other genotype performances is not the main purpose of check cultivars in the ongoing success of the Onion Trials Programme. It is more important that the connectivity of the resulting G×E matrices be maintained.

May and Kozub (1995) describe genotypes in either their second or third season of testing as check cultivars. They found that these varieties (chosen on the grounds of

their success in the first testing season) did not perform as well in subsequent testing. Given that the check variety recommendations in Section 9.3 have not yet been tested in a substantial number of trials, they are likely to be better understood after they have been used more frequently.

Lin and Binns (1985) noted that if any two check cultivars fall into the same group in clustering, or give the same interaction profiles according to the joint regression model, then there is an inefficiency induced. The later prominence of the AMMI modelling approach in G×E analyses would probably allow the Lin and Binns (1985) judgment to be expanded by removing the explicit reference to the joint regression model, but the point remains that there is little benefit in using two check cultivars that are expected to perform in a similar fashion. For example, the four check cultivars ('Dessex SS', 'Jenin HZ', 'Marix ZU', and 'Superex TK') found in Section 9.3 were to be found in four different first stage clusters for Onion Data II, but in only three Onion Data I first stage clusters (refer to Section 5.4). In that case the varieties 'Jenin HZ' and 'Marix ZU' were both included in genotype cluster 1, but have thus far been tested in only six common environments of both Onion Data I and II. When they have been tested together in more environments, it will be known if they are in fact as similar as the clustering of Onion Data I genotypes suggests, or are as dissimilar as indicated by the clustering of Onion Data II genotypes. If the former is found to be true, there would be good reason for not testing them together in any future trial.

The set of check cultivars should have good wide adaptability to the trials that are planned for the coming season. The exercise of Section 9.3 found the suggested check cultivars using all the trials from past seasons. If for some reason some of these trials are determined to have experienced extremely unusual, and therefore unlikely to be repeated, conditions they could be removed from that analysis. It is assumed that the set of trials that have been included represent the population of environments that need to be covered by the trials programme. If this is not the case some means of weighting results may be in order. No matter how the check cultivars are found, they may change after data from another season of trials has been included in analyses. There should, therefore, be an expectation that check variety recommendations are for the coming season only and will probably change from season to season.

Use of local varieties

Inclusion of local varieties in a trials programme will give little information about their international performance unless they are tested in a sufficient number of trials. They may, however, be beneficial to the collaborators' ongoing research interests. There have been situations throughout this investigation where a local variety may have been used to aid the analysis. In the imputation of Onion Data I and II, the observed minimum and maximum yields were used to 'trim' unrealistic imputed values. If the local variety yielded higher

than those genotypes included in the data being imputed, its yield could have been used in place of the maximum yield from within that data. The use of the adjusted superiority score $p_i^\star$ also used the environment maximum and could be enhanced by use of the overall observed maximum yield instead of the maximum of the data under consideration. The data from local varieties should also be used if the yield/density modelling of Section 10.5 is to be undertaken.

The major aim of any trials programme is the acquisition and transfer of knowledge. If a local cultivar is superior to others tested in a location, it may indicate the need for its further testing in a broader range of environments. It should then be considered for inclusion as a new variety as discussed earlier in this section.

G×E combinations that fail

There are two main reasons why knowledge that a crop will not bulb in a particular location should not be discarded. Whether this knowledge be gained from past experience or from observed data, such G×E combinations should be included in G×E analyses because:

1. Deleting these observations would break the missing at random condition.
2. Deletion of these observations contributes to the sparsity. These G×E combinations would then be imputed, possibly with values that are then known to be wildly inaccurate.

As observed with the data arising from the Onion Trials Programme, inclusion of zero yields allowed some trials to be included in Onion Data I or II. For example, addition of G×E combinations with zero yields allowed trial 'X04601' from Cameroon to be included in the Onion Data I set.

Differing views on the inclusion of lower yielding genotypes were found in the literature. Finlay and Wilkinson (1963) noted that "The practice of ignoring 'crop failures' will bias the selection towards types specifically adapted to high-yielding environments, and will pass over those with general adaptability." On the other hand, Gauch and Zobel (1996) present an argument for the non-inclusion of inferior genotypes, as this would make the chances of selecting superior genotypes more difficult. The more inferior genotypes that are added to a trial, the harder it will be to select a superior one, and the more superior genotypes added, the easier it will become to select a superior one. This does not necessarily mean that the best genotype will be selected though, and their judgement was based on a single environment.

The argument of Gauch and Zobel (1996) is justifiable when considering genotype selection in a regional context, but the aims of an international trials programme are to select on the benefits of any advantageous G×E interaction. There is, therefore, a need to ensure that the best understanding of G×E interaction is obtained for groups of genotypes in all different subsets of environments. This means finding where genotypes

are most suited as well as where they are least suited, and more particularly where their selection would lead to wastage of resources.

As many G×E combinations have been estimated via the imputation process, selections for a new trial are not guaranteed to be successful, and there is therefore an inherent risk of choosing to grow inferior genotypes that must be borne. Strategies for the removal of genotypes that are currently designated as inferior from future testing will need to be developed with great care. For example, it may be decided that a genotype that has performed lower than the mean performance in every environment should be omitted from future testing. The problem with this is that a genotype may not have been tested in the environments to which it is most suited. Knight (1970) raised this concern of confounding low performing genotypes with suboptimal growing conditions for genotypes.

It is in the nature of a collaborative trials programme covering a wide range of climatic, and edaphic conditions that varieties will not always perform well outside the original environment where they were developed.

The final reason for not discarding varieties that have not yet shown high yields, is that two-stage imputation does not restrict imputed yields of genotypes to be below average where they were not observed just because they were below average when they were observed. It is the inter-genotype relationships that will determine the imputed values, and if a genotype has not yet been tested in environments to which it is suited, while a similar genotype has been tested in those environments, the imputed yields will reflect the desirability of testing the first genotype further.

Selection of genotypes for testing in a new environment

In Chapter 9, genotypes were selected for a new environment using the imputed data found in Section 6.5. An implicit assumption of two-stage imputation is that data are missing at random, and that there is enough information about particular combinations of explanatory variables to allow the estimation of missing G×E combinations. More specifically, each set of similarly performing genotypes (from first stage clustering) should be represented in each mega-environment, so that imputations are based on the most desirable relationships, rather than relying on more tenuous relationships over groups of genotypes.

In Section 9.6, use of imputations based solely on inter-genotype relationships from first stage clustering was shown to be unable to provide imputations for all genotype groups in all mega-environments. Over time, organizers should aim to have tested each genotype in at least one environment from each mega-environment, or at the very least, one genotype from each first stage cluster in every mega-environment. This would then allow the use of the partial imputation process to give estimated performances of missing G×E combinations, and therefore, allow recommendations to be based on really (rather than nearly) similar G×E interaction profiles.

Setting this concern aside, however, there are other concerns that need to be addressed. First, many genotypes were given an imputed yield equal to the maximum observed yield of an environment. In the trial 'X04601' from Cameroon in 1994, 43 genotypes were given the same imputed value. It will be difficult to obtain distinct imputed values for these genotypes until they are grown in enough future trials to provide a better means of determining their differences. There is no guarantee though that the imputations for this environment can ever have few imputed values at the maximum observed yield, because the imputations are dependent on the varieties that were tested there in the first place. If the varieties tested at 'X04601' were varieties that performed at a lower level than most, it will be impossible to obtain a distinct imputed value for every genotype. The use of averaging across environments in the same mega-environment reduces the chance that a large number of genotypes will receive the same average relative performance.

There are a number of dangers when selecting the varieties that are likely to perform well in an environment. First, if no weak genotypes are recommended, it will be difficult to gauge the value of the tested varieties. A second, and more important danger is that of having resulting data that is not missing at random. If it is known that the tested variety combinations are all for high yielding G×E combinations, the missing at random condition will be broken. This is better explained with a long term view. If varieties are grown in only environments in which they are likely to succeed, there will be a confounding of the specific adaptation (beneficial G×E interaction) and the presence/absence of data. Similarity of interaction and similarity of tested environments will be impossible to separate.

The inclusion of check cultivars may in part ameliorate this concern, but steps need to be taken to ensure that the missing at random condition is not broken in the long term. In the short term, imputations will be imperfect and not all recommendations will in fact lead to successful testing of high performing G×E combinations.

Combining all the findings of this section does not necessarily lead to a suitable selection of varieties to recommend for testing in a new trial. The ability to link the data that will be gained from a new trial needs to be given priority to ensure that the new data will be included in the analyses that will occur at the end of the next season. These concerns are covered in the next section.

10.7 Enhancing the connectedness of the G×E matrix

In Chapter 9, genotypes were found that were expected to succeed in a new environment, using the imputed data arising from the Onion Trials Programme. Successful clustering of either genotypes or environments, and therefore imputation of missing yields, relied on the ability of available data to represent the differences within the set being clustered. If a genotype has common environments with few other genotypes, it is less likely that it will cluster with the genotypes it would have if data were complete. It must therefore be

a goal of trial programme organizers to ensure that genotypes (environments) are related to a sufficient number of other genotypes (environments). It is known that if all genotypes have sufficient linkage to one another, the same will be true for environments (John and Williams, 1995).

A regionally focused programme of variety trials progresses from season to season by comparing new varieties with the most successful ones from the past. The nature of the Onion Trials Programme's development since its initiation meant that this season to season focus was not possible. A regionally focused programme will almost certainly have sufficient connectedness of both environments and genotypes, but a trials programme similar to the Onion Trials Programme may not necessarily have connectedness of the same quality.

A first step would be to ensure that future trials grow as many varieties as possible, both to test the hypothesized good performers and to test varieties that will improve the connectedness of the G×E matrices. This section shows how G×E matrices can be improved by judicious genotype selections for new trials, both to prevent them from being discarded from future analyses, and to allow more trials, and therefore genotypes, from past trials to be included in the analyses.

John and Williams (1995) used graph theory to show the level of connectedness for incomplete block designs. A set of vertices, representing treatments, had edges drawn between them if they were in a single block together. This can easily be done for an incomplete G×E matrix using genotypes as treatments, and environments as the blocks. An incomplete block design relies on the notion that there is no interaction between treatments and blocks to allow the treatment effects to be resolved. The probable existence of G×E interaction means that using a single link between pairs of genotypes will not allow estimation of the similarity of their G×E profiles, and will therefore, overstate the connectedness of the G×E matrix.

Throughout this investigation, inter-genotype distance measures were deemed to have been accurately recorded if they were calculated using data from at least four environments. The reliability of these distance measures would have been enhanced if this threshold could have been increased.

Figures 10.1 and 10.2 graphically display the inter-connectedness of the genotypes and environments of Onion Data II. They were created using an S-PLUS program which generated the necessary LaTeXcommands. Genotypes and environments are numbered according to the codes given in Appendix A. In these figures, the number of common environments (genotypes) required to have an edge drawn between a pair of genotypes (environments) has been set to seven for illustrative purposes. Using this criterion, there were eleven genotypes and seventeen environments within Onion Data II that could not be linked to any other genotype or environment respectively. It is this problem that forced the choice of q in Section 4.6, when estimates of unobserved distances were made, and

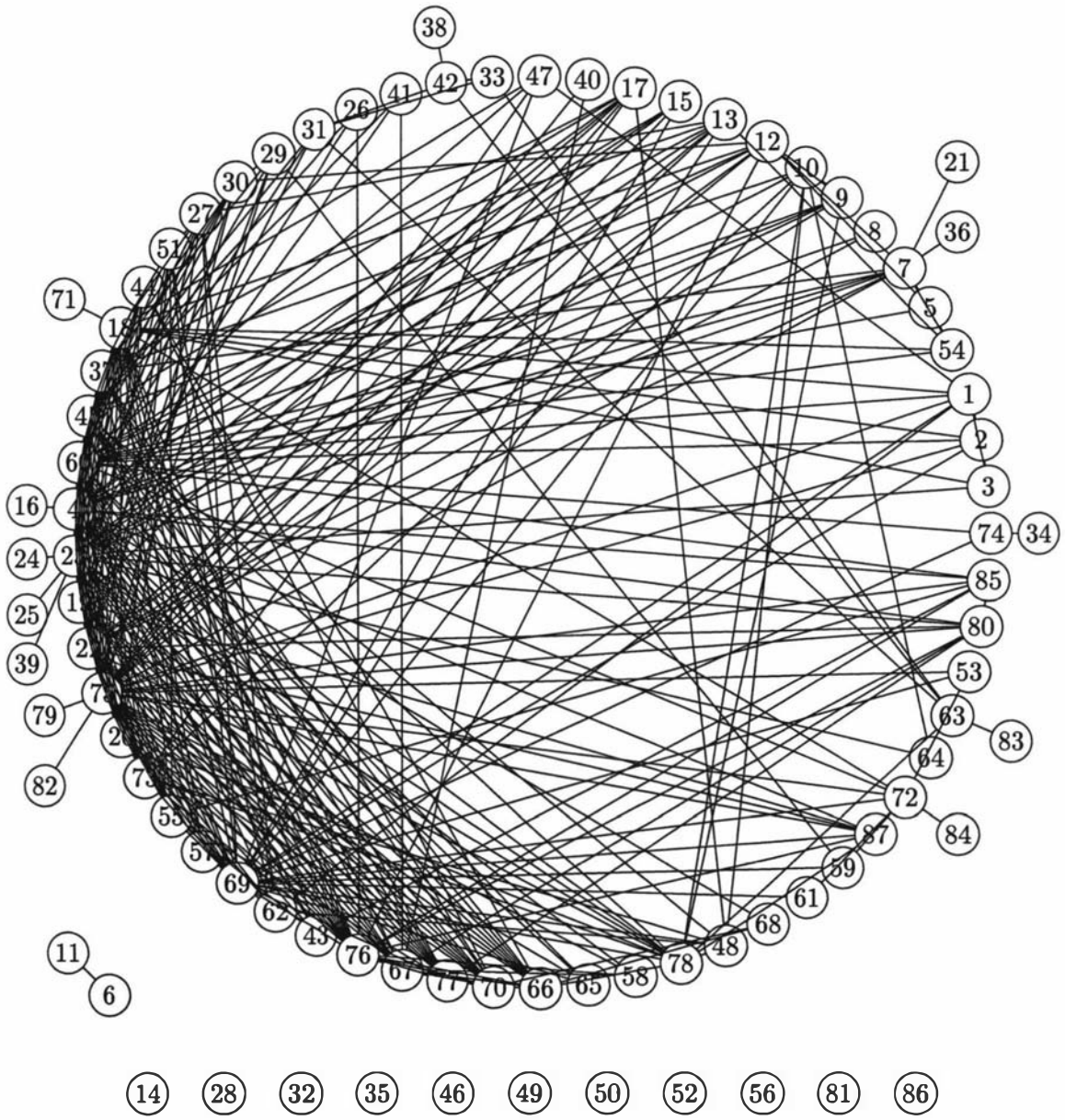


Figure 10.1: Diagram representing the inter-genotype connectedness of Onion Data II. The 87 genotypes are plotted as numbered nodes, while edges indicate that a pair of genotypes were grown together in at least seven of the 98 environments. As an example, genotype 74 (at the right of the figure) was tested in at least seven environments with three other genotypes, numbered 34, 60, and 69.

distances based on a small number of common environments were adjusted. Selecting values of q greater than four would have forced removal of some genotypes from Onion Data I and II, or a data splitting (seen by the connected pair of genotypes disconnected from the majority of genotypes in Figure 10.1).

Altering the threshold used to draw an edge in either Figure 10.1 or 10.2 would provide a series of diagrams that could be used to show the strength of inter-genotype (inter-environment) relationships. The same effect could be achieved by using a colouring of

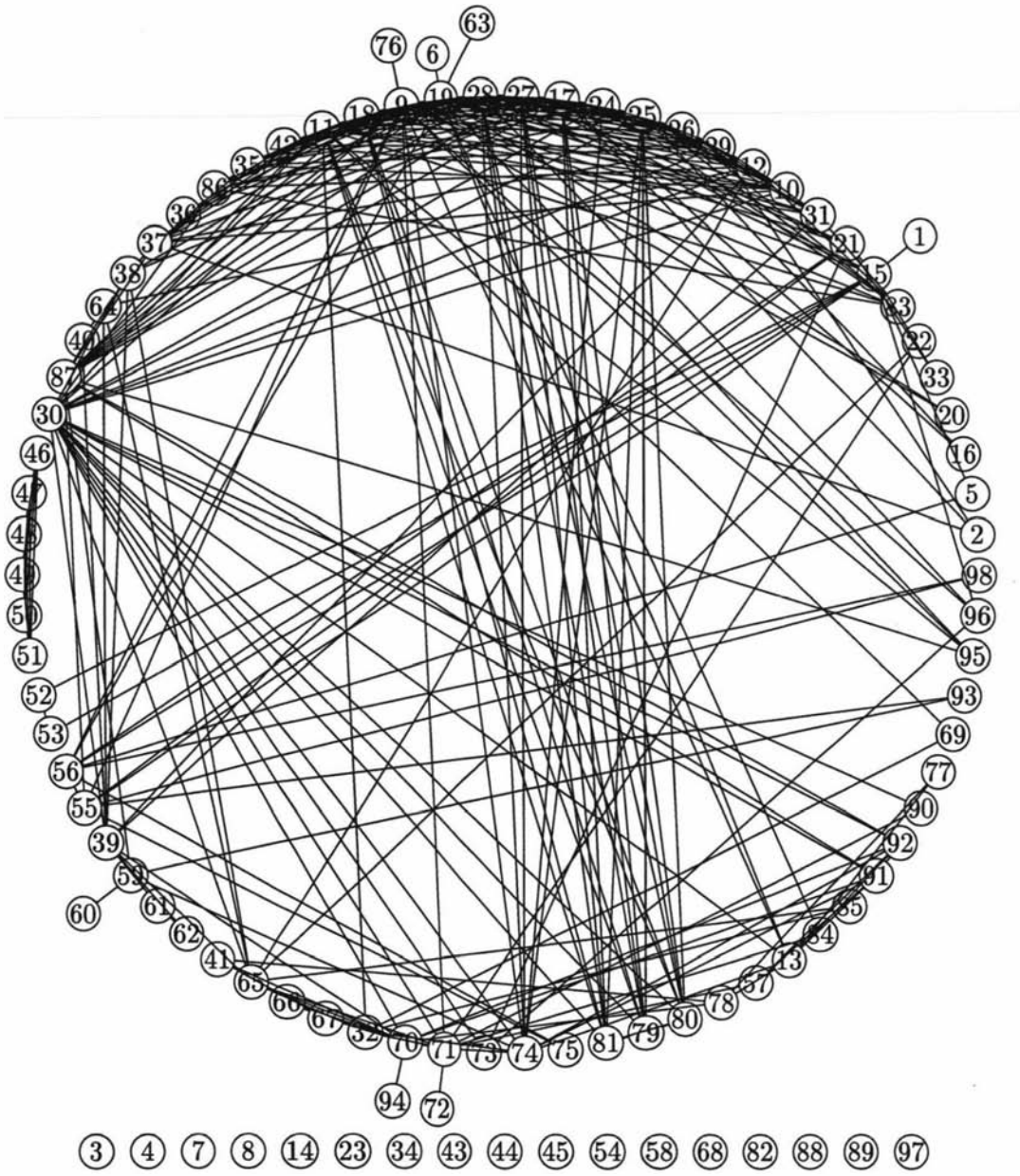


Figure 10.2: Diagram representing the inter-environment connectedness of Onion Data II. Nodes represent the 98 environments and edges show that at least seven of the 87 genotypes were grown in both environments in a pair.

edges, but this would lead to a reduction in clarity. The important relationships of the data are actually those involving genotypes (environments) which have little or no connectivity. The genotypes and environments that are not related to any other genotypes or environments, and those that are related to only one other genotype or environment need to be worked on to improve the overall connectedness of the G×E matrix. It may prove useful to establish if any of these genotypes have become obsolete, or are not likely to become available in future, so that they can be ignored for the development of the trials

Genotype pair	Genotype pair
4 Arad HZ and 79 Texas Grano LO	16 Dessex SS and 19 El Ad HZ
16 Dessex SS and 23 Gladalan Brown YA	16 Dessex SS and 76 Superex TK
19 El Ad HZ and 79 Texas Grano LO	23 Gladalan Brown YA and 34 Houston AS
37 Jenin HZ and 49 RAM 710 HZ	35 Hurricane RS and 42 Mercedes PS

Table 10.4: A list of the significantly altered genotype links that would be improved by the recommendations for the fictitious Yemeni trial. These will affect the inter-genotype structure displayed in Figure 10.1. Most notably, genotypes 35 and 49 would be removed from the list of disconnected genotypes.

programme. These genotypes could be removed from the data to be analysed, but may be crucial for ensuring linkage between all trials.

Selections can be adjusted for a new trial by re-drawing the diagrams to see how the current selections affect the linkage of the G×E matrices. Some seemingly minor alterations to the set of genotypes allocated to new trials may allow more genotypes to have significant relationships with other genotypes which would in turn strengthen the imputation results. For example, if the top ten varieties chosen in Section 9.5 and the four check cultivars ('Dessex SS', 'Jenin HZ', 'Marix ZU', and 'Superex TK') found in Section 9.3, were planned for distribution to the fictitious Yemeni trial, it would be worth investigating what other genotypes could be added to the recommendations to improve the G×E matrix. Using only the Onion Data II results the varieties would be numbered 4, 16, 19, 23, 30, 34, 35, 37, 40, 41, 42, 49, and 76 in Figure 10.1. In this instance, one of the check cultivars ('Superex TK') is actually recommended for testing in the new trial. There is therefore an extra opportunity to include another genotype to enhance the G×E matrix. Adding Genotype 79, better known as 'Texas Grano LO', would allow its comparisons to genotypes 4 'Arad HZ' and 19 'El Ad HZ' to be improved so that links would be formed in Figure 10.1. A complete list of the significantly altered inter-genotype links that would be improved by the recommendations for the fictitious Yemeni trial appear in Table 10.4. Of particular note are the new links established for genotypes 35 'Hurricane RS' and 49 'RAM 710 HZ' which currently have no significant linkage to the majority of genotypes. Selection of 16 'Dessex SS', 34 'Houston AS', and 79 'Texas Grano LO' will mean that the number of significant linkages they have to other genotypes will increase from one to four, two, and three respectively.

As well as saving resources, as genotypes and environments are discarded, duplication of effort can also be reduced. In future, pairs of genotypes that have been tested in a larger number of environments could be barred from being tested in the same trials. For example, genotypes 'Early Red HZ' and 'Red Synthetic HZ' have been grown in 27 and 25 Onion Data I and II environments respectively. Likewise 'Galil HZ' and 'Sivan HZ' have been tested in 24 common environments in both Onion Data I and II. The distances between these pairs of genotypes are currently measured with greater accuracy than any

other comparison, and improving this accuracy would be less useful than increasing the accuracy of other comparisons.

Once the set of genotypes for a new trial are determined, a check needs to be made to ensure that the trial itself is able to be linked to other trials in the programme. If check cultivars are used the chance of having an unlinked trial will be reduced, but there is no guarantee that the new trial will be linked unless the number of check cultivars is equal to or exceeds the level chosen for sufficient linkage.

Discussion thus far has concentrated on improving the connectedness of the current G×E matrices. Ideally the exercise shown above should be carried out on the entire data arising from the Onion Trials Programme, less the data for varieties that are obsolete, or those local varieties that are not available for testing in future. By working with the entire data set, genotype selections for new trials could be made so that more genotypes meet the criteria for inclusion in the next Onion Data I and II sets to be created.

Some importance must be given to the inclusion of data from trials that are not currently part of Onion Data I or II. Some environments will never be included in Onion Data I, let alone Onion Data II, because only six genotypes were tested, but others need to be examined more closely to see if they can be included by testing some of their genotypes in future trials. For example, a 1998 trial 'L01103' from Kenya, tested sixteen genotypes, but only four of them were grown in more than three environments. Another trial, 'A01606' from Yemen in 1994, grew fifteen genotypes, of which five have been included in more than ten trials thus far, but the genotype with the next highest usage was tested in only four environments.

It will prove difficult to include the data from these and other discarded trials, but it may be easier to include some of the currently discarded genotypes into a new Onion Data I or II, if they can be tested in enough future trials. For example, only five of the seven trials in which the variety 'Early Lockyer White YA' was tested were included in Onion Data I. Testing this variety in two more future trials will allow its inclusion in a future Onion Data I set; a third new trial will allow its inclusion in new versions of both Onion Data I and II. 'Early Lockyer White YA' was tested in two environments not yet included in Onion Data I or II in spite of those trials having tested eleven and twelve genotypes. The chances of including the data from these two trials would be improved if 'Early Lockyer White YA' was tested in at least two more environments.

There were surprisingly few genotypes that were tested at four or five trials in the Onion Trials Programme thus far, but many genotypes were tested in six trials. A small amount of effort to get these genotypes included in new versions of Onion Data I and II will widen the scope of the analyses presented in this investigation.

Use of graphical summaries such as those presented in Figures 10.1 and 10.2 could assist trials programme organizers with the enhancement of their programme and therefore its G×E matrices. In some instances it may prove useful to use what is known as the

Genotype	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
(1) Agrifound Dark Red AF	22	13	8	4	4	1	5	2
(2) Agrifound Light Red AF	13	19	11	4	5	2	5	1
(3) Agrifound Rose AF	8	11	13	2	3	3	2	0
(4) Arad HZ	4	4	2	36	5	3	8	7
(5) Australian Brown ST	4	5	3	5	14	1	11	3
(6) Belem IPA-9 IP	1	2	3	3	1	12	2	1
(7) Ben Shemen HZ	5	5	2	8	11	2	22	4
(8) Bon Accord HT	2	1	0	7	3	1	4	10

Table 10.5: The upper left portion of the full concurrence matrix for Onion Data I genotypes. Entries on the leading diagonal show how many times the genotype was to be found in the data set. Only the first eight genotypes are shown. Row and column numbers refer to genotype codes as used in figure 10.1.

concurrence matrix, which shows how many times certain treatments are to be found in the same block together. An example of a concurrence matrix can be found in Table 10.5 for the first eight genotypes of Onion Data I. This matrix presents the P_{ij} values used in the distance measure calculations of the imputation process, and was used to construct Figures 10.1 and 10.2.

Over the course of the Onion Trials Programme thus far, approximately ten trials from each year have been included in Onion Data I and II. If this pattern were to continue, the use of check cultivars will mean that these varieties will have greater linkage to many other varieties. This practice will lead to the concurrence matrix having different ranges of values:

1. Many varieties will not be linked together in a sufficient number of environments $0 \leqslant P_{ij} < q$, which is a direct consequence of sparsity.

2. The majority of varieties should be linked to check cultivars in a sufficient number of environments $P_{ij} > q$.

3. Some varieties will have sufficient linkage to other varieties that are not designated as check cultivars.

4. Pairs of check cultivars will have higher values of P_{ij}.

Strategically increasing P_{ij} values to the level q has been discussed above, but some thought also must be given to the number of genotypes with which each genotype has sufficient linkage.

When the results from this and the preceding chapter were combined as a flow chart (presented in Figure 10.3) to describe the process for developing a trials programme, two-stage imputation of the data seemed like a very small cog in a large machine. Two-stage imputation has, however, allowed the data arising from the Onion Trials Programme to

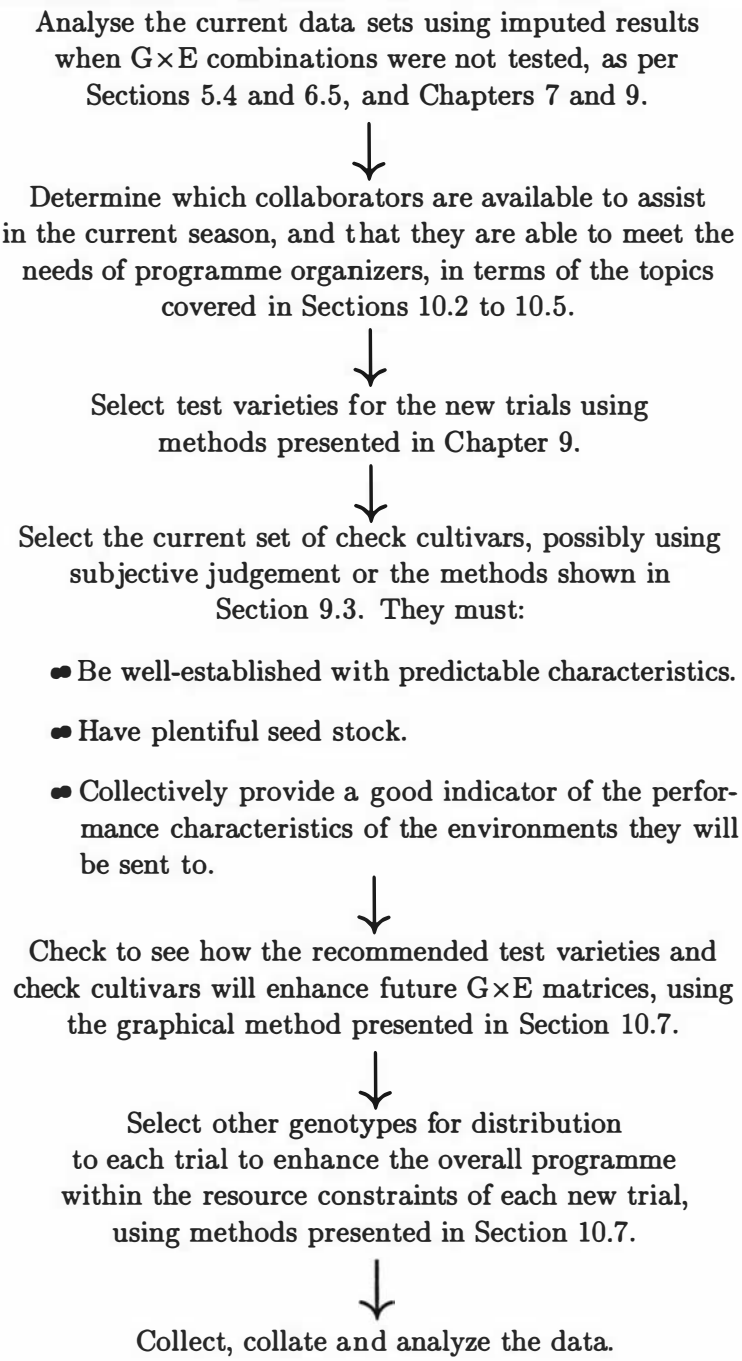


Figure 10.3: A flow chart depiction of the suggested method for selecting genotypes for testing in a new trial.

be analysed in a single analysis, and is therefore, the most essential cog in that machine. As new trials are added to the data used for imputing unobserved G×E combinations, the imputations are likely to change. By using the suggestions of this and the preceding sections, the ability of two-stage imputation to provide adequate estimates of unobserved G×E combinations should be enhanced for the following reasons:

1. Where distances are calculable, they will be measured over a greater number of common environments.

2. Unobserved distances will be estimated more accurately.

3. The resulting genotype clustering will approach the clustering that would arise from use of complete data.

4. The differences in level between genotypes deemed similar by first stage clustering will be measured with greater accuracy.

5. Greater numerical accuracy from distance estimates, and the use of the information from the right genotypes, will lead to better imputations.

The principal research question of this investigation was adjusted at the end of Chapter 8 to give,

> "Given a certain (possibly new) environment and using our current knowledge, which onion varieties should we test in order to find out which ones succeed in terms of their edible yield?"

which is in keeping with the above comments. As trials are added to the programme, imputations should become more accurate. If the accuracy of the imputations can be relied upon, it could be said that the trials programme has used resources efficiently by gaining the greatest information from the least expense of effort.

It is likely, given the evidence of Section 6.4, that imputations are imperfect estimates of unobserved G×E combinations. Improving the accuracy of imputations, however, is achievable and remains a worthwhile goal.

10.8 Summary

The development of two-stage imputation and its application to the data arising from the Onion Trials Programme has implications on the future development of that and any other international programme. From the discussion presented in this chapter, recommendations are summarized below.

For each new trial, programme organizers need to ensure that:

1. the collaborator collects sufficient covariate information to allow the trial's data to be included in the overall programme's statistical analysis. A list of covariates was presented in Section 10.3, from which organizers can determine those that are of particular interest to the trials programme.

2. The needs and abilities of collaborators can be managed within the needs of the trials programme. Minimum standards for the covariate information to be collected and experimental design for trials in the programme should be set, allowing collaborators to provide more if their ability and experience allows. These were discussed in Sections 10.2 to 10.5.

3. If planting density is allowed to vary in a trial, then its impact on yields must be removed using the modelling approaches discussed in Section 10.5.

4. When it is known that a variety will fail, this fact is recorded.

5. A suitable selection of genotypes is sent to the collaborator. This selection should include:

   - The current season's check cultivars, as per the discussion starting on page 294.
   - A set of genotypes recommended using current imputed results.
   - Genotypes that will enhance the overall connectedness of the G×E matrix. Use of the concurrence matrix or graphics similar to Figures 10.1 and 10.2 can assist with ensuring the link between future trials and the data currently in the G×E matrix. They can also be used to strategically choose genotypes that will allow previously discarded trial data to be included in the G×E matrix.
   - Genotypes new to the programme that will be sent to enough new trials to ensure their inclusion in the G×E matrix.

Two-stage imputation rests upon the assumption that data are missing at random. Testing only those genotypes chosen for their likely success in a new trial would break this condition if imputations were accurately estimating unobserved G×E combinations. To ensure that data remain missing at random, some varieties that are suspected of not succeeding should be tested. Concern over the breaking of the missing at random condition is well founded when taking a long term view, where imputations will become more accurate. In the short term, however, imputations are based on inter-genotype relationships that are not measured with great accuracy or are not based on the information that comes from the genotypes that would have been considered the most similar if data were complete.

This chapter has provided a set of guidelines and discussion pointers for trials programme organizers to consider when developing a programme for the coming season. It used the findings arising throughout the development of two-stage imputation and relied heavily on the lessons from experience in the data collection phase. The next (final)

chapter concludes the investigation by bringing all the developments of Chapters 2 to 10 together.

Chapter 11

Conclusion

11.1 Introduction

The purpose of this chapter is to bring together and summarize the findings of this investigation. It describes how the problems of sparsity in the data arising from an international programme of trials were overcome, allowing analysis and providing recommendations for continuing the programme.

There were three stages to this investigation. The first stage, reviewed in the next section, outlined the difficulty of using current methodology on the data arising from the Onion Trials Programme. The development of the new clustering and imputation methods, as well as the methods for comparing clusterings are then reviewed in Section 11.3. Finally, the consequences of applying the new methodology to the data arising from the Onion Trials Programme, and the implications for the ultimate success of the programme, are presented in Section 11.4.

11.2 Background to the investigation

Inspiration for the two-stage imputation process arose from the need to analyse a large and incomplete G×E matrix. A brief history of the Onion Trials Programme was given in Section 3.2. Standard methodologies employed in the past by G×E researchers (introduced in Chapter 2) were, for the most part, unusable due to the incompleteness of the data. Explanations for this incompleteness were given in Section 3.3

When working with complete G×E data where replicates are available, the model, given in (2.3),

$$Y_{ikr} = \mu + G_i + E_k + GE_{ik} + \epsilon_{ikr}$$

can be used to determine the existence of G×E interaction. Testing for the significance of the interaction term is difficult when replicate data are not available, and is made more so when not all G×E combinations are tested. In spite of this, it is likely that the data arising from the Onion Trials Programme have a significant G×E interaction component

as the environments and genotypes differed markedly.

An initial analysis of the data arising from the Onion Trials Programme was presented in Section 3.4, while Sections 3.7 and 3.8 described attempts to apply standard G×E methodology. Unfortunately, data for a large number of infrequently used genotypes and environments needed to be discarded before this modelling was undertaken; the creation of two data sets for use throughout the investigation was described in Section 3.6.

Standard G×E methodologies fall into two broad categories. Researchers are either seeking genotypes that succeed across the entire set of environments under consideration (wide adaptability), or ascertaining to which environments each genotype is most suited (specific adaptation).

Section 2.7 reviewed many stability parameters that have been developed to gauge the wide adaptability of genotypes by identifying those that perform in a constant, or predictable manner. Several of these stability measures were adapted for use with the incomplete data arising from the Onion Trials Programme in Section 3.7 with limited success, and to the fully imputed data in Section 9.3. The adjusted stability measures developed were:

1. The adjusted coefficient of variation, given in (3.4), which uses within-environment standardized yields instead of raw yields in its calculation.

2. The adjusted Wricke's ecovalence, given in (3.6), which uses appropriate averaging to ensure that genotypes grown in greater numbers of environments do not have higher scores than those grown in few environments.

3. The adjusted superiority score, given in (3.8), which also uses appropriate averaging.

Nonparametric stability measures have been proposed that use the ranks of genotypes within environments rather than raw yields, for example Hühn and Nassar (1989). These were not adjusted for incomplete data, but were applied in Section 9.3 with the imputed yields of Onion Data I and II to recommend check varieties.

Establishing where particular genotypes should be grown to take advantage of specific adaptation and optimize the total yield has been achieved in many ways. These methods can be classified by the strategy they use: either a model is employed that explains the existing G×E interaction, or some other means of establishing the dissimilarity of genotype performances across environments has been used. The most popular modelling approaches for explaining G×E interaction have been:

1. Multiple linear regression, such as those performed by Hardwick and Wood (1972) or Piepho *et al.* (1998).

2. Joint regression, using the model originally proposed by Yates and Cochran (1938) given in (2.5). The results of joint regression can be plotted to give a graphical

summary of the findings. An example of the contribution of Finlay and Wilkinson (1963), which plots the genotype coefficients against the mean yield of each genotype, can be seen in Figure 2.1. Other proposed extensions of the joint regression model were reviewed in Section 2.4. The joint regression model was not fitted to the data arising from the Onion Trials Programme because its use assumes that there is only one factor impacting on the G×E interaction, or that a combination of factors affect the G×E interaction in a linear fashion.

3. Factorial regression, using genotype and environment covariates as well as their interactions in a model such as (2.11), introduced in Section 2.5, and employed by Vargas *et al.* (1999). A factorial regression model was proposed in Section 3.5 that related the G×E interaction of a genotype to a quadratic function of the latitudes of environments. This model, given in (3.2), was shown to have limitations in its usefulness for determining which genotypes would succeed in tropical and subtropical locations.

4. Models that use principal components of the G×E matrix in some way. These include:

   (a) The additive main effects and multiplicative interaction (AMMI) model, given in (2.12), which has gained impetus in G×E research over the last twenty years although it had already been in use since the early 1970s in other disciplines.

   (b) The shifted multiplicative model.

   (c) The completely multiplicative model.

   Biplots have often been employed to show the outcome of fitting a multiplicative model to G×E data. In these plots, genotypes and environments are plotted using coordinates based on the principal components. The best environments for each genotype are those that have similar coordinates on the biplot. Examples of biplots were presented in Figures 9.2 and 9.3 for the fully imputed data arising from the Onion Trials Programme. Biplots have also been used as diagnostic tools for selection of the 'best' multiplicative model to employ (Bradu and Gabriel, 1978).

   Another problem with these multiplicative models is that they each give a number of models from which one must be selected.

Except for multiple regression, all of these models are in some way multiplicative in the way they explain the G×E interaction, as they have parameters that are dependent on both genotypes and environments. Joint regression and AMMI models have proved useful when multiple regression and factorial regression models have been limited by unavailability of sufficient covariate data.

Multiplicative models can be employed when working with incomplete data, but initial estimates of main effects and interactions are inaccurate. Digby (1979) proposed an

iterative approach for fitting the joint regression model, but it does not give the correct standard errors for parameters. Ng and Grunwald (1997) identified this problem and showed how correct errors can be found using nonlinear regression techniques. Both of these approaches update parameter estimates using parameter estimates from the previous iteration.

The Healy-Westmacott algorithm, shown to be an implementation of the EM algorithm by McLachlan and Krishnan (1997), can be used when fitting a general linear model to incomplete data. It uses the fitted values from a model to impute missing values, so that the model can be re-fitted, and iterates through these two steps until some convergence criterion is satisfied. In Section 3.8, the EM-AMMI model proposed by Gauch and Zobel (1990), was implemented to Onion Data I and II. This investigation showed that this approach was problematic. The imputed values were dependent on the model being employed, and the starting values used for the algorithm. In Section 6.2 it was argued that the EM-AMMI model would be more effective at imputing missing replicate data than missing G×E combinations.

Cluster analysis has often been employed as a simple and pragmatic means of establishing groups of genotypes that have performed in a similar way across environments. Once groups of genotypes have been established, the G×E matrix can be reduced in size by averaging within groups of genotypes that perform similarly. Simple graphical comparisons can then be viewed to determine which group of genotypes is most suited to each environment.

In some circumstances, it has proved more effective to cluster environments rather than genotypes to achieve the same aim (Ivory *et al.*, 1991), and occasionally to cluster both genotypes and environments to reduce both dimensions of the G×E matrix (Byth *et al.*, 1976; Corsten and Denis, 1990). Outcomes from each of these clustering approaches can be parameterized using models presented in Section 5.5, but as yet this idea has gained little traction in the literature covering G×E analyses.

Godfrey *et al.* (2001), in reporting the preliminary findings of this investigation, provided the first publication of a method for clustering incomplete data on the basis of G×E interaction profile similarity. The next section reviews this extension of cluster analysis methodology, and its use in a method to impute missing G×E data.

11.3 New theoretical developments

This section describes the methods developed in this investigation to analyze the data arising from the Onion Trials Programme. The major developments included:

1. A partitioning of Euclidean distance into two separate components that identify differences among genotype performances in two ways:

(a) Interaction distance measures the dissimilarity of G×E interaction profiles (referred to as 'shape').

(b) Main effect distance measures the dissimilarity of the average performance (referred to as 'level').

2. A two-stage clustering method which first clusters on the shape-similarity using interaction distance, and then re-clusters these groups of shape-similar genotypes using main effect distance.

3. An imputation method which uses the ideas of shape-similarity and the differences in level to estimate missing G×E yields.

4. The determination of mega-environments.

5. A graphical method for comparing cluster analyses.

In Chapter 4, main effect and interaction distance measures were developed for complete and incomplete data. These two distance measures were shown to be related to one another via a partition of Euclidean distance, for both complete and incomplete data, in Section 4.3. They were also shown to have expected values that are dependent on the difference being measured and not the number of common environments over which the calculations were made. This idea built on the work of Ouyang *et al.* (1995) who used squared Euclidean distance when working with incomplete data in a cluster analysis.

The ultimate purpose for developing these distance measures was to allow clustering of the genotypes of Onion Data I and II, but this was not immediately possible due to the sparsity of the data. Many distances were not calculable, and others were based on a small number of common environments. In Section 4.6 these unobserved distances were estimated using a strategy aimed at estimating an upper bound for unobserved distances. It was proposed to use this upper bound to adjust observed inter-genotype distance based on a low number of common environments.

Chapter 5 introduced the two-stage clustering method which used the two new distance measures in successive stages to identify genotypes that were similar in terms of their G×E interaction profiles, and then in terms of their similar performance across environments. A data set from Mungomery *et al.* (1974) was used as a pilot to trial two-stage clustering. Results for a complete and an incomplete data set (found by randomly deleting some of the G×E combinations) were compared in Section 5.3.

The data arising from the Onion Trials Programme were then clustered in Section 5.4 using this two-stage method. A model for two-stage clustering was presented in Section 5.5, which built on the models for other clustering methods. This model, given in (5.6), could be used to provide variables for a general linear model that could be employed to explain genotype yields.

Development of two-stage clustering and the partition of G×E performances into the two components for shape and level inspired the two-stage imputation process outlined in Chapter 6. After some background material covering existing imputation methods for general use was given in Section 6.1, the new algorithm for imputing missing G×E yields was proposed in Section 6.2. In that section, other imputation strategies suitable for use with G×E data were reviewed for their comparison to two-stage imputation via simulation testing. The Mungomery *et al.* (1974) data used in Section 5.3 were used again in Section 6.3 to illustrate two-stage imputation and compare its results with the nearest cluster imputation of Drake (1981).

Results from simulation testing were presented in Section 6.4 and extended the results published in Godfrey *et al.* (2001). The data being imputed were randomly deleted from a number of data sets from the G×E literature, and were therefore missing completely at random. This testing showed that the two-stage imputation method was superior to nearest cluster imputation (Drake, 1981) and the closest observation method, both described in Section 6.2. All three of these methods were superior to use of randomly selected data from the same environment.

Following the success shown by this testing, two-stage imputation was applied to the data arising from the Onion Trials Programme in Section 6.5. As expected, results from imputing Onion Data I and II were similar but not equal. In Section 8.8, however, the imputations were shown to be dependent on the particular G×E combinations that had been tested.

The future development of two-stage imputation was discussed in Section 6.6. This culminated in suggesting the use of the model for two-stage clustering in a multiple imputation process.

Theoretical development moved in a different direction in Chapter 7, to allow the environments of the Onion Trials Programme to be formed into mega-environments. Identifying the similarity of the eight sets of mega-environments was not easy due to the high number of mega-environments, and required development of additional methodology.

Methodology for comparing the mega-environment clusterings was presented in Chapter 8. Cluster influence diagrams were created to compare mega-environment groupings by:

1. Showing the number of mega-environments.

2. Showing the number of environments in each mega-environment.

3. Showing how many environments moved from each mega-environment in one clustering to the mega-environments of a second clustering.

4. Showing how many environments were added to each mega-environment in the clustering of the second set that were not contained in the first set of data.

Numerical summaries were presented in Section 8.3 to provide a means of gauging the similarity of the mega-environment clusterings. Use of the new graphical tool and the numeric summaries uncovered the most important finding of Chapter 8. This was that imputations found for Onion Data I and II were dependent on the set of G×E combinations for which data was available. The fact that Onion Data I and II were too sparse to expect that imputed values were made with sufficient accuracy, forced a change in the way imputations were viewed in terms of their relevance to establishing which onion varieties were likely to succeed in tropical and subtropical environments. Instead of making a bold statement over which onion variety would succeed in each environment, results needed to be qualified by the fact that the predictions were based on the information currently available. Once more trials have been added to the data for analysis, the imputed values will change, and in some circumstances may conflict with current predictions. By carefully selecting varieties for new trials so that current theories can be tested, changes in imputed values can on the whole be viewed as improvements.

Two-stage imputation has not provided a means of answering the principal research question because the data arising from the Onion Trials Programme was too sparse. The method would, however, be used to greater effect on data that are missing completely at random.

As a consequence, strategies for ensuring that data sets arising from the Onion Trials Programme approach a state of missing completely at random over time were developed in Chapter 10.

11.4 Implications and recommendations arising from the investigation

The information gained by imputing the missing entries of Onion Data I and II in Section 6.5, and that from establishing mega-environments in Chapter 7, was combined in Chapter 9 to recommend varieties for testing in a new trial. These recommendations were dependent on the data used (Onion Data I or II) and the set of past trials chosen to be similar to the new trial. When a new trial is planned, however, more than just the recommended varieties need to be tested.

First of all, the connectedness of the G×E matrix must be ensured. New trials can be compared to the existing G×E matrix, but it is recommended that a set of check varieties be selected for use in all trials in the coming season. Several methods were used in Section 9.3 to select check varieties, and these selections depended on the stability measure employed and the genotypes considered (Onion Data I or II).

A second consideration that may affect the genotypes selected for testing in a new trial is that of the linkage between existing genotypes in the overall data set. As discussed in Sections 10.6 and 10.7, some effort is required to ensure that each genotype is grown in

a sufficient number of admissible trials to ensure its inclusion in the data being analyzed. Each genotype needs to also be grown in a reasonable number of environments with a selection of other genotypes so that the clustering of genotypes is not constrained by the pattern of observed to unobserved G×E combinations.

The third consideration is that of the nature of the missingness and the impact this has on imputations. In Section 3.6, G×E combinations known to fail were determined to have a zero yield and included in the data being analyzed, in spite of the fact that they were not observed. This also had the effect of ensuring that the data did not break the missing at random condition; namely, data can be missing according to the G×E combination as long as it is not missing because of its expected yield. As a consequence G×E combinations recommended for future testing should not be chosen solely on the basis of their probable high yields.

At present, the imputed values of Onion Data I and II are dependent on the particular G×E combinations that have been tested. It was recognized that the data arising from the Onion Trials Programme are not missing completely at random, and therefore that the accuracy of imputed values remains questionable. This concern may in fact be the least worrisome of the three because recommendations based on current estimates will therefore not necessarily lead to testing of only high yielding G×E combinations. This concern may need to be reconsidered once more data has been included in the analysis.

Consideration of these three issues should lead to the development of the Onion Trials Programme so as to:

1. Increase the chance that trials not currently being used in the analysis can be incorporated into a future analysis, thus expanding its scope.

2. Improve the accuracy of imputations by improving the average quality of the distance measure estimates and therefore the clustering of genotypes.

3. Allow imputation of unobserved G×E combinations in future by not breaking the missing at random condition.

As well as the overall improvement of the G×E matrix, the Onion Trials Programme will be enhanced by the collection of enough covariates to allow researchers to fully understand all the genotypic and environmental factors that determine the success of certain G×E combinations. Suggestions for the covariates to be collected were presented in Section 10.3. Other issues relevant to the planning of a trials programme were discussed in Sections 10.2, 10.4, and 10.5 which refer to the selection of new environments, the design to be employed in new trials, and addressing any changes in the planting density used within a trial respectively.

11.5 Summary

Times have changed since Lord Rutherford was credited with saying, "If your experiment needs statistics, you ought to have designed a better experiment." This volume contains the developments in statistical theory required to avoid drawing the same conclusion when dealing with the data arising from the Onion Trials Programme. This chapter reviewed the significant findings of the investigation.

The idea of imputing over 80% of observations in a data set, as was achieved for Onion Data I and II, initially seemed unrealistic, but Einstein once said, "If at first the idea is not absurd, then there is no hope for it". His statement summarizes the spirit in which this investigation was carried out. Development of two-stage imputation has provided a relatively simple, but effective, means of allowing analysis of an incomplete G×E data set.

Appendix A

Data from the Onion Trials Programme

The Onion Trials Programme included a total of 400 genotypes. Their names are sometimes descriptive, and on occasion are not the names which they are marketed under; for example, 'Red Synthetic' produced by Hazera Genetics in Israel is marketed as 'Ofir' (Currah, 2002). Some cultivar names appear quite similar, and can be distinguished by the code for their source, usually the relevant seed company. The cultivars used in this work are listed in Table A.1, while the names of the remaining varieties are listed in Table A.2.

Trials for which data was made available were run in over fifty countries. Only 109 and 98 of the trials for which data were made available were included in Onion Data I and IIrespectively. Codes for these appear in Table A.3. Location names, altitudes, latitudes, and the year in which trials were run are presented in Table A.4. Mean and variances of the square roots of yield are also presented in that table, while the proxy variables used in Section 7.4 are presented in Table A.5.

Codes		Genotype name	Codes		Genotype name
1	1	Agrifound Dark Red AF	31	28	HA-222 HZ
2	2	Agrifound Light Red AF	32	29	HA-226 HZ
3	3	Agrifound Rose AF	33		HA-230 HZ
4	4	Arad HZ	34	30	HA-489 HZ
5	5	Australian Brown ST	35	31	HA-675 HZ
6	6	Belem IPA-9 IP	36	32	HA-817 HZ
7	7	Ben Shemen HZ	37		HA-891 HZ
8	8	Bon Accord HT	38	33	HA-950 HZ
9	9	Brownsville AS	39	34	Houston AS
10		Cadix ZU	40	35	Hurricane RS
11	10	Colossal PVP SS	41	36	IRAT-69 MA
12	11	Composto IPA-6 IP	42		Jaguar PS
13	12	Creamgold YA	43	37	Jenin HZ
14	13	Creole Red PRR PS	44		Kano Red NI
15	14	Dehydrator No 3 SS	45	38	Linda Vista PS
16	15	Deko HZ	46	39	Lockyer Gold NW
17	16	Dessex SS	47	40	Marathon HZ
18	17	Early Lockyer Brown YA	48	41	Marix ZU
19	18	Early Red HZ	49	42	Mercedes PS
20	19	El Ad HZ	50		Mr Max RC
21	20	Equanex PS	51		N-53 LO
22	21	Extra Early Creamgold NW	52		Nasik Red LO
23		Eytan HZ	53		Nikita RC
24	22	Galil HZ	54	43	Niv HZ
25	23	Gladalan Brown YA	55	44	NuMex BR-1 RC
26	24	Gladalan White YA	56	45	Ori HZ
27	25	Granex 33 AS	57		PS 8392 PS
28	26	Granex 429 AS	58	46	Pera IPA-4 IP
29		Granex Yellow TK	59		Primero SS
30	27	Granoble PS	60	47	Pusa Red AF

Table A.1: The names of the 104 and 87 onion varieties included in Onion Data I and II respectively. The left column for each variety gives the number assigned to it for Onion Data I, while the right number is used in Onion Data II, after 17 varieties were removed. *Continued on page 320.*

		Genotype name			Genotype name
61	48	Pyramid SA	83	69	Rio Raji Red RC
62	49	RAM 710 HZ	84	70	Rio Ringo RC
63	50	RS 209 RS	85		Riviera AS
64	51	Red Bombay RS	86	71	Rojo SS
65	52	Red Burgundy Imp NE	87	72	Rouge de Tana TS
66	53	Red Comet PS	88		Savages Flat White YA
67	54	Red Creole AS	89	73	Savannah Sweet PS
68	55	Red Creole Credo RS	90	74	Serrana AS
69	56	Red Creole LO	91	75	Sivan HZ
70	57	Red Creole PRR PVP SS	92	76	Superex TK
71	58	Red Creole SA	93	77	Supply YA
72		Red Creole Select NE	94		Texas Grano 438 AS
73	59	Red Star PS	95	78	Texas Grano 502 PRR AS
74	60	Red Synthetic HZ	96	79	Texas Grano LO
75	61	Redbone AS	97	80	Tropic Ace TK
76	62	Regal PVP SS	98	81	Tropic Gold NW
77	63	Regia AS	99	82	Tropicana RS
78	64	Ringer Grano SS	100	83	Utopia AS
79	65	Ringo RS	101	84	Violet de Galmi TS
80	66	Rio Blanco Grande RC	102	85	Yellow Creole SS
81	67	Rio Bravo RC	103	86	Yellow Granex Imp PRR SS
82	68	Rio Hondo RC	104	87	Yodalef HZ

Table A.1: *Continued from page 319.* The names of the 104 and 87 onion varieties included in Onion Data I and II respectively. The left column for each variety gives the number assigned to it for Onion Data I, while the right number is used in Onion Data II, after another 17 varieties were removed.

Genotype names
0-0037 LO, 0-0038 LO, Adama Red LO, Aldobo ZU, Alix ZU,
Aloubassa LO, Angaco INTA LO, Apachi-F1 LO, Arka Nikethan LO,
Arka Pragati LO, Aspen PS, Atlas HZ, B-780 LO, BGS 140 BJ,
BGS 66 BJ, BGS 71 BJ, BGS 82 BJ, BGS 83 BJ, BGS 84 BJ, BGS 85 BJ,
BGS 95 BJ, Bafteem Improved-1 LO, Bafteem Improved-2 LO,
Bafteem LO, Bafteem Yellow LO, Baia Dura AC, Baia Dura AG763 AC,
Baia Periforme AC, Baia Periforme AG558 AC, Barak HZ, Barke LO,
Bawku Red GH, Blanc de Galmi LO, Blanc de Soumarana MA,
Bola Precoce AC, Boldor LO, Bombay Red LO, Bombay White PO,
Bon Accord LO, Bronco HM, Brownsville LO, CV 19 FA, CV 20 FA,
CV 30 LO, CV 90 FA, Cadillac PS, Candy PS, Capri BJ, Caraibe TS,
Centrex SH, Chariot YA, Chata IPA-5 IP, Chula Vista PS,
Composite 1 LO, Composite 4 LO, Composto IPA-6 AS,
Composto IPA-6 Redonda IP, Conquista AC, Contessa AS, Contessa PVP AS,
Corona LO, D 77 LO, DPX 945 LO, Damk LO, Desi Red VR, Don Victor RC,
Dongola-4 LO, Dorata di Parma LO, E Tx Yellow Grano 502 PRR NG,
E Y Tx Grano 502 PRR VM, E515 YA, Early Creamgold YA,
Early Golden Globe YA, Early Lockyer White YA, Early Supreme SS,
Early Yellow LO, Early Yellow Premium HT, Egyptian HZ, Egyptian ZR,
El Hilo LO, Encino AS, Endeavour YA, Ex-Dala LO, Excalibur RC,
Explorer RS, Faridpur Bhati LO, First Edition LO, Flare BJ,
Footlong White HN, Franciscana IPA-10 IP, Geesa LO, Gindin Tasa LO,
Gladiator YA, Gladstone LO, Gold Rush LO, Golden Brown LO,
Golden Mosque LO, Gorgia Red LO, Goudanir LO, Granale LO,
Granex 2000 HZ, Granex HZ, Grano 2000 HZ, Grano F1 2000 HZ,
Grano HZ, HA-1208 HZ, HA-220 HZ, HA-234 HZ, HA-508 HZ, HA-642 HZ,
HA-688 HZ, HA-815 HZ, HA-870 HZ, HA-888 HZ, HA-944 HZ, HA-95 HZ,
Hazera 508 HZ, Henry's Special AB, Henry's Special SS, Hojem HT,
Hojem LO, Hojem MF, Hojem SA, Hojem ST, Jaune Espagnol VM,

Table A.2: Genotypes not included in either Onion Data I or II. *Continued on page 322.*

Genotype names
Jaune de Valence VM, Jubileu AC, Jubiley 50 LO, K-1 SL, K1 LO,
Kalpitiya SL, Kamleen LO, Kathmandu local LO, La Joya AS, Liberty BJ,
Local LO, Local Red LO, Mallajh LO, Marquesa AS, Marquis SH,
Miltry Onion LO, Moab HZ, Moonlight RS, Mutuali IPA-8 IP, N-53 AF,
Nasik Red AF, Nasik Red PO, Nasik Red SI, Nitzan ZR, Noflaye TS, Ole AB,
Orient BJ, PRS 11390 PS, PS 13589 PS, PS 492 PS, PS 8489 PS,
PSR 11390 PS, PSR 1190 PS, PSR 2091 PS, Pantanoso LO, Patriot LO,
Payola SH, Pera IPA-6 AC, Perla PVP AS, Phulkara VR, Pira Ouro AC,
Poona Red PO, Primavera PS, Pusa Madhavi LO, Pusa Red 90 Yala LO,
Pusa Red IN, Pusa Red LO, Pusa Red SI, Pusa Safed LO, Pyramid HT,
Pyramid LO, RAM 735 HZ, RC line RC, RCS1903 RC, RCS1919 RC,
RCS2211 RC, RCS2302 RC, RCS9306 RC, RCSX-1941 RC, RD 77 LO, RS 204 RS,
RS 218 RS, RS 226 RS, RS 266 RS, RS 303 RS, RS 392 PS, RS 505 RS,
RS 506 RS, RS 513 RS, RS 514 RS, RS 533 RS, Radar BJ, Radium LO,
Rampure IN, Red Bandana HM, Red Baron LO, Red Bombay BK, Red Bombay LO,
Red Comet RS, Red Creole (Nepal) LO, Red Creole C-5 NE, Red Creole NL,
Red Granex SS, Red Kano NI, Red Onion AS, Red Pinoy EW, Red Pinoy LO,
Red Pinoy PH, Red Poona LO, Red Wave PS, Redwing BJ, Ringer Grano Imp RC,
Rio Bonita RC, Rio Corona RC, Rio Enrique RC, Rio Jefe RC, Rio Plata RC,
Rio Redondo RC, Rio San Juan RC, Rio Selecto RC, Rio Solo RC,
Rio Sonora RC, Rio Unico RC, Rio Zorro RC, Riverside HZ, Robin BJ,
Robust SS, Rojo 38057 SS, Roxa IPA-3 IP, Rustler HM, SSC 6008 SH,
SSC 6060 SH, Saggai-1 LO, Samaru 5 LO, Senshyu IT, Solist BJ,
Special 38 AB, Star 5504 SA, Stetson HM, Swat PK, Sweet Dixie RC,
Sweet Georgia RC, Sweet Sunrise RC, Synthetic LO, Taherpuri LO,
Tainan 1 LO, Tainung Sel.3 LO, Tarmagon LO, Tex-400 JP,
Texas Early Grano 502 RS, Texas Early Grano 502 ST, Texas Early Grano LO,
Texas Grano 502 PRR ST, Tontal INTA LO, Torrens White YA, Trimontium LO,
Tropi Red LO, Tropic Brown LO, Tropicana, Tropicana LO, Valcatorce LO,
ValeOuro IPA-11 IP, Vatikiotiko LO, Veronique LO, Violet de Galmi LO,
Violet de Galmi MA, Violet de Galmi No 5 TS, Violet de Galmi No MA 027 TS,
Violet de Galmi VM, Wallon Brown LO, White Creole PRR PVP SS,
White Creole PRR RS, White Creole PRR SS, White Creole RS, White Creole SS,
XPH 6074 AS, XPH 6700 AS, XPH 6712 AS, XPH 8407 AS, YHO 30 YA,
YHO 34 YA, YHO 37 YA, Yakouri LO, Yellow Novex SS, Z-204 NE, Z-218 AB,
Z-235 AB, Z-250 AB, Z-513 NE, Z512 YA, Z516 YA.

Table A.2: *Continued from page 321.* Genotypes not included in either Onion Data I or II.

Codes		Trial	Codes		Trial	Codes		Trial
1	1	A00301	38	38	O02701	75	69	X04804
2	2	A00501	39	39	O02702	76	70	X05701
3	3	A01601	40	40	O02703	77	71	X07201
4	4	A01604	41	41	O02709	78	72	X07202
5	5	A03101	42	42	O07601	79	73	X10801
6	6	A03401	43	43	O07603	80	74	X10901
7	7	A03602	44	44	O07604	81	75	X13801
8	8	A03603	45	45	O11601	82	76	X13901
9	9	A04001	46		P07200	83	77	X15101
10	10	A04101	47	46	P07201	84	78	X15103
11	11	A04102	48	47	P07202	85	79	X15105
12	12	A04103	49	48	P07203	86	80	X15106
13	13	A04202	50	49	P07204	87	81	X15107
14	14	A05902	51	50	P07205	88	82	Y01302
15	15	D02501	52	51	P07206	89	83	Y01401
16	16	D03601	53	52	R00101	90	84	Y01501
17	17	D03602	54	53	R00102	91	85	Y03402
18	18	D03603	55	54	R00401	92		Y03404
19	19	D03604	56	55	R00701	93	86	Y07401
20	20	D03605	57	56	R00702	94	87	Y07403
21	21	D03610	58	57	W05201	95	88	Y07405
22	22	D07101	59		W13204	96	89	Y07501
23	23	D08401	60	58	W16502	97	90	Y13101
24	24	F02404	61		W16503	98		Y13102
25	25	F02405	62	59	W16506	99		Y13901
26	26	F02406	63	60	W16508	100		Z00101
27	27	F02407	64		W17501	101	91	Z03401
28	28	F02408	65		W17502	102	92	Z03402
29	29	F02701	66	61	W17503	103	93	Z03501
30	30	F02702	67	62	W17504	104		Z03901
31	31	F04006	68	63	W22801	105	94	Z03903
32	32	J00301	69	64	W23001	106	95	Z09601
33	33	J08801	70	65	W23003	107	96	Z09603
34	34	L01101	71	66	W23005	108	97	Z10501
35	35	L03701	72	67	W23007	109	98	Z13503
36	36	L05101	73	68	W23009			
37	37	O02700	74		X04601			

Table A.3: The 'Tricodes' of the 109 and 98 trials included in Onion Data I and II respectively. The left column for each trial gives the number assigned to it for Onion Data I, while the right number is used in Onion Data II, after another eleven trials were removed.

Trial code	Country	Site	Year	Lat	Alt	Used	Mean	Variance
A00301	Barbados	St Philip	1995	13	50	28	2.12	0.24
A00501	Nigeria	Kadawa	1991	12	476	16	1.11	0.04
A01601	Yemen	Seiyun	1992	16	600	24	1.42	0.11
A01603	Yemen	Seiyun	1993	16	600	12	1.45	0.02
A01604	Yemen	Seiyun	1993	16	600	13	1.52	0.17
A01606	Yemen	Seiyun	1994	16	600	15	2.13	0.10
A03101	Malaysia	Klang	1998	3	3	32	0.91	0.05
A03401	Belize	Cayo	1990	17	75	10	0.57	0.03
A03602	Australia	Gatton	1996	28	90	23	1.10	1.41
A03603	Australia	Gatton	1997	28	95	18	2.20	0.06
A04001	India	Nasik	1993	20	550	37	1.10	0.12
A04101	Bangladesh	BAU	1990	27	19	17	0.75	0.09
A04102	Bangladesh	BAU	1991	27	19	32	1.07	0.06
A04103	Bangladesh	BAU	1992	27	19	27	0.82	0.11
A04202	Mauritania	Kaedi	1994	16	17	25	1.38	0.09
A05901	Fiji	Sigatoka	1991	18	11	7	1.04	0.05
A05902	Fiji	Sigatoka	1991	18	11	20	1.77	0.01
C04401	PNG	Laloki	1993	9	30	11	1.69	0.06
C04402	PNG	Laloki	1993	9	30	8	1.44	0.09
D02501	Egypt	Giza	1996	30	19	56	0.92	0.07
D03601	Zimbabwe	Marondera	1989	18	1630	19	3.24	0.16
D03602	Zimbabwe	Marondera	1990	18	1630	30	2.75	0.18
D03603	Zimbabwe	Matopos	1990	20	1365	23	3.13	0.28
D03604	Zimbabwe	Marondera	1991	18	1630	42	3.04	0.32
D03605	Zimbabwe	Matopos	1991	20	1365	22	3.91	0.07
D03610	Zimbabwe	Marondera	1999	18	1630	27	1.69	0.08
D07101	Burkina	Bobo	1993	11	405	15	1.12	0.50
D08401	Mexico	Hermosillo	1992	20	149	14	2.76	0.14
F02404	Botswana	Sebele	1993	25	994	25	2.56	0.52
F02405	Botswana	Sebele	1993	25	994	26	1.96	0.23
F02406	Botswana	Sebele	1993	25	994	24	1.20	0.09
F02407	Botswana	Sebele	1994	25	994	24	1.75	0.57
F02408	Botswana	Sebele	1994	25	994	26	2.15	0.70
F02701	Tanzania	Morogoro	1992	7	524	17	1.08	0.15
F02702	Tanzania	SUA	1994	7	524	22	1.36	0.06
F02703	Tanzania	SUA	1996	7	524	6	0.87	0.04
F04006	Uganda	Kawanda	1993	0	1320	25	0.72	0.04
J00301	India	Pune	1994	19	559	22	1.57	0.09
J08001	Benin	Kargui	1998	12	200	11	1.50	0.04
J08801	Mauritius	Richelieu	1998	20	66	30	1.82	0.08
L01101	Kenya	Thika	1997	1	1549	16	1.23	0.06

Table A.4: Summary information for the 123 environments of the Onion Trials Programme. Square roots of yield were used for means and variances. *Continued on page 325.*

Trial code	Country	Site	Year	Lat	Alt	Used	Mean	Variance
L01103	Kenya	Thika	1998	1	1549	16	1.03	0.03
L03701	Lesotho	Maseru	1992	29	1500	18	1.19	0.23
L05101	South Africa	Roodeplaat	1993	26	1164	20	2.67	0.16
O02700	Sri Lanka	FCRDI	1991	8	137	17	1.30	0.07
O02701	Sri Lanka	FCRDI	1992	8	137	23	1.19	0.06
O02702	Sri Lanka	FCRDI	1992	8	137	20	0.90	0.01
O02703	Sri Lanka	FCRDI	1993	8	137	17	1.29	0.04
O02707	Sri Lanka	FCRDI	1994	8	137	12	1.14	0.16
O02709	Sri Lanka	FCRDI	1994	8	137	18	0.52	0.03
O07601	Mozambique	Umbeluzi	1991	26	10	20	1.86	0.12
O07603	Mozambique	Umbeluzi	1997	26	10	16	1.60	0.05
O07604	Mozambique	Umbeluzi	1997	26	10	18	1.41	0.05
O11601	Senegal	St Louis	1993	16	2	29	1.99	0.05
P07200	Nepal	Lumle	1990	28	1675	7	2.34	0.32
P07201	Nepal	Lumle	1991	28	1650	11	2.19	0.22
P07202	Nepal	Tapu	1991	28	950	12	1.51	0.15
P07203	Nepal	Lopre	1991	28	2250	12	1.03	0.21
P07204	Nepal	Rishingpatan	1991	28	430	12	1.77	0.12
P07205	Nepal	Keware	1991	28	1100	12	2.29	0.30
P07206	Nepal	Sigana	1991	28	1240	12	1.57	0.11
R00101	Australia	Emerald	1996	24	178	31	2.80	0.06
R00102	Australia	Emerald	1997	24	178	30	2.80	0.67
R00401	Tunisia	Sahline	1999	36	10	25	1.26	0.16
R00701	India	Rajgurunagar	1998	19	554	31	1.22	0.18
R00702	India	Rajgurunagar	1998	19	554	26	1.81	0.20
W05201	Bulgaria	Plovdiv	1995	42	200	19	0.93	0.04
W13204	Nepal	PAC	1992	27	1747	7	1.29	0.02
W16502	Brazil	Belem PE	1992	9	305	16	1.50	0.26
W16503	Brazil	Belem PE	1993	9	305	9	1.23	0.12
W16506	Brazil	Belem	1994	9	305	21	1.47	0.32
W16508	Brazil	Belem	1995	9	305	18	1.86	0.15
W17501	Greece	Athens	1993	38	30	8	2.22	0.13
W17502	Greece	Athens	1993	38	30	8	2.00	0.06
W17503	Greece	Athens	1994	38	30	13	2.33	0.36
W17504	Greece	Athens	1994	38	30	12	1.79	0.09
W17505	Greece	Athens	1995	38	30	6	2.20	0.41
W17506	Greece	Athens	1995	38	30	6	2.16	0.33
W22801	Italy	Catania	1990	38	20	11	1.23	0.06
W23001	Ethiopia	Melkassa	1993	8	1550	20	1.18	0.08
W23003	Ethiopia	Melkassa	1993	8	1550	18	1.58	0.09
W23005	Ethiopia	Melkassa	1994	8	1550	15	0.96	0.03

Table A.4: *Continued from page 324.* Summary information for the 123 environments of the Onion Trials Programme. *Continued on page 326.*

Trial code	Country	Site	Year	Lat	Alt	Used	Mean	Variance
W23007	Ethiopia	Melkassa	1996	8	1550	18	1.09	0.05
W23008	Ethiopia	Melkassa	1996	8	1550	8	1.10	0.04
W23009	Ethiopia	Melkassa	1997	8	1550	21	1.06	0.39
X04601	Cameroon	Maroua	1994	11	370	9	1.18	0.83
X04804	Zambia	Mazabuka	1993	16	978	20	1.92	0.10
X05701	Nigeria	Sokoto	1994	13	300	21	1.28	0.07
X07201	Nepal	Chitwan	1993	28	228	25	0.99	0.05
X07202	Nepal	Rampur	1994	28	228	10	1.48	0.02
X10801	P R China	Jinan	1998	37	250	22	1.94	0.13
X10901	New Caledonia	La Foa	1994	21	38	27	1.98	0.17
X13801	Mali	Bamako	1997	13	320	21	1.48	0.10
X13901	Senegal	CDH	1994	15	20	24	1.47	0.17
X15101	Guinee	Bareng	1993	11	1000	20	1.52	0.04
X15103	Guinee	Bareng	1994	11	1000	26	1.23	0.04
X15105	Guinee	Ley-Tolin	1995	11		22	1.49	0.06
X15106	Guinee	Tolo	1995	11		22	0.99	0.08
X15107	Guinee	Sanama	1995	11		22	0.96	0.02
Y01302	Argentina	San Juan	1997	32	618	24	2.03	0.14
Y01401	Uruguay	Sagayo	1993	35	27	22	1.55	0.09
Y01501	Kenya	Njoro	1992	0	2225	14	0.86	0.03
Y03402	C Ivoire	Ferke	1994	10	323	19	0.44	0.08
Y03404	C Ivoire	Ferke	1998	10	323	17	1.20	0.04
Y07401	Taiwan	AVRDC	1992	23	50	23	2.14	0.28
Y07403	Taiwan	AVRDC	1993	23	50	27	2.11	0.15
Y07405	Taiwan	AVRDC	1996	23	50	24	2.73	0.35
Y07501	Honduras	EAP	1996	14	780	32	1.95	0.05
Y13101	St Helena	Longwood Field	1997	16	470	8	1.53	0.42
Y13102	St Helena	Mulberry Gut	1997	16	410	8	2.64	0.10
Y13901	Nigeria	Bauchi	1995	10	609	15	1.99	1.25
Z00101	Thailand	Maejo	1991	18	750	10	1.15	0.66
Z03401	Pakistan	Jarma	1994	34	580	20	0.85	0.03
Z03402	Pakistan	Jarma	1995	34	580	20	0.52	0.04
Z03501	Philippines	NOMIARC	1996	8	800	19	2.33	0.09
Z03901	Ghana	Kumasi	1997	7	255	10	0.84	0.02
Z03903	Ghana	Kumasi	1998	7	255	20	1.08	0.06
Z09601	Korea	Pusan	1994	35	4	18	3.01	0.03
Z09603	Korea	Pusan	1995	35	4	25	2.41	0.07
Z10501	Cape Verde	San Domingos	1996	15	247	13	0.97	0.14
Z12201	Malawi	Bvumbwe	1996	16	1150	12	1.28	0.05
Z13501	Nigeria	Samaru	1997	11	686	9	0.59	0.09
Z13503	Nigeria	Samaru	1997	11	686	27	0.95	0.05

Table A.4: *Continued from page 325.* Summary information for the 123 environments of the Onion Trials Programme.

Trial	Heat unit variables				Photoperiod variables			
code	1	2	3	4	1	2	3	4
A00301	20.91	21.09	21.56	22.91	11.63	11.94	12.35	12.77
A00501	17.36	13.94	17.42	24.22	11.70	11.59	11.90	12.41
A01601	20.02	16.49	17.86	24.65	11.76	11.35	11.94	12.92
A01603	20.06	14.77	16.31	24.22	11.87	11.35	11.73	12.77
A01604	20.03	14.76	16.13	24.09	11.86	11.35	11.70	12.76
A01606	21.92	15.56	16.31	22.42	11.87	11.36	11.62	12.59
A03101	25.14	24.66	24.12	23.58	12.28	12.36	12.40	12.34
A03401	18.48	19.02	20.88	22.89	11.22	11.64	13.10	13.10
A03602	13.31	9.88	8.98	13.73	10.98	10.55	11.09	12.36
A03603								
A04001	13.79	13.74	16.84	22.09	11.04	11.22	11.83	12.71
A04101	20.33	15.91	13.48	18.23	11.38	10.69	10.93	11.99
A04102	15.09	14.06	20.19	22.86	10.66	11.04	12.17	13.37
A04103	16.73	12.30	16.33	20.05	10.78	10.70	11.45	12.53
A04202	23.37	20.39	23.54	27.05	11.37	11.37	11.86	12.60
A05901	18.67	17.07	17.83	17.87	11.50	11.16	11.26	11.75
A05902	16.86	17.64	17.58	19.06	11.34	11.24	11.80	12.64
C04401	21.45	21.55	21.00	21.43	11.76	11.70	11.82	12.12
C04402	21.55	20.97	21.66	22.12	11.70	11.82	12.15	12.50
D02501								
D03601								
D03602								
D03603								
D03604								
D03605								
D03610								
D07101	19.74	18.65	21.98	23.95	11.62	11.64	11.90	12.33
D08401	11.66	10.53	14.53	20.19	11.28	11.20	12.06	13.09
F02404	17.88	13.41	10.08	14.95	12.05	11.00	10.91	11.94
F02405	16.22	11.08	10.70	17.02	11.56	10.80	11.10	12.23
F02406	15.23	10.59	10.89	17.61	11.36	10.77	11.25	12.43
F02407	17.83	11.25	7.65	13.72	12.02	11.09	10.83	11.81
F02408	15.05	9.28	8.18	15.05	11.45	10.81	10.97	11.96
F02701	22.00	20.64	18.74	16.88	12.30	12.09	11.90	11.86
F02702	16.15	16.97	19.01	20.28	11.80	11.92	12.22	12.48
F02703	17.31	16.12	17.19	18.79	11.84	11.86	12.01	12.29
F04006	21.14	20.00	19.16	18.98	12.20	12.20	12.20	12.20
J00301	19.19	15.42	13.53	18.06	11.62	11.20	11.19	11.78
J08001								
J08801	20.77	17.86	17.68	19.02	11.37	11.05	11.43	12.27
L01101	15.24	15.09	15.42	15.70	12.20	12.20	12.20	12.20

Table A.5: Artificial covariates used for clustering environments in Section 7.4. These values have been created using formulae given in (7.1) and (7.2). *Continued on page 328.*

Trial	Heat unit variables				Photoperiod variables			
code	1	2	3	4	1	2	3	4
L01103	14.75	13.33	12.28	13.68	12.20	12.20	12.20	12.20
L03701	6.44	4.96	11.09	17.95	10.69	10.81	12.40	13.86
L05101	12.18	7.64	8.51	14.61	11.33	10.73	11.21	12.53
O02700	23.43	23.39	24.03	22.44	12.68	12.63	12.40	12.01
O02701	23.36	23.10	23.20	22.93	12.68	12.67	12.50	12.16
O02702	19.74	19.88	21.50	22.80	11.80	11.84	12.05	12.28
O02703	23.93	23.49	23.15	23.73	12.55	12.68	12.66	12.40
O02707	24.00	23.50	23.50	23.82	12.65	12.69	12.56	12.35
O02709								
O07601	18.50	15.61	14.19	17.36	11.36	10.75	10.80	11.91
O07603	18.00	14.24	14.27	19.28	11.48	10.70	11.10	12.58
O07604								
O11601	22.22	21.05	22.00	25.35	11.46	11.31	11.58	12.34
P07200	4.80	6.91	12.47	14.61	10.55	11.36	12.67	13.80
P07201	7.68	4.08	12.59	13.92	10.74	10.87	12.38	13.78
P07202	11.66	7.86	11.80	18.53	10.77	10.67	11.77	13.22
P07203	4.74	0.35	7.72	11.99	10.74	10.88	12.32	13.74
P07204	16.10	10.24	12.55	20.18	10.81	10.60	11.53	13.09
P07205	10.66	6.91	11.78	17.91	10.77	10.72	11.95	13.50
P07206	10.60	6.04	11.73	17.80	10.75	10.76	12.03	13.53
R00101	13.36	10.91	12.69	16.48	10.93	10.89	11.51	12.54
R00102								
R00401								
R00701								
R00702								
W05201								
W13204	4.29	8.37	12.73	14.81	10.87	11.89	13.07	13.86
W16502	21.94	20.76	19.52	19.39	11.94	11.76	11.70	11.86
W16503	21.82	20.60	19.44	19.43	11.93	11.75	11.72	11.87
W16506	22.15	21.57	20.87	19.78	12.18	11.95	11.77	11.71
W16508	21.60	19.59	19.10	20.37	11.82	11.70	11.75	11.97
W17501	9.85	6.59	7.12	14.28	10.25	10.01	11.84	14.07
W17502	7.88	5.98	7.79	14.24	9.93	10.25	12.13	14.10
W17503	11.51	5.31	7.03	15.71	10.59	9.96	11.93	14.31
W17504	6.35	5.47	9.52	18.11	9.86	10.63	12.93	14.66
W17505	9.48	4.91	5.85	16.51	10.59	9.96	11.93	14.31
W17506	6.80	4.58	6.49	17.28	9.95	10.36	12.44	14.46
W22801	11.60	8.45	10.74	17.75	9.97	11.54	13.63	14.56
W23001	15.06	16.28	17.24	16.15	11.89	12.23	12.56	12.69
W23003								
W23005	14.35	13.63	16.94	17.60	11.87	11.75	12.00	12.44

Table A.5: *Continued from page 327.* Artificial covariates used for clustering environments in Section 7.4. These values have been created using formulae given in (7.1) and (7.2). *Continued on page 329.*

Trial	Heat unit variables				Photoperiod variables			
code	1	2	3	4	1	2	3	4
W23007	17.27	17.43	17.02	16.39	12.07	12.45	12.69	12.51
W23008								
W23009	15.93	20.99	15.75	17.62	11.90	11.74	11.87	12.17
X04601	21.68	19.65	22.35	27.08	11.70	11.62	11.85	12.34
X04804	12.34	11.66	15.54	20.29	11.36	11.42	11.94	12.75
X05701	21.93	19.60	20.03	24.82	11.67	11.50	11.71	12.25
X07201	15.82	12.24	15.52	23.15	10.85	10.64	11.58	13.21
X07202	13.89	9.70	16.16	23.73	10.74	10.70	11.79	13.20
X10801								
X10901	12.29	13.27	13.66	17.30	11.07	11.59	12.37	13.14
X13801	19.84	19.95	24.43	27.70	11.52	11.58	11.95	12.44
X13901	20.92	18.43	16.82	15.85	11.48	11.42	11.77	12.31
X15101	16.08	14.35	15.28	19.27	11.75	11.60	11.82	12.31
X15103	14.33	15.01	15.84	19.40	11.62	11.63	11.93	12.47
X15105	14.70	15.00	17.27	19.01	11.62	11.67	12.05	12.52
X15106	15.92	16.88	19.15	19.92	11.62	11.65	11.97	12.49
X15107								
Y01302	15.78	6.87	6.01	10.98	11.63	10.41	10.64	12.26
Y01401	13.29	7.28	6.58	13.00	11.22	10.08	11.11	13.70
Y01501	11.67	10.41	10.79	11.07	12.20	12.20	12.20	12.20
Y03402								
Y03404	20.39	20.97	23.59	24.33	11.70	11.76	12.03	12.40
Y07401	17.96	14.67	14.32	19.00	11.33	10.87	11.50	12.85
Y07403	20.97	15.42	14.00	16.16	11.77	10.95	11.13	12.23
Y07405	20.09	14.87	15.02	17.61	11.61	10.90	11.34	12.47
Y07501	19.32	15.78	16.60	18.84	11.93	11.53	11.45	11.91
Y13101								
Y13102	14.80	14.89	13.01	11.70	12.75	11.95	11.38	11.68
Y13901	19.13	18.70	21.28	25.74	11.68	11.65	11.92	12.33
Z00101	17.60	15.72	18.85	24.59	11.27	11.30	11.90	12.67
Z03401	10.25	6.58	11.36	20.55	10.39	10.34	11.83	13.56
Z03402	13.64	7.24	14.19	23.31	10.60	10.26	11.90	13.89
Z03501	18.17	19.27	19.50	19.33	12.10	12.40	12.65	12.67
Z03901	22.22	22.13	23.80	22.33	12.30	12.50	12.60	12.56
Z03903	22.80	22.33	22.09	20.81	12.20	12.38	12.55	12.60
Z09601	13.02	0.71	2.55	10.05	11.67	10.11	11.63	14.04
Z09603	12.03	0.00	0.00	11.80	11.71	10.08	11.47	13.99
Z10501	19.66	20.15	19.69	19.40	11.48	11.91	12.39	12.93
Z12201	11.42	9.81	13.13	16.30	11.36	11.36	11.74	12.40
Z13501	21.17	19.42	19.35	19.35	12.76	12.88	12.68	12.36
Z13503	19.20	16.88	18.55	21.47	11.74	11.60	11.75	12.10

Table A.5: *Continued from page 328.*Artificial covariates used for clustering environments in Section 7.4. These values have been created using formulae given in (7.1) and (7.2).

Appendix B

References

Abou-El-Fittouh, H.A., J.O. Rawlings, and P.A. Miller (1969) "Classification of environments to control genotype by environment interactions with an application to cotton." *Crop Science* **9**, 135–140.

Alagarswamy, G., E.J. van Oosterom, and S.C. Sethi (1996) "International multi-environment trials at the International Crops Research Institute for the Semi-Arid Tropics (ICRISAT)" in Cooper, M., and G.L. Hammer (Eds) (1996) *Plant adaptation and crop improvement.* CAB International, Wallingford, United Kingdom, 165–174.

Anderberg, M.R. (1973) *Cluster Analysis for Applications.* Academic Press, New York.

Baker, R.J. (1969) "Genotype-environment interactions in yield of wheat." *Canadian Journal of Plant Science* **49**, 743–751.

Baker, R.J. (1988) "Tests for crossover genotype-environmental interactions." *Canadian Journal of Plant Science* **68**, 405–410.

Baril, C.P., J.B. Denis, and P. Brabant (1994) "Selection of environments using simultaneous clustering based on genotype×environment interaction." *Canadian Journal of Plant Science* **74**, 311–317.

Baril, C.P., J.B. Denis, R. Wustman, and F.A. van Eeuwijk (1995) "Analysing genotype by environment interaction in Dutch potato trials using factorial regression." *Euphytica* **82**, 149–155.

Basford, K.E. (1982) "The use of multidimensional scaling in analysing multi-attribute genotype response across environments." *Australian Journal of Agricultural Research* **33**, 473–480.

Basford, K.E., and J.W. Tukey (1998) *Graphical approaches to multiresponse data: Illustrated with a plant breeding trial.* Chapman & Hall, London.

Becker, H.C., and J. Leon (1988) "Stability analysis in plant breeding." *Plant Breeding* **101**, 1–23.

Bleasdale, J.K.A. (1966) "The effects of plant spacing on the yield of bulb onions (Alioum Cepa L.) grown from seed." *Journal of Horticultural Science* **41**, 145–153.

Boyd, D.A., L.T.K. Yuen, and P. Needham. (1976) "Nitrogen requirement of cereals."

Journal of Agricultural Science **87**, 149–162.

Bradu, D., and K.R. Gabriel (1978) "The biplot as a diagnostic tool for models of two-way tables." *Technometrics* **20**, 47–68.

Brancourt-Hulmel, R.M., V. Biarnes-Dumoulin, and J.B. Denis (1997) "Points de repère dans l'analyse de la stabilité et de l'interaction génotype-milieu en amélioration des plantes." *Agronomie* **17**, 219–246.

Breese, E.L. (1969) "The measurement and significance of genotype-environment interactions in grasses." *Heredity* **24**, 27–44.

Brewster, J.L. (1990) "The influence of cultural and environmental factors on the time of maturity of bulb onion crops." *Acta Horticulturae* **267**, 289–296.

Brewster, J.L. (1994) *Onions and other vegetable alliums.* CAB International, Wallingford, UK.

Brewster, J.L. (1997) "Onions and garlic" in Wien, H.C. (Ed.) *The Physiology of Vegetable Crops.* CAB International, Walingford, UK.

Brewster, J.L., and P.J. Salter (1980) "The effect of plant spacing on the yield and bolting of two cultivars of over wintered bulb onions." *Journal of Horticultural Science* **55**, 97–102.

Byth, D.E., R.L. Eisemann, and I.H. DeLacy (1976) "Two way pattern analysis of a large data set to evaluate genotype adaptation." *Heredity* **37**, 215–230.

Chatfield, C., and A.J. Collins (1980) *Introduction to multivariate analysis.* Chapman & Hall, London.

Cochran, W.G., and G.M. Cox (1957) *Experimental Designs* (2nd edition), John Wiley & Sons Inc., New York.

Cooper, M., D.E. Byth, and I.H. DeLacy (1993) "A procedure to assess the relative merit of classification strategies for grouping environments to assist selection in plant breeding regional evaluation trials." *Field Crops Research* **35**, 63–74.

Cooper, M., and I.H. Delacy (1994) "Relationships among analytical methods used to study genotypic variation and genotype-by-environment interaction in plant breeding multi-environment experiments." *Theoretical and Applied Genetics* **88**, 561–572.

Cornelius, P.L. (1993) "Statistical tests and retention of terms in the additive main effects and multiplicative interaction for cultivar trials." *Crop Science* **33**, 1186–1193.

Cornelius, P.L., D.A.V. Sanford, and M.S. Seyedsadr (1993) "Clustering cultivars into groups without rank-change interactions." *Crop Science* **33**, 1193–1200.

Corsten, L.C.A., and J.B. Denis (1990) "Structuring interaction in two-way tables by clustering." *Biometrics* **46**, 207–215.

Crossa, J. (1988) "A comparison of results obtained with two methods for assessing yield stability." *Theoretical and Applied Genetics* **75**, 460–467.

Crossa, J. (1990) "Statistical analysis of multi-location trials." *Advances in Agronomy* **44**, 55–85.

Crossa, J., P.L. Cornelius, M. Seyedsadr, and P. Byrne (1993) "A shifted multiplicative

model cluster analysis for grouping environments without genotypic rank change." *Theoretical and Applied Genetics* **85**, 577–586.

Crossa, J., H.G. Gauch Jnr., and R.W. Zobel (1990) "Additive main effects and multiplicative interaction analysis of two international maize cultivar trials." *Crop Science* **30**, 493–500.

Currah, L. (2002a) "Onions in the tropics cultivars and country reports." In Rabinowitch, H.D., and L. Currah (Eds) *Allium Crop Science Recent Advances.* CAB International, Wallingford, United Kingdom. pp. 379–407.

Currah, L. (2002b) personal communication.

Currah, L. (2003) personal communication.

Currah, L., A.J.R. Godfrey, M.A. Nichols, G.R. Wood (2001) "Improving the methods used in collaborative onion trials in order to facilitate genotype×environment analyses — Lessons from experience." in Randall, W.M. (Ed.)*Proceedings of the Third International Symposium on Edible Alliaceae.* Athens, Georgia, USA, 30 October to 3 November 2000, 26–31.

Currah, L., G.M. Green,and J.E. Orchard (1997) *International collaborative short-day onion trials, 1990–95.* Supplement to *Onion Newsletter for the Tropics.* Natural Resources Institute, The University of Greenwich, Chatham, United Kingdom.

Currah, L., D.J. Midmore, and G.R. Wood (1999) "Relating experimental onion yields to environmental factors across the tropics and sub-tropics." *Allium Improvement Newsletter* **8**, 46–49.

de Ruiter, J.M. (1986) "The effects of temperature and photoperiod on onion bulb growth and development" *Proceedings of the Agronomy Society of N.Z.* **16**, 93–100.

DeLacy, I.H., K.E. Basford, M. Cooper, J.K. Bull, and C.G. McLaren (1996) "Analysis of multi-environment trials — An historical perspective." in Cooper, M., and G.L. Hammer (Eds) *Plant Adaptation and Crop Improvement.* CAB International, Wallingford, United Kingdom, 39–124.

DeLacy, I.H., M. Cooper, and P.K. Lawrence (1990) "Pattern analysis over years of regional variety trials: Relationship among sites." in Kang, M.S. (Ed.) *Genotype by Environment Interaction and Plant Breeding.* Louisiana State University, Baton Rouge, 189–213.

DeLacy, I.H., and P.K. Lawrence (1988) "Combining pattern analysis over years — Classification of locations." *Proceedings of the 9th Australian Plant Breeding conference*, Wagga Wagga, 175–176.

Denis, J.B., and J.C. Gower (1996) "Asymptotic confidence regions for biadditive models: Interpreting genotype-environment interactions." *Applied Statistics* **45**, 479–493.

Digby, P.G.N. (1979) "Modified joint regression analysis for incomplete variety×environment data." *Journal of Agricultural Science* **93**, 81–86.

Drake, W.D. (1981) *The GEBEI analysis package.* Queensland Department of Primary Industries, Misc. Pub. 81022, Brisbane.

Eberhart, S.A., and W.A. Russell (1966) "Stability parameters for comparing varieties." *Crop Science* **6**, 36–40.

Eskridge, K.M. (1990) "Selection of stable cultivars using a safety first rule." *Crop Science* **30**, 369–374.

Eskridge, K.N., O.S. Smith, and P.F. Byrne (1993) "Comparing test cultivars using reliability functions of test check differences from on-farm trials." *Theoretical and Applied Genetics* **87**, 60–64.

Everitt, B.S. (1993) *Cluster Analysis* (3rd edition), Edward Arnold, London.

Finlay, K.W., and G.N. Wilkinson (1963) "The analysis of adaptation in a plant-breeding programme." *Australian Journal of Agricultural Research* **14**, 742–754.

Flores, F., M.T. Moreno, and J.I. Cubero (1998) "A comparison of univariate and multivariate methods to analyze GxE interaction." *Field Crops Research* **56**, 271–286.

Flores, F., M.T. Moreno, A. Martinez and J.I. Cubero (1996) "Genotype-environment interaction in faba bean: Comparison of AMMI and principal coordinate models." *Field Crops Research* **47**, 117–127.

Fowlkes, E.B., and C.L. Mallows (1983) "A method for comparing two hierarchical clusterings." *Journal of the American Statistical Association* **78**, 553–569.

Fox, P.N., and A.J. Rathjen (1981) "Relationships between sites used in the interstate wheat variety trials" *Australian Journal of Agricultural Research* **32**, 691–702.

Fox, P.N., and A.A. Rosielle (1982) "Reducing the influence of environmental main effects on pattern analysis of plant breeding environments." *Euphytica* **31**, 645–656.

Francis, T.R., and L.W. Kannenberg (1978) "Yield stability studies in short-season maize. I. A Descriptive Method for Grouping Genotypes." *Canadian Journal of Plant Science* **58**, 1029–1034.

Frappell, B.D. (1973) "Plant spacing of onions." *Journal of Horticultural Science* **48**, 19–28.

Freeman, G.H. (1973) "Statistical methods for the analysis of genotype-environment interactions." *Heredity* **31**, 339–354.

Freeman, G.H. (1975) "Analysis of interactions in incomplete two-way tables." *Applied Statistics* **24**, 46–55.

Freeman, G.H. (1985) "The analysis and interpretation of interactions." *Journal of Applied Statistics* **12**, 3–10.

Freeman, G.H., and B.D. Dowker (1973) "The analysis of variation between and within genotypes and environments." *Heredity* **30**, 97–109.

Freeman, G.H., and J.M. Perkins (1971) "Environmental and genotype-environmental components of variability VIII: Relations between genotypes grown in different environments and measures of these environments." *Heredity* **27**, 15–23.

Ganesalingam, S., and G.J. McLachlan (1978) "The efficiency of a linear discriminant function based on unclassified initial samples." *Biometrika* **65**, 658–662.

Gauch Jnr, H.G. (1988) "Model selection and validation for yield trials with interaction." *Biometrics* **44**, 705–715.

Gauch Jnr, H.G. (1990) "Full and reduced models for yield trials." *Theoretical and Applied Genetics* **80**, 153–160.

Gauch Jnr, H.G. (1992). *Statistical analysis of regional yield trials: AMMI analysis of factorial designs.* Elsevier Science Publishers B.V., Amsterdam.

Gauch Jnr, H.G., and R.W. Zobel (1990) "Imputing missing yield trial data." *Theoretical and Applied Genetics* **79**, 753–761.

Gauch Jnr, H.G., and R.W. Zobel (1996) "Optimal replication in selection experiments." *Crop Science* **36**, 838–843.

Gauch Jnr, H.G., and R.W. Zobel (1997) "Identifying mega-environments and targeting genotypes." *Crop Science* **37**, 311–326.

Ghaderi, A., E.H. Everson, and C.E. Cress (1980) "Classification of environments and genotypes in wheat." *Crop Science* **20**, 707–710.

Gibbons, J.D. (1970) *Nonparametric statistical inference.* McGraw-Hill, New York.

Godfrey, A.J.R., G.R. Wood, S. Ganesalingam, M.A. Nichols,and C.G.Qiao (2001) "Two-stage clustering in genotype-by-environment analyses with missing data." *Journal of Agricultural Science (Camb)* **139**, 67–77.

Godfrey, A.J.R., G.R. Wood, and M.A. Nichols (1999) "'Gotta know your onions' or 'A new approach for clustering cultivars in genotype-environment analysis with sparsity in the data'" Abstracts of the New Zealand Statistical Association 50th Anniversary Conference held in Wellington, New Zealand, July 4–7, 1999.

Gollob, H.F. (1968) "A statistical model which combines features of factor analytic and analysis of variance techniques." *Psychometrika* **33**, 73–115.

Goodchild, N.A., and W.J.R. Boyd (1975) "Regional and temporal variations in wheat yield in Western Australia and their implications in plant breeding." *Australian Journal of Agricultural Research* **26**, 209–217.

Goodman, L.A., and W.H. Kruskal (1954) "Measures of association for cross-classification." *Journal of the American Statistical Association* **49**, 732–764.

Gower, J.C., (1985) "Measures of similarity, dissimilarity, and distance." in Kotz, S., N.L. Johnson, and C.B. Read (Eds) *Encyclopedia of Statistical Sciences* **5**. John Wiley & Sons, New York.

Gower, J.C., (1990) "Three-dimensional biplots." *Biometrika* **77**, 773–785.

Gusmão, L., J.T. Mexia, and J. Baeta (1992) "Trimmed joint regression: A new approach to the joint regression analysis for cultivar relative-performance evaluation." *Theoretical and Applied Genetics* **84**, 735–738.

Hardwick, R.C., and J.T. Wood (1972) "Regression methods for studying genotype-environment interactions." *Heredity* **28**, 209–222.

Hegemann, V., and D.E. Johnson (1976) "The power for two tests for nonadditivity."

Journal of the American Statistical Association **71**, 945–948.

Hill, J. (1975) "Genotype-environment interactions — A challenge for plant breeding." *Journal of Agricultural Science* **85**, 477–493.

Hühn, M., and R. Nassar (1989) "On tests of significance for nonparametric measures of phenotypic stability." *Biometrics* **45**, 997–1000.

Ivory, D.A., S. Kaewmeechai, I.H. DeLacy, and K.E. Basford (1991) "Analysis of the environmental component of genotype×environment interaction in crop adaptation evaluation." *Field Crops Research* **28**, 71–84.

Jalaluddin, M., and S.A. Harrison (1993) "Repeatability of stability estimators for grain yield in wheat." *Crop Science* **33**, 720–725.

John, J.A., and E.R. Williams (1995) *Cyclic and computer generated designs* (2nd edition). Chapman & Hall, London.

Johnson, G.R. (1977) "Analysis of genotypic similarity in terms of mean yield and stability of environmental response in a set of maize hybrids." *Crop Science* **17**, 837–842.

Johnson, D.E., and F.A. Graybill (1972) "An analysis of a two-way model with interaction and no replication." *Journal of the American Statistical Association* **67**, 862–868.

Kang, M.S. (1998) "Using genotype-by-environment interaction for crop cultivar development." *Advances in Agronomy* **62**, 199–252.

Kang, M.S., D.P. Gorman, and H.N. Pham (1991) "Application of a stability statistic to international maize yield trials." *Theoretical and Applied Genetics* **81**, 162–165.

Kang, M.S., J.D. Miller, and L.L. Darrah (1987) "A note on relationship between stability variance and ecovalance." *Journal of Heredity* **78**, 107.

Kang, M.S., and H.N. Pham (1991) "Simultaneous selection for high yielding and stable crop genotypes." *Agronomy Journal* **83**, 161–165.

Keisling, T.C. (1982) "Calculation of the length of day." *Agronomy Journal* **74**, 758–759.

Kempton, R.A. (1984) "The use of biplots in interpreting variety by environment interactions." *Journal of Agricultural Science* **103**, 123–135.

Kempton, R.A., J.C. Seraphin, and A.M. Sword (1994) "Statistical analysis of two-dimensional variation in variety yield trials." *Journal of Agricultural Science* **122**, 335–342.

Knight, R. (1970) "The measurement and interpretation of genotype-environment interactions." *Euphytica* **19**, 225–235.

Krzanowski, W.J. (1988) *Principles of multivariate analysis: A user's perspective.* Oxford University Press, Oxford.

Lawrence, P.K., and I.H. DeLacy (1993) "Classification of locations in regional cotton variety trials where trial entries change over years." *Field Crops Research* **34**, 195–207.

Lefkovitch, L.P. (1980) "Conditional clustering." *Biometrics* **36**, 43–48.

Lefkovitch, L.P. (1985) "Multi-criteria clustering in genotype-environment interaction problems." *Theoretical and Applied Genetics* **70**, 585–589.

Lin, C.S. (1982) "Grouping genotypes by a clustering method directly related to genotype-environment interaction mean square." *Theoretical and Applied Genetics* **62**, 277–280.

Lin, C.S., and M.R. Binns (1985) "Procedural approach for assessing cultivar-location data: Pairwise genotype-environment interactions of test cultivars with checks." *Canadian Journal of Plant Science* **65**, 1065–1071.

Lin, C.S., and M.R. Binns (1988a) "A superiority measure of cultivar performance for cultivar×location data." *Canadian Journal of Plant Science* **68**, 193–198.

Lin, C.S., and M.R. Binns (1988b) "A method of analysing cultivar×location×year experiments: a new stability parameter." *Theoretical and Applied Genetics* **76**, 425–430.

Lin, C.S., and M.R. Binns (1989) "Comparison of unpredictable environmental variation generated by year and by seeding-time factors for measuring type IV stability." *Theoretical and Applied Genetics* **78**, 61–64.

Lin, C.S., and M.R. Binns (1991a) "Assessment of a method for cultivar selection based on regional trial data." *Theoretical and Applied Genetics* **82**, 379–388.

Lin, C.S., and M.R. Binns (1991b) "Genetic properties of four types of stability parameter." *Theoretical and Applied Genetics* **82**, 505–509.

Lin, C.S., and M.R. Binns (1994) "Concepts and methods for analysing regional trial data for cultivar and location selection." *Plant Breeding Reviews* **12**, 271–297.

Lin, C.S., M.R. Binns, and L.P. Lefkovitch (1986) "Stability analysis: Where do we stand?" *Crop Science* **26**, 894–900.

Lin, C.S., and G. Butler (1990) "Cluster analyses for analysing two-way classification data." *Agronomy Journal* **82**, 344–348.

Lin, C.S., and M.J. Morrison (1992) "Selection of test locations for regional trials of barley." *Theoretical and Applied Genetics* **83**, 968–972.

Lin, C.S., and B. Thompson (1975) "An empirical method of grouping genotypes based on a linear function of the genotype-environment interaction." *Heredity* **34**, 255–263.

Little, R.J.A. (1988) "Missing data adjustments in large surveys." *Journal of Business and Economic Statistics* **6**, 287–296.

Little, R.J.A., and D.B. Rubin (1987) *Statistical Analysis with Missing Data.* John Wiley & Sons, New York.

McLachlan, G.J., and K.E. Basford (1988) *Mixture Models: Inference and Applications to Clustering.* Marcel Dekker, New York.

McLachlan, G.J., and T. Krishnan (1997) *The EM Algorithm and its Extensions.* John Wiley & Sons, New York.

Malhotra, N.K. (1987) "Analyzing marketing research data with incomplete information on the dependent variable." *Journal of Marketing Research* **24**, 74–84.

Manly, B.F.J. (1994) *Multivariate Statistical Methods: A Primer* (2nd edition). Chapman & Hall, London.

May, K.W., and G.C. Kozub (1995) "Genotype×environment interactions for two-way

barley grain yield and implications for selection of test locations." *Canadian Journal of Plant Science* **75**, 571–575.

Menz, K.M. (1980) "A comparative analysis of wheat adaptation across international environments using stochastic dominance and pattern analysis." *Field Crops Research* **3**, 33–41.

Milliken, G.A., and D.E. Johnson (1989) *Analysis of Messy Data Volume II: Nonreplicated Experiments.* Chapman & Hall, New York.

Moro, J., and J.B. Denis (1997) "Selecting genotypes by clustering, for qualitative genotype by environment interaction, using a non-symmetric inferiority score." *Agronomie* **17**, 283–289.

Mungomery, V.E., R. Shorter, and D.E. Byth (1974) "Genotype×Environment interactions and environmental adaptation. I Pattern analysis — Application to soya bean populations." *Australian Journal of Agricultural Research* **25**, 59–72.

Ng, M.P. (2001) "Quadratic extension of a joint-regression model." *Abstracts from the 52nd Annual New Zealand Statistical Association Conference*, held in Christchurch, New Zealand, December 10–13, 2001, 49.

Ng, M.P., and G.K. Grunwald (1997) "Nonlinear regression analysis of the joint-regression model." *Biometrics* **53**, 1366–1372.

Ng, M.P., and E.R. Williams (2001) "Joint-regression analysis for incomplete two-way tables." *Australian and New Zealand Journal of Statistics* **43**, 201–206.

Nichols, M.A. (1970) "A note on the reciprocal yield-density model." *Horticultural Research* **10**, 88–90.

O'Neill, M.E., and K. Mathews (2000) "A weighted least squares approach to Levene's test of homogeneity of variance." *Australian and New Zealand Journal of Statistics* **42**, 81–100.

Ouyang, Z., R.P. Mowers, A. Jensen, S. Wang, and S. Zheng (1995) "Cluster analysis for genotype×environment interaction with unbalanced data." *Crop Science* **35**, 1300–1305.

Perkins, J.M., and J.L. Jinks (1968) "Environmental and genotype-environmental components of variability III. Multiple lines and crosses." *Heredity* **23**, 339–356.

Pham, H.N., and M.S.Kang (1988) "Interrelationships among and repeatability of several stability statistics estimated from international maize trials." *Crop Science* **28**, 925–928.

Piepho, H.P. (1994a) "Older and recent approaches for assessing yield stability in agricultural crops." *Tropenlandwirt, Beiheft* **52**, 103–115.

Piepho, H.P. (1994b) "Missing observations in the analysis of stability." *Heredity* **72**, 141–145.

Piepho, H.P. (1994c) "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. (1994d) "Partitioning genotype-environmental interaction in regional yield

trials via a generalised stability variance." *Crop Science* **34**, 1682–1685.

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., J.B. Denis, and F.A. van Eeuwijk (1998) "Predicting Cultivar Differences Using Covariates." *Journal of Agricultural, Biological and Environmental Statistics* **3**, 151–162.

Ramey, T.B., and A.A. Rosielle (1983) "HASS Cluster analysis: A new method of grouping genotypes for environments in plant breeding." *Theoretical and Applied Genetics* **66**, 131–133.

Rand, W.M. (1971) "Objective criteria for the evaluation of clustering methods." *Journal of the American Statistical Association* **66**, 846–850.

Rubin, D.B. (1976) "Inference and missing data." *Biometrika* **63**, 581–592.

Sayedsadr, M., and P.L. Cornelius (1992) "Shifted multiplicative models for nonadditive two-way tables." *Communications in Statistics: Simulation and Computation* **21**, 807–832.

Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data.* Chapman & Hall, London.

Seif, E., J.C. Evans, and L.N. Balaam (1979) "A multivariate procedure for classifying environments according to their interaction with genotypes." *Australian Journal of Agricultural Research* **30**, 1021–1026.

Shorter, R., D.E. Byth, and V.E. Mungomery (1977) "Genotype × environment interaction and environmental adaptation. ii assessment of environmental contributions." *Australian Journal of Agricultural Research* **28**, 223–235.

Shukla, G.K. (1972) "Some statistical aspects of partitioning genotype-environmental components of variability." *Heredity* **29**, 237–245.

Siegel, S. (1956) *Nonparametric Statistics for the Behavioral Sciences*, McGraw-Hill Inc., New York.

Snee, R.D. (1982) "Nonadditivity in a two-way classification: Is it interaction or nonhomogeneous variance." *Journal of the American Statistical Association* **77**, 515–519.

St-Pierre, C.A., H.R. Klinck, and F.M. Gauthier (1967) "Early generation selection under different environments as it influences adaptation of barley." *Canadian Journal of Plant Science* **47**, 507–517.

Theil, H. (1972) *Statistical Decomposition Analysis.* North Holland Publishing Co., Amsterdam.

Tukey, J.W. (1949) "One degree of freedom for non-additivity." *Biometrics* **5**, 232–242.

van Eeuwijk, F.A. (1995a) "Linear and bilinear models for the analysis of multi-environment trials: I. An inventory of models." *Euphytica* **84**, 1–7.

van Eeuwijk, F.A. (1995b) "Multiplicative interaction in generalised linear models." *Biometrics* **51**, 1017–1032.

Vargas, M., J. Crossa, F.A. van Eeuwijk, and M.E. Ramírez (1999) "Using partial least

squares regression, factorial regression, and AMMI models for interpreting genotype×environment interaction." *Crop Science* **39**, 955–967.

Verma, M.M., G.S. Chahal, and B.R. Murty (1978) "Limitations of conventional regression analysis: A proposed modification." *Theoretical and Applied Genetics* **53**, 89–91.

Voltas, J., F.A. van Eeuwijk, A. Sombrero, A. Lafarga, E. Igartua, and I. Romagosa (1999) "Integrating statistical and exophysiological analyses of genotype by environment interaction for grain filling of barley I. Individual grain weight." *Field Crops Research* **62**, 63–74.

Wade, L.J., C.G. McLaren, B.K. Samson, K.R. Regmi, and S. Sarkarung (1996) "The importance of environmental characterization for understanding genotype by environment interactions" in Cooper, M., and G.L. Hammer (Eds) (1996) *Plant adaptation and crop improvement.* CAB International, Wallingford, United Kingdom, 549–562.

Ward Jr, J.H. (1963) "Hierarchical grouping to optimize an objective function." *Journal of the American Statistical Association* **58**, 236–244.

Westcott, B. (1986) "Some methods of analysing genotype-environment interaction." *Heredity* **56**, 243–253.

Westcott, B. (1987) "A method for assessing the yield stability of crop genotypes." *Journal of Agricultural Science* **108**, 267–274.

Williams, E.R., D.J. Lucket, P.E Reid and N.J. Thomson (1992) "Comparison of locations used in cotton-breeding trials." *Australian Journal of Experimental Agriculture* **32**, 739–746.

Williams, E.R., and J.T. Wood (1993) "Testing the significance of genotype-environment interaction." *Australian Journal of Statistics* **35**, 359–362.

Wright, A.J. (1971) "The analysis and prediction of some two factor interactions in grass breeding." *Journal of Agricultural Science* **76**, 301–306.

Wright, A.J. (1976) "The significance for breeding of linear regression analysis of genotype-environment interactions." *Heredity* **37**, 83–93.

Yan, W., and L.A. Hunt (1998) "Genotype by environment interaction and crop yield" *Plant Breeding Reviews* **16**, 135–178.

Yates, F., and W.G. Cochran (1938) "The analysis of groups of experiments." *Journal of Agricultural Science* **28**, 556–580.

Yau, S.K. (1991) "Need of scale transformation in cluster analysis of genotypes based on multi-location yield data." *Journal of Genetics and Breeding* **45**, 71–76.

Yau, S.K. (1995) "Regression and AMMI analyses of genotype×environment interactions: an empirical comparison." *Agronomy Journal* **87**, 1211–126.

Zobel, R.W., M.J. Wright, and H.G. Gauch Jnr (1988) "Statistical analysis of a yield trial." *Agronomy Journal* **80**, 388–393.