Evaluation of alternative sugarcane selection strategies

Thesis submitted by

Vanja Calija-Zoppolato BSc(Hons) UK

in June 2004

for the degree of Doctor of Philosophy
in the Mathematics and Statistics and Tropical Plant Science within the Schools of
Mathematical and Physical Sciences and Tropical Biology
James Cook University of North Queensland

Statement of Access

I, Vanja Calija-Zoppolato, the author of this thesis, understand that James Cook University of North Queensland will make it available for use within the University Library and, by microfilm or other means, allow access to users in other approved libraries.
All users consulting this thesis will have to sign the following statement:

In consulting this thesis I agree not to copy or closely paraphrase it in whole or in part without consent of the author, and to make proper written acknowledgment for any assistance, which I have obtained from it

Beyond this, I do not whish to place any restrictions on access to this thesis.

16. 08. 2005

Vanja Calija-Zoppolato Date

Abstract

The international competitiveness and success of the Australian sugar industry, which is one of the world's largest exporters of raw sugar depends on increased cane yield and advanced farming practices. One of the key drivers for a sustainable sugar industry is therefore, to increase cane yield through designing efficient breeding programs, that aim at producing new and improved varieties of cane. Selection for superior genotypes is the most important aspect of sugarcane breeding programs, and is a long and expensive process. It consists of a number of stages where at each stage some genotypes are chosen for further selection and some are discarded from future selection. Designing a selection system is a complex task, with varying parameters at each stage. While studies have investigated different components of selection independently, there has not been a whole system approach to improve the process of selection.

The aim of this research was to develop a tool for the optimisation of selection systems. The problem of designing an efficient selection system has two components: firstly, evaluating the performance of selection systems and secondly, deciding on a combination of selection variables that will select the most promising genotypes. These two components were designated sub-objectives, one and two respectively.

To address the first sub-objective, data on previous selection trials was collected and used to predict gain for different selection designs. The value that is used to compare the performances of different selection systems is what was called in this thesis, the genetic gain for economic value $\widetilde{G}$, a measure based on the estimate of a potential economic value of a genotype if planted as a cultivar. The connection between $\widetilde{G}$ and choice taken for selection variables at various stages is complex and not expressible by a simple set of

formulas. Instead, a computer based stochastic simulation model SSSM (Sugarcane Selection Simulation Model) was developed.

To eliminate as many simplifying assumptions as possible and bring the study as close to real life as possible, the quantitative genetics of sugarcanes relevant to selection was studied. Furthermore, a specific sugarcane-breeding region was targeted, the Burdekin region (Australia). To ensure the accuracy of the SSSM, it performance was verified and the sensitivity analysis was performed to identify those variance parameters to which it is most sensitive.

By, developing the SSSM this study approached the problem as an integrated system, where if one parameter changes the state of the whole system changes. Furthermore, by creating an accurate selection simulation model a new methodology for evaluating alterative sugarcane selection strategies was obtained. A new methodology that tests the performance of different selection designs prior to their field trials and also tests the impact any change in the estimated variance components may have on selection, will be a potential money saver for the industry. Furthermore, the SSSM can be directly applied to any region targeted by sugarcane breeding programs or to other clonally propagated crops.

The second sub-objective was addressed by the development of the optimisation algorithm called ASSSO (Algorithm for the Sugarcane Selection Simulation Optimisation), a combination of dynamic programming and branch-and-bound. The ASSSO was applied to the Burdekin region to identify selection designs that maximise selection outputs. Apart from providing a new approach to the problem of optimising selection system, the ASSSO also presents a new application of dynamic programming and branch-and-bound.

The ASSSO identified a number of alternative selection systems that are significant improvements to the practices currently used in the Burdekin region. Nevertheless, the purpose of this research was not to suggest that the intuitions and experiences of plant breeders can be replaced by the set of guidelines obtained using a computer simulation,

but rather to validate the benefits of a joint venture between mathematicians and plant breeders.

Acknowledgments

First of all I wish to thank my family for helping me through all difficulties I encountered during this research. I am grateful to my supervisory team: Dr Leone Bielig, Prof Dr Danny Coomans, Dr Andrew Higgins and Dr Phillip Jackson. Their contribution to this research and constant guidance has been invaluable.

I would like to thank Professor Bob Lawn personally as well as the entire CRC Sugar for their help and support. I also wish to thank Mr Mike Cox and Mr Terry Morgan for their expertise and comments throughout the research. I would like to express my great appreciation to both past and present members of the Mathematics and Statistics staff at the JCU, above all to Dr Trevor Waechter, Dr Christine Ormond and Dr Michael Steel.

Contents

List of Figures

List of Tables

List of Acronyms

Table 1

List of acronyms used throughout the thesis

ASSSO	Algorithm for Sugarcane Selection System, Optimisation
BSES	Bureau of Sugar Experiment Station
CCS	Commercial Cane Sugar
CSIRO	Commonwealth Scientific and Industrial Research Organisation
CSR	formerly known as Colonial Sugar Refinery and today the name of a privately owned sugar research company
DP	Dynamic Programming
MANOVA	Multivariate Analysis of Variance
NMG	Net Merit Grade
OR	Operations Research
SRDC	Sugar Research and Development Corporation
SSSM	Sugarcane Selection Simulation Model
TCH	Tonnes Cane per Hectare – cane yield
TSH	Tonnes Sugar per Hectare – sugar yield

List of Symbols

Table 2

List of the symbols frequently used throughout the thesis

k	starting number of genotypes in the case of clonal selection or starting number of genotypes per family in case of family selection;
f	number of families to start selection;
b	family selection to be used in stage one $\in \{TRUE, FALSE\}$;
n	number of stages;
p_z	plot size for each stage $z = 1,2,...,n$;
s_z	number of sites (locations) for each stage $z = 1,2,...,n$;
r_z	number of replications per location for each stage $z = 1,2,...,n$;
d_z	selection index for each stage $z = 1,2,...,n$;
t_z	selection intensity for each stage $z = 1,2,...,n$;
y_{ijk}	phenotypic (observed) value of a genotype i being planted in the environment j in plot size k;
μ	grand mean of the trait observed;
g_i	genetic effect of genotype i;
v_{ij}	environmental effect of genotype i being planted in environment j
c_{ik}	competition effect of genotype i being planted in plot size k;
x_{ij}	genotype by environment interaction effect of genotype i being planted in environment j;
e_{ijk}	error effect;
σ_P^2	phenotypic variance;
σ_g^2	genetic variance (true yield);
σ_f^2	between-family genetic variance;
σ_w^2	within-family genetic variance;
σ_v^2	environmental variance;
σ_x^2	genotype by environment interaction variance;
σ_c^2	competition variance;
σ_τ^2	genotype by competition interaction variance;
σ_e^2	error variance;
σ_A^2	additive genetic variance (breeding value);
σ_D^2	dominance part of the non-additive genetic variance;
σ_I^2	epistasis part of the non-additive genetic variance;
$\text{cov}(g,v)$	covariance between genotype and environment;
$\text{cov}(g,c)$	covariance between genotype and competition;
H^2	broad sense heritability of a trait;

h^2	narrow sense heritability of a trait;
R	response to selection;
S	selection differential;
δ	proportion of the genetic variability attributable to the between families;
$\rho_{g,g+c}$	genetic correlation between a plot size and pure stand;
$\rho_{g,c}$	genetic correlation between genotype and competition;
$\rho_{CCS,TCH}$	phenotypic correlation between CCS and cane yield (TCH);
E_i	economic value of genotype i;
R_i	return from sugar produced when growing genotype i commercially;
C_i	costs associated with growing genotype i commercially;
G_i	genetic effect for economic value of genotype i;
$\overline{G}$	mean genetic effect for economic value of a population;
$\widetilde{G}$	genetic gain for economic value;
$\widetilde{G}_o$	the highest genetic gain for economic value given all optimisation constraints;
$\widetilde{G}_z$	the highest genetic gain for economic value obtained through stages 1,2,3,..., z;
U_{z,P_0}	upper bound for selection stage z given the starting population of genotypes P_0;
A_z	alternative combination of parameters to be used at stage z;
$\overline{B}$	budget available to perform selection system;
C_z	cost of stage z;
$\overline{C}_z$	cost of stages 1,2,3,..., z;
N	the total number of nodes to be branched;
$\widehat{M}_z$	planting material required to plant stage z;
$\breve{M}_z$	planting material available for stage z;
P_z	population of genotypes selected through stages 1,2,3,..., z;

Statement of Sources

Declaration

I declare that this thesis is my own work and has not been submitted in any form for another degree or diploma at any university or other institution of tertiary education. Information derived from the published or unpublished work of others has been acknowledged in the text and a list of references is given.

16.08.2005

Vanja Calija-Zoppolato

Date

Publications Arising From the Thesis

Referred International Journal Publications

V. Calija, A.J. Higgins, P.A. Jackson, L.M. Bielig, and D. Coomans, An Operations Research Approach to the Problem of the Sugarcane Selection, Annals of Operations Research 108 (2001) 123-142.

Refereed Conference Papers

V. Calija, P. Jackson, A.J. Higgins, L. Bielig, and D. Coomans, Simulating and optimising sugarcane selection In: *Proceedings of the 15th National Conference of the Australian Society of Operations Research Inc. ASOR Queensland Branch and ORSJ Hokkaido Chapter Joint Workshop on Operations research from theory to real life: Contributed papers,* Ed. E. Kozan, School of Mathematical Sciences, Queensland University of Technology, Brisbane Vol. I (1999) 260-268

V. Calija, P. Jackson, A. Higgins and L. Bielig, Simulating and optimising selection systems, 11th Australian Plant Breeding Conference, Adelaide, SA, April 19-23, 1999

Chapter 1

Introduction

Selecting few candidates from a large starting population is a problem that is common to animal and plant breeding, as well as to chemical and pharmaceutical multi-stage screening. In all these situations, the starting population to which selection is applied is narrowed down over a number of stages, where at each stage some candidates are advanced to further stages while other are discarded from future selection.

In animal breeding for example, a one or two-stage selection is used to select a set of sires and a set of dams for possible mating (Dekkers, 1998). In this case, the proportion of animals to be selected is set and then selection is based on some index. In the pharmaceutical industry, in drug development, a large number of alternative compounds need to be screened prior to identifying those that deserve further attention.

Plant breeding, as exemplified in Figure 1.1 for sugarcane, is a process that generally aims to produce new cultivars (commercially used varieties) that are more profitable than predecessor cultivars. In sugarcane breeding programs, a number of varieties that possess desired characteristics, are selected as parents. They are crossed, producing a new generation and population of seedlings. Seedlings that belong to the same parental cross represent a family. These seedlings then enter the selection program. Selection can be based on individual (clonal) or family performance. In the case of individual selection, each individual genotype is assessed and accordingly selected or discarded. Therefore, all varieties produced from the annual crosses are the starting population from which plant breeders select varieties to enter the selection program. If family

selection is practiced, entire families are selected or discarded based on the average performance of the genotypes that constitute them. Individual seedlings are then usually selected from the selected families for further selection as individual genotypes in subsequent stages of selection.

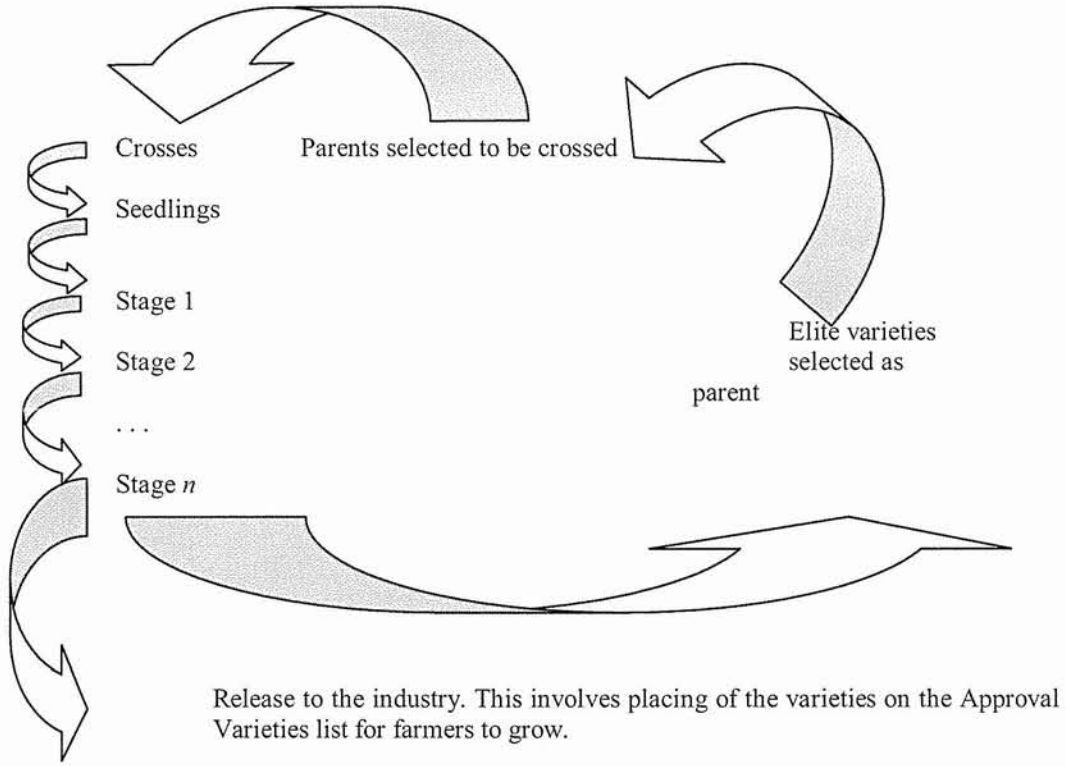

Figure 1.1: Diagram outlining the general structure of sugarcane breeding programs. Firstly, varieties with desired characteristics are crossed, and secondly, progeny seedling varieties from those crosses that possess desired properties are selected (stages $1,2,...,n$).

The selection system, described by stages $1,2,...,n$ in Figure 1.1, is the most expensive and lengthy part of sugarcane breeding and new cultivar development, taking ten to fifteen years. Berding and Bull (1997) presented data for forty-four Queensland cultivars and showed that the average time required from initial seedling planting to farm release was 14.8 years.

The aim of this thesis was to help design efficient selection practices and therefore improve breeding programs. The following section gives an introduction into the sugarcane breeding as well as the issues connected to designing selection systems.

1.1 Sugarcane breeding programs

In Australia, sugarcane is grown in a narrow coastal strip stretching over 2,100 kilometres from Mossman, north of Cairns in Queensland to Grafton, south of Lismore in New South Wales (Figure 1.2). The cane grows for 10 to 18 months before being harvested. Harvesting or crushing season is usually between June and December. The first sugarcane harvested is referred to as plant cane. After the first harvesting it is then allowed to regrow as a ratoon crop. Depending on the quality of the crop and the economics of replanting, three to five ratoon crops are grown before replanting.

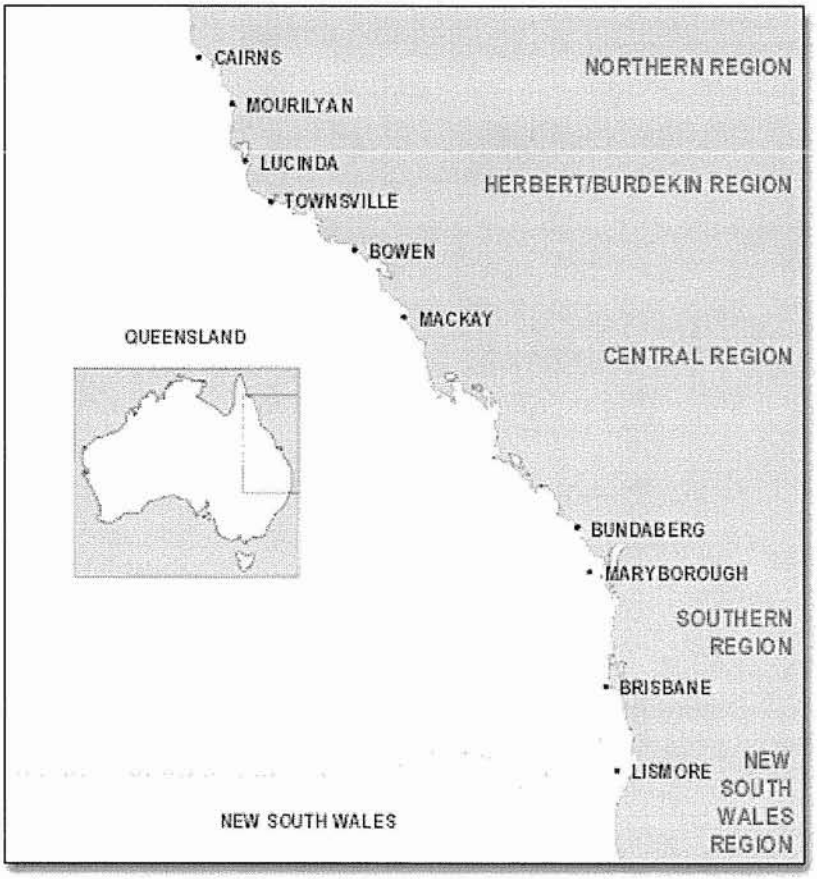

Figure 1.2: Map of the major Australian sugarcane growing regions accessed from the http://www.sri.org.au/sugarindustry1.html

Raw sugar is a major contributor to the Australian economy, valued at $2 billion annually. Approximately eighty percent of annual raw sugar product of Australia is exported, making Australia one of the world's largest sugar exporters. Within Australia, sugar ranks second to wheat in terms of export crop and second to beef in terms of farm production. Directly or indirectly, the Australian sugar industry employs approximately 41,000 people. There are approximately 7,200 family-owned cane farms in Australia, and twenty-nine self–contained raw sugar mills are annually crushing over 41 million tonnes of cane (http://www.sri.org.au/sugarindustry1.html).

Advanced farming practices and the development of new cane varieties meant that Australia maintained its international competitiveness up to 2004. Sugarcane breeding programs have been developing new and improved varieties for sugar industries around the world since the early 1900s. An ongoing varietal improvement has been a major technology driver of improved efficiency and economic sustainability in sugar industries.

The general objective of most sugarcane improvement programs is to develop new varieties, which are able to increase profits through increased cane yield [TCH (Tonnes Cane per Hectare)] and/or commercially extractable sucrose [termed CCS (Commercial Cane Sugar) in Australia]. Cane yield and CCS are both industry standard measurements taken to help determine economic value of sugarcane (Skinner *et al,* 1987). The two measurements may be used to calculate a selection index, which is a measurement used for ranking genotypes under evaluation for the purpose of selection. A selection index often used in the industry is sugar yield - TSH (Tonnes Sugar per Hectare) calculated as the product of CCS and cane yield. Note that throughout the thesis the two measurements will be referred to as cane yield and CCS, the terms most commonly used amongst plant breeders.

In sugarcane, a variety proposed for commercial use from breeding programs is usually identified after a long selection process starting with thousands of varieties that are narrowed down to the few regarded as best. The selected varieties are further evaluated and only if they appear to be commercially more valuable than existing cultivars, are they then released to the industry. Each sugarcane variety constitutes a unique

combination of genes and as such is often referred to as the genotype, the term adopted throughout this thesis.

Sugarcane genotypes are selected based on their observed that is phenotypic performance. Phenotype is a joint result of the effects of genes within a genotype and effects of the environment in which it grows (Falconer and Mackay, 1996). The genetic component of a phenotype, ie. its genetic value, cannot be observed directly and can only be inferred from observations on the phenotype for characters of interest. Furthermore, when commercially used, a genotype occupies an extensive area of land and exists in a pure stand, whereas, in selection trials many different genotypes are planted on limited land and they have to compete for available sunlight and nutrients. Thus, performance accurately measured in a selection trial may not correlate highly with performance in commercial fields due to differences in competitive ability in small plots. Selection trials with small plots are also subject to random error effects due to soil heterogeneity across the field that trials are conducted in, or other factors causing variation among measurements not related to genetic effects. Thus, the challenge plant breeders face is to design selection trials, where each genotype occupies limited areas of land, that would allow effective and efficient selection of genotypes that are superior to existing cultivars when planted in pure stand.

In selection trials sugarcane genotypes occupy only small plots. To get an insight into genotypes' performance in a pure stand, it is necessary to test them in different plot sizes and planting designs. Thus, a number of different selection trials are required. However, because testing the entire starting population of genotypes in experimental designs involving large areas would require substantial amounts of resources, at each selection trial the set of genotypes to be tested is narrowed down. Furthermore, separate breeding regions within the Australian sugarcane industry (Figure 1.2) were developed for climatically and environmentally different regions (Hogarth and Mullins, 1989).

1.1.1 An example of a sugarcane selection system: the Burdekin region

The selection systems in sugarcane breeding programs are typically made up of between three and six stages, with the best performing genotypes being released to industry at the end of the last stage. In the following example (Figure 1.3), the process used in the CSR breeding program targeting the Burdekin region (Figure 1.2) is outlined. It involves three stages and includes seven years of selection, three years of propagation and finally the year of cultivar release (Jackson *et al,* 1992). Only an approximate number of progenies at each stage are given. Note that this selection system involves a year of family selection, a standard practice in the Australian sugarcane industry.

In stage one, sixty seedlings per cross are planted at a particular site in four replicates per cross, so that fifteen seedlings from each cross constitute a single row plot (Figure 1.3). In the first year, the sixty families, with the highest sugar yield are selected. Cane was harvested and weighed to give cane yield. All stalks available are further crushed to give juice to assess CCS.

In the second year, thirty seedlings from within each of the sixty selected families are selected based on their visual appearance in ratoon crop. Stage two starts with planting of the 1800 genotypes selected in stage one (year 3, Figure 1.3). Two single plots of each genotype are planted at a site. The one hundred genotypes with the highest sugar yield are selected in plant crop. This is followed by one year of propagating (year 4, Figure 1.3) enough material to plant the stage 3 trials.

In stage three, the one hundred genotypes selected in the previous stage are planted into trials with four row plots and two replicates of each genotype per trial (Figure 1.3). Trials are located at four different sites in the region. Genotypes are evaluated in plant (year 5), first ratoon (year 6) and second ratoon (year 7) crops. The best performing genotypes are released to the industry after further testing with remaining promising genotypes being used in crosses in future years. Performance was assessed based on the weighed cane yield of all available stalks, whereas only two middle rows (30 stalks) were crushed to obtain an estimate of its CCS.

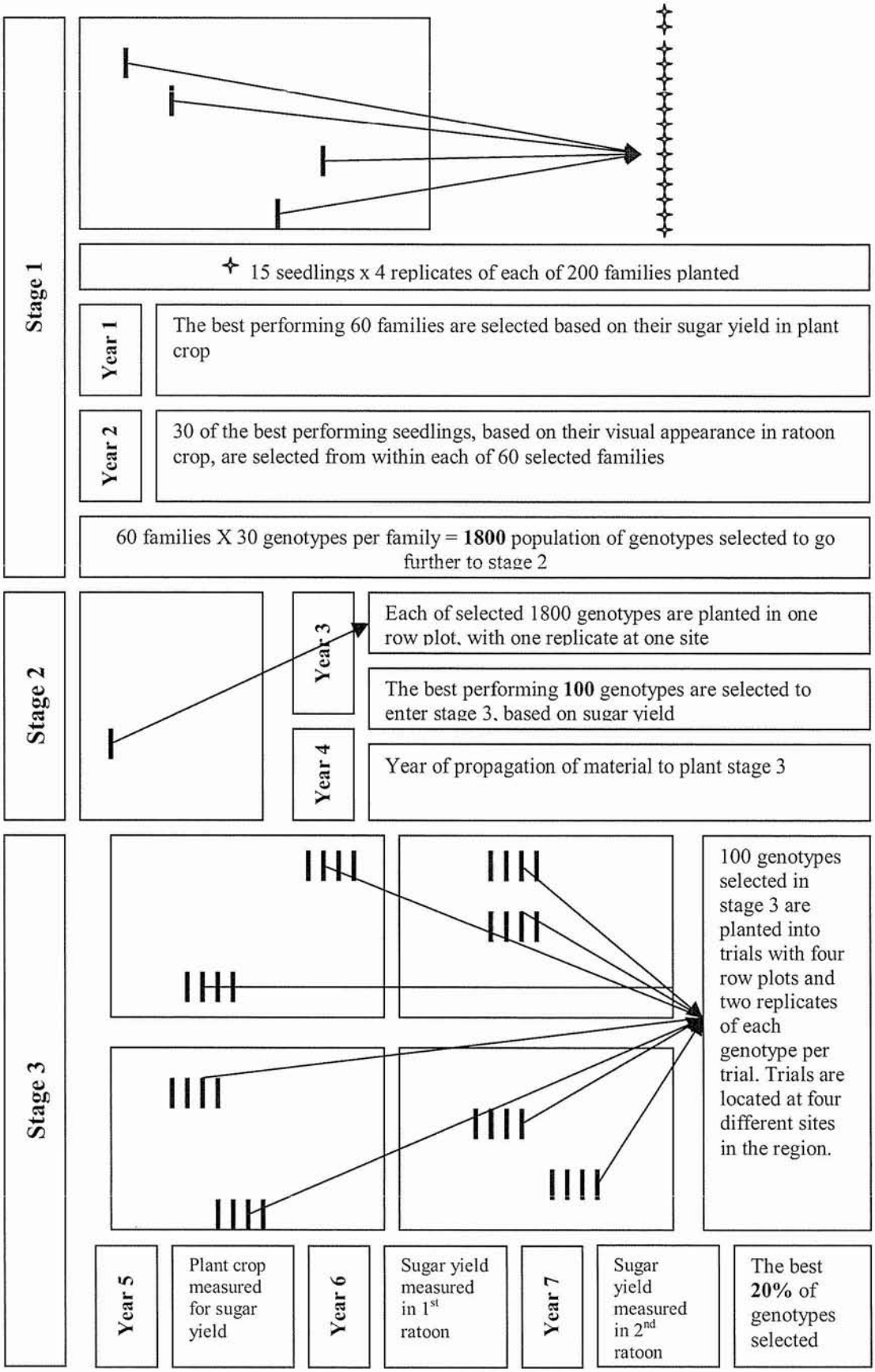

Figure 1.3: Diagram outlining the typical sugarcane selection system from the Burdekin breeding region

In addition to the seven years of selection detailed in Figure 1.3, the process of cultivar development may be prolonged by time required to make the initial crosses, to select further ratoon crops, to test for disease resistance, to assess performance at a number of sites which represent a range of environmental conditions, and to propagate sufficient planting material for distribution to growers.

1.1.2 Selection variables

Logically, the larger the starting population of genotypes, the greater the likelihood that it will contain elite genotypes that are better than current cultivars. The larger the number of rows per plot, the more that conditions for growing will resemble a pure stand situation and thus the more reliable that result would be expected to be. Also, a greater number of sites and replicates per site will reduce the impact of random experimental error effects on affecting estimates of genotype means in the trial. However, due to the expenses associated with testing genotypes at many sites, with many replicates per site and larger plot sizes, a larger starting population means that fewer field trial designs are practicable and less adequate field trial designs are possible given any particular budget. Therefore, for a given amount of resources available at any particular stage of selection, there is usually a trade-off between population size that can be screened and effectiveness of the selection processes.

Designing an effective selection system is a complex task because of the number of variables to be considered at each selection stage (McIntosh, 1935). To help mathematically formulate the sugarcane selection model, selection variables that are necessary to define a selection system are given below:

- family selection to be used in the stage 1, $b \in \{TRUE, FALSE\}$
- number of genotypes to start the selection in case of individual (clonal) selection or the number of genotypes per family in case of family selection, k
- number of families to start the selection in case of family selection, f
- number of stages, n
- plot size for each stage $z = 1,2,...,n$, p_z

- number of sites (locations) for each stage $z = 1,2,...,n$, s_z
- number of replications per site for each stage $z = 1,2,...,n$, r_z
- selection intensity that is the percentage of genotypes to be selected for each stage $z = 1,2,...,n$, t_z
- selection index for each stage $z = 1,2,...,n$, d_z.

1.2 Rationale for the study

Designing efficient selection systems has been the subject of constant research, and has aimed at identifying possible optimal selection practices. Although traditional breeding methods have done much to increase the productivity of the sugarcane industry, the pressure on breeding programs to produce efficient cultivars at reduced costs is ever increasing (Irvine, 1994). Skinner *et al* (1987) presented a review of research on separate aspects of sugarcane selection designs, and discussed difficulties in choosing the optimal combination of selection design variables (Section 1.1.2). Computer simulation experiments were advocated as a way to approach the problem of optimal selection design, because field trials that extend over large areas of land and stretch over many years are prohibitively expensive and time consuming. Skinner (1961) used mathematical modelling to represent competitive situations in different plot sizes and statistically estimate efficiency of alternative selection programs. Skinner (loc. cit.) emphasised that because many combinations of selection variables gave approximately similar results, to examine a breeding program thoroughly it is not sufficient to compare a few alternative systems in field trials. He further stressed that empirical experiments comparing systems that differ only in the value of some of their variables would often yield no significant improvements to existing methods and would be doomed to failure. Goldringer *et al* (1996) also argued that, because experimental comparison of selection practices is difficult and time consuming theoretical comparison should be used to obtain guidelines.

Jackson *et al* (1996) also concluded that many studies on sugarcane selection compare specific selection methods in particular situations, so are of limited usefulness in obtaining an overview into effective selection practices. They suggested a more effective approach is to develop an understanding of the key genetic and statistical

parameters affecting selection effectiveness, and to use simple models to explore and compare different selection design options.

The literature indicates which genetic parameters need to be estimated in order to allow the application of simple genetic models. These models may be used to predict genetic gain for selection, and present a broader perspective of selection options, which in turn allows for enhanced selection system optimisation. Focused field evaluation of options selection design could be used to validate and/or refine models and used to more confidently further assess options.

The value of mathematical involvement in designing optimal selection systems was highlighted by Gauch and Zobel (1996) and Martin and McBlain (1991). While studying different aspects of selection, and selection system design variables (Section 1.1.2), both Gauch and Zobel (1996) and Martin and McBlain (1991) emphasised that a change in any of the selection variables in any of its stages affects all other stages and thus selection overall. Cooper and Podlich (1999) argued the case for modelling breeding programs in order to identify optimal designs. Therefore, computer simulation may be one approach to deal with such complexity and permit a holistic examination of multistage selection systems.

Computer simulation has already been used to examine the effect of some variables on total genetic gain. These include the number of stages and the selection intensity (Young, 1976 and Finney, 1966). In recent years a number of studies (Hardwick and Stout, 2002, 2001, 1999; Wang and Leung, 1998; Yao and Venkatraman, 1998 and Yao *et al*, 1996) have been conducted that give a unified theoretical approach to the design of optimal multistage screening designs. However, these researchers did not apply nor develop a simulation model to perform a comprehensive investigation of selection systems, and concentrated only on the effect that selection intensity and the number of stages or cost constraints have on the total genetic gain. Furthermore, there is an uncertainty about the practical relevance and applicability of their studies to breeding programs, because some important factors such as the genotype by environment interaction and competition effects in small plots were neglected.

Mariotti and Faver (1983) showed that computer technologies could reproduce similar experimental situations to those observed in real populations. Although the study concentrated only on the percentage of recovery of the best genotypes, it nevertheless showed that by describing the phenotypic performance of a genotype as a linear model of its genetic and environment component, it is possible to reproduce experimental situations similar to those observed in real populations.

Optimising a multi-stage selection process is mathematically very complex as it involves many variables each with a different influence on the final result and thus they can not be joined together by simple mathematical relationships. This has forced scientists to observe only a portion of parameters at a time. Young (1976, 1974, and 1972), Finney (1966, 1958), Curnow (1961) and Robertson (1957) all studied general selection of few candidates from a large starting population. They all suggested that certain selection systems are optimal under a wide variety of conditions, providing that selection is performed on normally distributed populations. They optimised selection systems either with respect to only one variable or they considered a single stage of selection. However, the above studies did not address any particular plant or animal breeding program therefore involved many generalisations and simplifying assumptions.

The joint use of a simulation model of a real selection system, and mathematical involvement may be useful. However, it would appear that not nearly enough attention has been devoted to the problem of optimising an actual selection system from the mathematical viewpoint. By mathematically expressing the connection between selection variables and phenotypic values for CCS and cane yield it would be possible to predict the performance of selection and thus evaluate alternative selection systems.

There is an indication that many breeding programs might be inefficient because they are based at least partly on tradition (Mamet and Domaingue, 1999), or based around particular patterns of thinking among breeders about the optimal structure of selection systems affected by past practices. Expressing sugarcane selection systems mathematically and approaching the problem of selection system design using mathematics should enable a completely objective approach to its optimisation. Mathematical modelling can be used to define selection systems and capture their

complexities. For example, appropriate models could explain how the state of the entire system changes if one parameter changes, thus approaching the problem as an integrated system rather than a series of subcomponents.

Due to the differences and individuality of different sugarcane breeding programs, a mathematical approach would be applicable in agriculture, only by addressing one specific selection system at the time. The new field of research addressed in this thesis firstly used simulation modelling to represent a real sugarcane selection system, namely that used in the Burdekin region, in Queensland, Australia (Figure 1.2). Dynamic programming and branch-and-bound optimisation techniques are then applied to identify selection system designs that maximise selection gain. Optimisation is performed within necessary practical constraints such as available budget and availability of planting material per genotype in early stages.

1.3 Aim of the thesis

The aim of this research was to develop a tool for the optimisation of selection systems and to demonstrate how optimisation techniques can be used to enhance a selection system across all its stages. However, the problem of designing an efficient selection system consists of two parts: firstly, simulating and evaluating the performance of selection systems through changing different design parameters, and secondly, deciding on the combination of design variables for a selection system that will be most successful. This study was accordingly defined through these two sub-objectives. Steps taken to address the two sub-objectives were:

- to identify the main selection design variables and input parameters needed to approximately represent a selection system in 'real life',
- to gather information and analyse the data from the Burdekin region necessary to define the input parameters required,
- to develop and test a simulation model that represents the selection system and allows design variables to be varied and the impact of these variations on the selection results to be observed,

- to develop and implement procedures for identifying an optimally designed selection system,
- to apply the optimisation method to the selection system practiced in the Burdekin region,
- to validate the optimal selection system obtained.

1.4 Structure of the thesis

Chapter 2 provides information relating to the problem of designing selection systems in breeding programs and factors affecting the quality of selection. It focuses on sugarcane breeding programs in particular, and examines some issues affecting decisions about how selection systems are designed. It also documents some statistical analyses performed on data collected from the breeding programs in the Burdekin region, to estimate variance components of the effects underlying the expression of phenotypes.

Chapter 3 details the development of the Sugarcane Selection Simulation Model (SSSM). The SSSM utilised the information gathered on the Burdekin region to represent selection systems for the region. The development of the simulation model SSSM was monitored by plant-breeders from the Burdekin region to assess the applicability of the research to the real life situation. The possible applicability of the simulation model SSSM is widespread, from evaluating selection systems prior to field trials, and analysing stage performance separately, to predicting how the selection would be affected when some variance components that describe the targeted region change. It can be customised easily for other sugarcane breeding regions as well as for other crops. Chapter 3 also presents validation of the simulation model results. Furthermore, an examination of the sensitivity of the SSSM to variation in most of the key input parameters is also presented in this chapter. A sensitivity analysis was performed in order to identify those input parameters of the SSSM that most significantly impact upon the simulation results.

Within this study, the Algorithm for Sugarcane Selection System Optimisation (ASSSO) was applied to the SSSM (Figure 1.4) and is detailed in Chapter 4. This involved the development of a new algorithm that is a combination of dynamic

programming and branch-and-bound operations research techniques. Dynamic programming was used globally, optimising the overall selection system whereas branch-and-bound was used locally to identify optimal stage designs that define states for dynamic programming optimisation. Chapter 4 also examines other operations research techniques that may be applied to optimise sugarcane selection systems. However, the scope of this study was not to compare performances of different optimisation techniques when applied to the problem, but rather to demonstrate that mathematical methods can be used to optimise selection systems. This is significant because it involved developing a new application of operations research techniques.

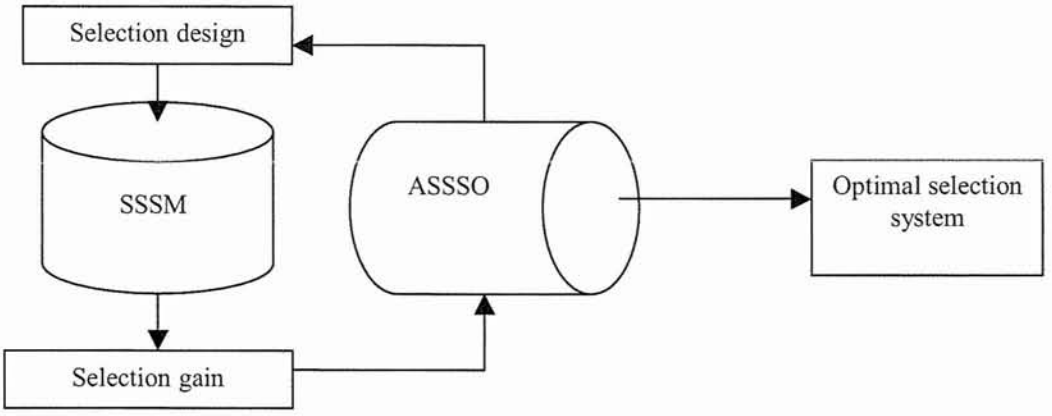

Figure 1.4: Diagram outlining the flow of information between the Sugarcane Selection Simulation Model (SSSM) and the Algorithm for Sugarcane Selection System Optimisation (ASSSO)

Chapter 4 also details the execution of the algorithm and the maximum number of nodes/iterations needed for the ASSSO to reach the optimal solution. In this chapter, the upper bound to the problem and constraints to the application are also presented.

Chapter 5 documents a range of selection systems resulting from application of the ASSSO to the Burdekin region. It further details the hierarchical clustering and regression tree methods used to analyse the optimal solutions as well as the examination of the convergence of the algorithm towards the optima.

Chapter 6 summarises and examines the significance of the optimisation study's findings and its importance to both operations research and plant breeding. This chapter recommends future research areas of priority.

Chapter 2

Selection systems in sugarcane breeding programs

2.1 Introduction

Based on a typical sugarcane selection program from the Burdekin region (Section 1.1.1) and in accordance with the aim of the thesis (Section 1.3), this chapter presents an analysis of (i) the main variables affecting phenotypic expression of cane yield and CCS and (ii) the way design parameters in the selection systems impact on phenotype and on gain from selection. Cane yield and CCS exhibit continuous variation and are thus quantitative or metric traits. To understand how selection design variables (Section 1.1.2) influence the two measured traits and thus permit a mathematical representation of the system, a linear model partitioning phenotype into component effects is discussed in Section 2.2. This model is then used as a basis for simulating selection and examining the impact of varying selection trial designs (Section 2.3).

Statistical analyses performed on the data gathered from the field trials from the Burdekin region are presented in Section 2.4. Estimates of the key variance components that contribute to the phenotypic variation are also given in this section.

2.2 Partition of phenotypic variance

The observed (phenotypic) value of a genotype for any trait is a result of the joint action of genes and the environmental conditions under which the genotype has grown. The phenotypic value, y_{ijk}, of any measure trait (CCS or cane yield in this case) of a genotype i, planted in environment j, in plot size k, may be partitioned further, and thus expressed as the linear model:

$$y_{ijk} = \mu + g_i + v_{ij} + c_{ik} + x_{ij} + e_{ijk} \qquad (2.1)$$

where μ is the grand mean of the trait observed; g_i is the genetic effect of genotype i; v_{ij} is the environmental effect of genotype i being planted in environment j; c_{ik} is the competition effect of genotype i being planted in plot size k; x_{ij} is genotype by environment interaction effect of genotype i being planted in environment j; and e_{ijk}, random error (Falconer and Mackay, 1996). These effects are not correlated except the genetic effect g_i and competition effect c_{ik}, which may be related.

Note that the linear model (2.1) as well as all the reasoning given in this section concerns the expression of the phenotypic values observed on genotypes in the case of selection trials only, where many different genotypes are planted alongside each other. On the other hand, when cultivars are grown in a pure stand, where the same genotype occupies extensive areas of land, competition effect is of no importance and thus the above linear model would change accordingly.

To understand the genetics of quantitative traits, such as CCS and cane yield it is necessary to analyse the partitioning of the phenotypic or observed variance into its components attributable to different causes. Different effects in the linear model (2.1) each have a variance, thus the phenotypic variance σ_P^2 can be partitioned into the genetic variance σ_g^2, the environmental variance σ_v^2, genotype by environment interaction variance σ_x^2, competition variance σ_c^2, the covariance between genotype

and competition $\mathrm{cov}(g,c)$, interaction between genotype and competition σ_τ^2 and error σ_e^2 (Falconer and Mackay, 1996):

$$\sigma_P^2 = \sigma_g^2 + \sigma_v^2 + \sigma_x^2 + \sigma_c^2 + 2\,\mathrm{cov}(g,c) + \sigma_\tau^2 + \sigma_e^2 \qquad (2.2)$$

Here the phenotypic variance σ_P^2 is the variance of the observed values for CCS or cane yield for the population of genotypes. The genetic variance σ_g^2 is the variance attributable to the genetic difference in the population. The magnitude of the genetic variance can not be estimated directly from observations on a single population but rather it requires a specifically designed experiment in which all other components of the above linear model are partitioned.

The environmental variance σ_v^2 is the portion of the phenotypic variance that is due solely to the difference in the environment in which genotypes are grown, where different environments are defined by different climatic conditions and soil types. However, generally all genotypes under comparison in a selection system are grown in the same selection trials, so that the environmental variance σ_v^2 does not affect changes in relative performance of genotypes. Thus, for the purposes of ranking and selecting this variance can be omitted from the above partitioning.

However, each trial is performed in a different environment, thus environmental variance is still present within the partitioning (2.2) through the presence of the interaction between genotype and environment σ_x^2. This interaction indicates that specific differences in environment has different effects on different genotypes (Falconer and Mackay, 1996). The way in which the environment impacts does change from genotype to genotype, thus the interaction between genotypes and environment σ_x^2 is an important component of the phenotypic variance σ_P^2 (Rattey and Kimbeng, 2001; McRae and Jackson, 1995; Mirzawan *et al*, 1993; Bull, 1992; and Jackson and Hogarth, 1992).

Apart from the components of phenotypic variance mentioned above, competition between neighbouring genotypes is another source of variation. The competitiveness is the ability of a genotype to compete with other neighbouring genotypes. The competition variance σ_c^2 is partly a component of the genetic and partly of the environmental nature. The competitive ability of a genotype depends on its genetic ability. It is of no importance for cultivars since they are grown in a pure stand. However, in selection trials where genotypes are planted in smaller plots, competition effect plays an important role (Jackson and McRae, 2001; 1998, Skinner *et al* 1987; and Skinner, 1961).

The covariance $\text{cov}(g,c)$ represents the correlation between genotype and competition. It is possible that 'better' performing genotypes are more competitive (Falconer and Mackay, 1996). Skinner *et al* (1987) reported that although there is some correlation between genetic and competitive effects $\rho_{g,c}$, it is not significant enough to indicate that 'better' genotypes are more competitive. For the Burdekin region, Jackson and McRae (2001) confirmed Skinner *et al* (1987) findings and reported that the correlation $\rho_{g,c}$ was highly variable in different environments. Nevertheless, the fact that the correlation between genotype and competition does exist suggests that covariance is a component in the partitioning of the phenotypic variance (2.2).

In selection trials, genotypes are planted in smaller plots with each their replicate being surrounded by a different set of neighbouring genotypes randomly assigned. Thus, the interaction between genotype and competition σ_τ^2 can be omitted from the partitioning (2.2), as the impact of the neighbourhood on the genetic performance is not systematic.

The partitioning of the phenotypic variance σ_P^2 (2.2) can therefore be re-expressed as the following partitioning that gave rise to the linear model (2.1) used throughout the thesis as the expression of the phenotypic value y_{ijk}:

$$\sigma_P^2 = \sigma_g^2 + \sigma_x^2 + \sigma_c^2 + 2\,\text{cov}(g,c) + \sigma_e^2 \qquad (2.3)$$

2.2.1 Heritability and the response to selection

The genetic variance σ_g^2 itself can further be partitioned to give an indication of genetic properties of the observed population and resemblance between genotypes when planted in different selection trials (Falconer and Mackay, 1996):

$$\sigma_g^2 = \sigma_A^2 + \sigma_D^2 + \sigma_I^2 \tag{2.4}$$

where σ_A^2 is additive genetic variance (breeding value), and the sum of σ_D^2 (dominance) and σ_I^2 (epistasis) is the non-additive genetic variance.

The total genetic variation σ_g^2 is used to determine the degree of the genetic determination or broad sense heritability H^2 of a trait, as the proportion of the total phenotypic variance, which is passed between generations (parent to offspring) after sexual reproduction (Falconer and Mackay, 1996):

$$H^2 = \frac{\sigma_g^2}{\sigma_P^2} \tag{2.5}$$

The additive genetic variation σ_A^2 is used to determine the narrow sense heritability h^2 of a trait, as the proportion of the total phenotypic variance due to the additive genetic factors (Falconer and Mackay, 1996):

$$h^2 = \frac{\sigma_A^2}{\sigma_P^2} \tag{2.6}$$

The heritability is always in the [0,1] range, as σ_P^2 is always more than σ_g^2 or σ_A^2. The closer the heritability is to 1 the larger proportion of the phenotypic variance σ_P^2 is attributable to the genetic variance σ_g^2 in the case of the broad sense heritability and attributable to additive genetic variance σ_A^2 in the case of the narrow sense heritability.

Consequently the closer the heritability is to 1, the more important it is to choose a breeding program that utilises the additive genetic variability in the case of the narrow sense heritability and total genetic variability the broad sense heritability. In sugarcane, following production of seedlings from crossing, genotypes are propagated through clonal reproduction. Thus, within a selection system, genotypes are passed from one stage to another through clonal propagation, so that all genetic variance is transmitted. Thus, the broad sense heritability H^2 is that relevant for the individual genotype selection in sugarcane breeding programs.

In the following sections of the thesis, the heritability H^2 of a trait was used to demonstrate the effect that a change in any of the selection variables (Section 1.1.2) will have on the measured trait in:

$$R = H^2 S \tag{2.7}$$

where, R is the response to selection and represents the change in the population means before and after selection, S is the selection differential, which represents the difference in measurement between the average of the selected genotypes and the average of the population of genotypes to start with (Falconer and Mackay, 1996).

Thus, for those traits that have higher heritability (close to 1) the response to selection is also "high", ie it is the major part of the phenotypic variability of that trait. Of course that is also dependent on the selection differential. By contrast, those traits that have lower heritability values (close to 0) also have a "lower" response to selection and thus, in case of the high costs of measuring that trait is not advisable to select for it.

2.3 The relationship between measured traits and selection design variables

The selection design variables (Section 1.1.2) and way in which they affect the magnitude of each of the effects in the linear model (2.1) are examined below.

2.3.1 Individual selection versus family selection

Individual selection discards or selects individual genotypes based on their phenotypic performance, whereas family selection discards or selects entire families based on the average performance of the genotypes that constitute them.

Australian sugarcane breeders, unlike those in other countries, have been practicing family selection since the late 1980s. Hogarth (1971) suggested that family selection would be superior to individual seedling selection in early stages of selection. Hogarth and Mullins (1989) argued that combined with mechanical harvesting and weighing, family selection was labour efficient in early stages compared with individual clonal selection, as did Kimbeng and Cox, (2003), Kimbeng *et al*, (2000), McRae and Jackson (1995), Cox and Hogarth (1993a), Jackson *et al*, (1992) and Bull, (1992).

Jackson *et al* (1996) used simulation to compare performances of family selection, individual genotype selection, and a combination of the two. The superiority or otherwise of family selection over individual selection was dependent on assumptions made, such as the proportion of genetic variance contained between families versus that within families. Although different situations require different selection designs, in most cases a combination of family and individual selection was superior.

When family selection is practiced, the genetic variance σ_g^2 is partitioned further into between-family variance σ_f^2 and within-family variance σ_w^2 to give:

$$\sigma_P^2 = \sigma_f^2 + \sigma_w^2 + \sigma_c^2 + \sigma_x^2 + 2\,\mathrm{cov}(g,c) + \sigma_e^2 \qquad (2.8)$$

Jackson and McRae (1998), Skinner *et al.* (1987), and Brown *et al* (1968) all reported the proportion of the genetic variability attributable to the between families δ to be approximately thirty percent for both CCS and canc yield, giving:

$$\begin{aligned} \sigma_f^2 &= \delta \cdot \sigma_g^2 = 0.30 \cdot \sigma_g^2 \\ \sigma_w^2 &= (1-\delta) \cdot \sigma_g^2 = 0.70 \cdot \sigma_g^2 \end{aligned} \tag{2.9}$$

2.3.2 Plot size and the affect it has on the relationship between the genetic effect and the competition effect

The genetic effect g_i represents the value of a genotype when planted in a pure stand across the targeted environments. However, in selection trials, unlike pure stands, genotypes occupy comparatively smaller plot sizes such as one or two rows that are 5 to 10 meters long. All harvested material is used to weigh TCH, with only five to fifteen stalks crushed to give the CCS estimate. In these situations, genotypes have to compete with neighbouring genotypes for water, nutrients and sunlight and this gives rise to competition effects c_{ik}. To accurately rate the relative values of any set of genotypes in a pure stand they need to be tested in a larger plot size such as a four-row plot.

By reducing the competition effects, larger plot sizes are eliminating the variance of competition effects and thus are allowing better prediction of the performance of genotypes in a pure stand. The competition effect c_{ik}, which is of no importance for cultivars since they are grown in pure stand, therefore can play an important role in selection trials. Competition effects have been found to have a greater effect on cane yield than on CCS (Jackson and McRae, 2001, 1998; McRae and Jackson, 1998; and Skinner, 1961). Jackson and McRae (1998) and Skinner *et al* (1987) reported that in the one row plot, variance due to competition effects for cane yield was greater than genetic effects. Jackson and McRae (2001, 1998) and McRae and Jackson (1998) suggested that in small plots, selection that emphasises CCS would be more effective than selecting for sugar yield or economic value, both of which give a strong weighting to both cane yield as well as CCS. They also suggested that the impact that competition has on phenotypic value should be minimised as soon as possible in selection systems. That is, multiple

row plots should be used in selection trials as soon as there is enough propagation material to do so. However, this clearly involves tradeoffs with number of genotypes or sites which could be used, given finite resources for conducting a selection system.

The effect that different plot sizes have on selection is indicated with the formula (2.3). Jackson and McRae (2001) predicted that selection based on sugar yield using 1-row plots would result in 56% gain in the relative economic gain whereas using 2-row plots in 68% gain. Large plots reduce the variance caused by competition, and because measurements such as cane yield are more precise in large plots, error variance is also reduced. A consequence of the reduction in error and competition variances is that a larger proportion of phenotypic variance will be due to genetic variance and hence H^2 will increase. Therefore, selection will be more effective as described in the response prediction formula 2.7.

The objective of selection is to select genotypes that will have high CCS and cane yield in a pure stand, both of which are traits that can not be measured directly in selection trials. Thus, as suggested by Jackson and McRae (2001) the selection in small plots can be regarded as an indirect selection (Falconer and Mackay, 1996), which utilises a correlated trait. In this case, the traits under selection, ie. cane yield and CCS values in small plots, are used to improve the target trait, ie. performance in a pure stand.

2.3.3 Number of sites and its impact on the interaction between genotype and environment

Genotype by environment (GE) interaction may be defined as the failure of genotypes to perform the same, relative to each other, when grown in different environments. Variance due to GE interaction is often significant in breeding programs, and it complicates selection because performance measured in any one or a few environments may not adequately represent performance across the wider set of environments targeted in the breeding program. The GE interaction effects may be subdivided into genotype by site (genotype by location, GL) interaction; genotype by years (GY) interaction; genotype by crop (GC) interaction; and the second order interaction of: genotype by location by year (GLY), genotype by crop by year (GCY), genotype by crop by site

(GCL); as well as the fourth order interaction (GCLY). Because sugarcane genotypes are grown at a site first as plant crops and then also as ratoon crops, the GC interaction is present. When both plant and ratoon crops are grown consecutively at a site over several years, crop and year effects are confounded, and referred to as crop-year effects.

In selection trials, when a single crop is grown in a year, the impacts of GY and GC interaction is small and thus the dominant interaction is the GL. In that case, the number of sites s used in that selection trial reduces the magnitude of the GE interaction variability from σ_x^2 to σ_x^2/s. In such situations therefore, increasing the number of sites, years and crops sampled would respectively reduce the magnitudes of GL, GY and GC interactions. Subsequently, to gain an accurate assessment of the genetic potential of genotypes, testing across a number of sites and/or years is required. This however, increases the size and complexity of selection programs.

The impact of GE interaction differs between countries as well as between regions within each country. Within Australia, the importance of GE interactions and its impact on selection was reported for the Southern region (Figure 1.2) (Mirzawan *et al*, 1993, and Bull, 1992), the Bundaberg region (Figure 1.2) (Bull *et al*, 1992; Hogarth and Bull, 1990; and Bull and Hogarth, 1990), the Herbert region (Figure 1.2) (McRae and Jackson, 1995; and Jackson and Hogarth, 1992), and the Burdekin region (Figure 1.2) (Kimbeng *et al*, 2000). Regardless of the region in which they were conducted, all these studies indicated that GC-Y interaction was small compared with the GL interaction. This therefore suggests that choice and number of sites is an important consideration in selection systems. In the Burdekin region however, in contrast to these findings, the GL interaction was found to be of relatively minor importance due to the similarity of soil types on cane land used in the region (Jackson *et al,* 1995 and McRae and Jackson, 1995). Family by environment interaction was also found to be relatively unimportant in the Burdekin region (McRae and Jackson, 1995). Only GC-Y interaction was reported to be a significant source of variation in advanced stage trials for the Burdekin region (Rattey and Kimbeng, 2001). Consequently, the first two stages of selection in the Burdekin region are currently conducted at one site only.

2.3.3.1 Ratooning performance

When commercially used, sugarcane is normally ratooned three to six times. Ratooning performance is therefore an important criterion for selecting new varieties. Currently, genotypes are selected according to their plant crop performance throughout selection, with the sole basis for selection in the early stages being plant crop performance, and ratooning performance being an important selection factor at later stages.

It has been suggested that only at later stages of selection, when competition is eliminated by planting genotypes in larger plot sizes, was it worthwhile testing genotypes for their ratooning abilities (Jackson and Hogarth, 1992, and Skinner *et al,* 1987). In a study on materials (trial designs and genotypes) representing early or middle stages of selection, Jackson and Hogarth (1992) showed that the marginal gains from selecting in ratoon crops in such stages would not justify the extra time and cost associated with this activity, as selection based on plant crop alone gave similar results to selection based on data from first and second ratoon crop. Jackson and McRae (2001), Mirzawan *et al* (1993), and Jackson (1992) found a high correlation between plant and ratoon crop in early stage trials, suggesting that there is little advantage in selectin genotypes fro their ratoon crop performance. This result was due to the large size of genetic variation compared with GC-Y interaction and the presence of a high correlation in error effects across year-crop within trials. Thus, one way to shorten a selection program is by putting less emphasis on ratoon performance (Mamet and Domaingue, 1999; Miller, 1994; and Cox and Hogarth, 1993b).

To grow a ratoon crop, the number of selection variables (Section 1.1.2) that are required to define it, is narrowed down to only two: selection index and intensity, since other parameters (eg. plot size, replicate number) have already been set at planting. Furthermore, since the whole fields previously planted are ratooned, no genotype needs to be discarded at any level of ratooning, leaving the decision on which genotype to select for after the specified number of ratoon crop have been grown ie three to five years. This in turn creates a third selection variable to be considered at each stage: the number of ratoon crops to be grown. Bearing in mind that the number of stages is an unknown variable, each stage needs thus to be further tested for ratooning, which would undoubtedly make the representation of the selection system mathematically even more

complex. The genotypes should thus, only be tested for the ratooning abilities at later stages of selection. In order to reduce the complexity of the simulation model and subsequently optimisation process, only plant crops were considered here. Consequently, this thesis focuses on the portion of the GE interaction that is the GL interaction. Options involving whether or not to grow ratoon trials in early stages of selection were not considered.

2.3.4 The number of replicates and its effect on error

The variability unassigned to any specific source of variation is assigned to error. Error variance is estimated from deviations between replicates of the same genotype within trials.

The total number of replicates, which is the number of sites s by the number of replicates per site r, reduces the error variance, as more replicates will allow more precision in estimating CCS and cane yield. Thus, the error variance is reduced from σ_e^2 to $\sigma_e^2/(s \cdot r)$ (Milliken and Johnson, 1984). Consequently, the heritability increases, and according to the formula (2.7), the response to selection increases too. But as the number of replicates and sites increases so does the cost of the trial.

2.3.5 Selection index

Breeders use some form of selection index throughout selection. Sugar yield or TSH is sometimes used as the selection criterion in sugarcane breeding programs.

$$TSH = TCH \cdot CCS \tag{2.10}$$

Another selection criterion frequently used by the Australian Bureau of Sugarcane Experimental Station (BSES) is NMG (Net Merit Grade), an economic index calculated using the following formula:

$$NMG = [(\frac{TSH}{TSH_{cul}} + (CCS - CCS_{cul}) \cdot 0.03] \cdot NMG_{cul} \tag{2.11}$$

Here, $TCH_{cul}, CCS_{cul}, NMG_{cul}$ are mean values of cane yield, CCS and NMG, respectively for standard cultivars used in the region.

Visual evaluation is a selection criterion normally practiced when genotypes are selected from within families planted in the first stage of selection. Since there is only a single seedling available for each genotype, it is considered a reliable estimator of the performance of the genotype. Furthermore, there is a high cost related to CCS sampling or cane yield measurement for a single seedling. Thus, rather than measuring genotypes, they are selected based on their visual appearance in comparison to other genotypes in the field.

Currently, sugarcane selection is based on one selection index (sugar yield or NMG), with visual selection being practiced when selecting genotypes from within families. However, it was suggested by Cuenya and Mariotti (1994) and Skinner *et al* (1987) that each stage should be addressed separately in order to determine the selection index to be used. Cuenya and Mariotti (1994) and Skinner *et al.* (1987) both argued that because CCS and cane yield are affected differently by the environment in which the genotype has grown, and accordingly different adjusting or weighing should be given to CCS and cane yield. Jackson and McRae (2001) added that greater emphasis should be given to CCS in the early stages of selection, since in early stages when small plots are used, cane yield is grossly affected by competition among neighbouring genotypes. Kimbeng *et al* (2001a,b; 2000) have found that elite genotypes were found from within families that score high CCS values and low cane yield or vice versa, thus putting further emphasis on the importance of changing the attitude towards the choice of the selection index used at each stage.

2.3.6 Selection intensity

Selection intensity is the proportion of the original population of genotypes that belongs to the selected group. On one hand, it is expensive to discard inferior genotypes late in selection, and on the other hand, selection intensity cannot be too severe in the early part of selection as it may result in discarding promising genotypes because data obtained in early stages is unreliable for reasons discussed above.

According to the response formula (2.7), selection differential S directly affects the response of selection, and because selection intensity impacts on the magnitude of the selection differential it also impacts on the response formula. However, there are limits to improving the response to selection through increasing selection intensity. Skinner (1969) put the greatest emphasis on the sequence of selection intensities used throughout selection for selecting superior genotypes. The second place in the importance was put on the starting population of genotypes, which should be of sufficient quality to ensure the presence of superior genotypes at the end of the selection process.

Kimbeng *et al* (2000) and Cox and Hogarth (1993a) both found that selecting at most the top 30-40% of families at stage one should be targeted for individual genotype selection. However, Kimbeng *et al* (2001a,b; 2000) found that even medium scoring families could produce elite genotypes. Kimbeng *et al* (2001b) have carried out a simulation study to identify optimum selection intensities for both family and individual genotypes selection for the central Queensland region. They concluded that selecting 40, 30 25 and 10% genotypes from within the best 10, 20, 30 and 40% families would results in the optimum returns.

2.3.7 The size of the starting population

In the case of clonal selection, the size of the starting population is determined by a single variable: the number of genotypes k entering selection system. In the case of family selection, on the other hand, it is determined by the number of families f as well as by the number of genotypes k per family. In practice, the size of the starting population is limited by two factors: the resources available and the selection intensities to be used at each stage of selection.

2.3.8 The number of stages

Due to the presence of the genotype by environment interaction, each genotype needs to be tested in a sufficient number of sites before it can be recommended for commercial use. However, at the beginning of selection there is a single plant of each genotype available. By planting larger plot sizes and more replicates with each consequent stage, propagation of planting material and selection are simultaneously happening.

Thus, the number of stages that are necessary for an efficient selection system must be at least as many that will allow sufficient production of cane for planting material to allow for sufficient evaluation in precise trials in final stages. From this point of view, the number of stages n could not practically be less than three.

2.3.9 Phenotypic correlation and genetic correlation between traits

There are two main types of correlation between traits: genetic and phenotypic correlation. Phenotypic correlation measures the relationship between the observed values of two traits, while genetic correlation measures the association between genetic effects of two traits that can not be directly measured (Falconer and Mackay, 1996). Some genes may increase the value of both traits while others may increase the value of one trait and concurrently decrease the value of the other. If the genetic correlation between two traits is zero it indicates that selection for one trait will have little impact on changing another trait. If genetic correlation is negative or positive, selection for one

trait will impact on the other, in a way related to whether correlation was positive or negative.

Brown *et al* (1969) showed that the genetic correlation between CCS and cane yield is virtually zero, so that selection for one of those traits only will not have a great effect on selection for the other trait. Thus, selecting simultaneously for the two traits is likely to be successful. Jackson and McRae (2001) suggested that while selecting simultaneously for the two traits, depending on the plot size used in the trial, weighing of CCS and cane yield used in the selection index should change. The genetic correlation between cane yield in small plots were found law compared to the genetic correlation between CCS in small plots, indicating that when small plots are used, selection could be based solely on CCS (Jackson and McRae, 2001).

Researchers in India have reported on the other hand that the phenotypic correlation between CCS and cane yield to be significantly positive (Pillai and Ethirajan, 1993), indicating that taking genes and environment together, the cane yielding higher cane tonnage tends to be higher in sugar content.

2.3.10 Effects of plot size

The effects of different plot sizes on genotypic performance can be expressed in terms of the genetic correlation between plot sizes. Plant breeders use the genetic correlation between performances in different plot sizes and those in pure stand as an indicator of the presence of the competition effects. Thus, performance in small plots, highly affected by large competition effects, would have a low genetic correlation. Whereas, performance in large plots the genetic correlation would be small or absent. In order to mathematically define the effect selection variables (Section 1.1.2) have on variance parameters, the connection between the genetic correlation for different plot sizes and the competition variance is considered below.

The competition effect c_{ik} equals zero in a pure stand, and when a genotype is planted in a plot size its genetic effect g_i is enlarged by the competition effect c_{ik}. The magnitude of the competition effect c_{ik} of a genotype in a given plot size could thus be

estimated as the difference between the phenotype expressions y_{ijk} of the genotype when planted in a pure stand and that when planted in the particular plot size. The genetic correlation between a particular plot size and a pure stand could thus be written as $\rho_{g,g+c}$ ie it could be considered as the correlation between the genotype effect g_i and the sum of the genotype and competition effect $g_i + c_{ik}$. Statistical theory provides the following link between the genetic variance σ_g^2, genetic correlation between plot sizes $\rho_{g,g+c}$, correlation between genotype and competition $\rho_{g,c}$ and the competition variance σ_c^2.

Let the distributions A and B be:

$$A = \begin{bmatrix} g \\ c \end{bmatrix} \sim N(\mu, \Sigma_A) \text{ and } B = \begin{bmatrix} g \\ g+c \end{bmatrix} \sim N(\mu, \Sigma_B) \tag{2.12}$$

where:

$$\Sigma_A = \begin{bmatrix} \sigma_g^2 & \text{cov}(g,c) \\ \text{cov}(g,c) & \sigma_c^2 \end{bmatrix} \text{ and } \Sigma_B = \begin{bmatrix} \sigma_g^2 & \text{cov}(g,g+c) \\ \text{cov}(g,g+c) & \sigma_{g+c}^2 \end{bmatrix} \tag{2.13}$$

If a transformation $L : A \rightarrow B$ exists such that $LA = B$, then $\Sigma_B = L\Sigma_A L'$ (Walpole, 1990). The matrix $L = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$ is such that $LA = B$, therefore:

$$\Sigma_B = \begin{bmatrix} \sigma_g^2 & \sigma_g^2 + \text{cov}(g,c) \\ \sigma_g^2 + \text{cov}(g,c) & \sigma_g^2 + \sigma_c^2 + 2\,\text{cov}(g,c) \end{bmatrix} \tag{2.14}$$

Hence,

$$\text{cov}(g, g+c) = \sigma_g^2 + \text{cov}(g,c) \text{ and } \sigma_{g,g+c}^2 = \sigma_g^2 + \sigma_c^2 + 2\,\text{cov}(g,c) \tag{2.15}$$

Using the correlation formula

$$\rho_{x,y} = \frac{\mathrm{cov}(x,y)}{\sqrt{\sigma_x^2}\sqrt{\sigma_y^2}} \tag{2.16}$$

$$\rho_{g,g+c} = \frac{\sigma_g^2 + \rho_{g,c}\sqrt{\sigma_g^2}\sqrt{\sigma_c^2}}{\sqrt{\sigma_g^2}\sqrt{\sigma_g^2 + \sigma_c^2 + 2\rho_{g,c}\sqrt{\sigma_g^2}\sqrt{\sigma_c^2}}} \tag{2.17}$$

Solving the above for σ_c^2 gives the formula for the competition variance:

$$\sigma_c^2 = \left(\frac{-2\rho_{g,c}\sqrt{(\sigma_g^2)^3}(\rho_{g,g+c}^2 - 1) \pm \sqrt{\left(2\rho_{g,c}\sqrt{(\sigma_g^2)^3}(\rho_{g,g+c}^2 - 1)\right)^2 - 4(\sigma_g^2)^3(\rho_{g,g+c}^2 - \rho_{g,c}^2)(\rho_{g,g+c}^2 - 1)}}{2\sigma_g^2(\rho_{g,g+c}^2 - \rho_{g,c}^2)} \right)^2 \tag{2.18}$$

From the discussion above, the relevance of any simulation model to breeding programs depends on the use of realistic statistical parameters when generating the effects in the equation 2.1. The following section summarises statistical analysis of data gathered from field experiments in the Burdekin region. This permitted an estimation of the variance components of the four effects as described by the equation 2.1, an investigation of the relationships between these four effects and an estimation of relationships between CCS and cane yield.

2.4 Estimation of statistical parameters

Two sets of data were obtained from selection trials in the Burdekin region. The first set of data A, was provided by the Commonwealth Scientific and Industrial Research Organisation (CSIRO), Davies Laboratory, Townsville, Queensland, Australia. This data set, that resulted from a collaborative research project between the privately owned sugar research company, CSR, the Bureau of Sugar Experiment Station (BSES), CSIRO and Sugar Research and Development Corporation (SRDC), was obtained from unselected populations of genotypes planted in three different plot sizes. The analyses detailed in Section 2.4.1 and 2.4.2 aimed to estimate:

- competition variance σ_c^2 (Section 2.3.2),
- genetic variance σ_g^2 in both unselected and selected populations (Section 2.3.2),
- error variance σ_e^2 (Section 2.3.4),
- genetic correlation between pure stand and different plot sizes $\rho_{g,g+c}$ (Section 2.3.10),
- the correlation between genotype and competition $\rho_{g,c}$ in different plot sizes and subsequently determine the competitive effect in different plot sizes (Section 2.3.2).

The second, data set B, was provided by BSES, Brisbane, Queensland, Australia. This data set was obtained from final stage selection trials conducted in the Burdekin region between 1995 and 1997. These data were used primarily to estimate the genotype x site (GL) variance component (Section 2.3.3) and subsequently, the $\hat{\sigma}_x^2/\hat{\sigma}_g^2$ ratio. It was also used to estimate expected error and genetic variances obtained during advanced stage trials of the kind conducted by BSES.

Table 2.1 summarises all the parameters that were estimated from the field experiment data. All of the estimates were used to parameterise or validate the simulation model. All statistical analyses were performed on the plant crop data. As explained in Section 2.3.3.1, in order to simplify the development of the simulation model and because ratooning is normally practiced at later stages of selection only, ratooning was not included in the simulation model SSSM and thus results from analysis of the ratoon data

are not presented here. All analyses were performed using the statistical package SPSS® 10.0.

Table 2.1

Summary of parameters estimated on the basis of: data set A (parameters estimated that depend on plot size) and the data set B (those that do not depend on plot size)

Parameter estimated	Symbol	Data Set
Genetic variance	σ_g^2	A, B
Genotype by environment interaction variance	σ_x^2	B
Error variance for different plot sizes	σ_e^2	A
Genetic correlation between pure stand and different plot sizes	$\rho_{g,g+c}$	A
Correlation between genotype and competition in different plot sizes	$\rho_{g,c}$	A

2.4.1 Experimental design, analysis and summary of results for the plot size experiment – data set A

Data set A was a part of the experimental data analysed by Jackson and McRae (2001). The data re-analysed for this study relates to the plant crop data (Section 2.3.3.1), since only plant crops were used in the development of the simulation model.

In a field experiment, fifty unselected seedlings were taken at random and planted in three plot sizes in order to test the effect plot sizes had on observed values of genotypes. The population of seedlings used in the experiment derived from sugarcane breeding programs conducted by CSR and BSES in the Burdekin Region. Three genotypes were taken at random from eleven randomly selected crosses from CSR breeding program, and five randomly selected crosses from BSES breeding program. In addition, two Hawaiian genotypes, previously not evaluated in this region, were used.

As shown in Figure 2.1, plot sizes were designated as small – one-row plot 10 meters long, medium – two-row plot 10 meters long and large – six-row plots 20 meters long. The trial was established in two blocks, with each plot size being represented in each

block (I and II). To give more precise estimates of the performance of genotypes in small plots, genotypes were replicated in small plots twice within each block, while medium and large plots were represented only once in each block. Thus, the analysis was performed twice for each small plot size replicate. The genotypes were randomly allocated to plots in each block. CCS and cane yield were measured in the plant crop and first ratoon crop. In the small and medium plots all material available was used to measure CCS and cane yield while in the large plot only the middle two rows were used, reducing the impact of competition effects on the measurements.

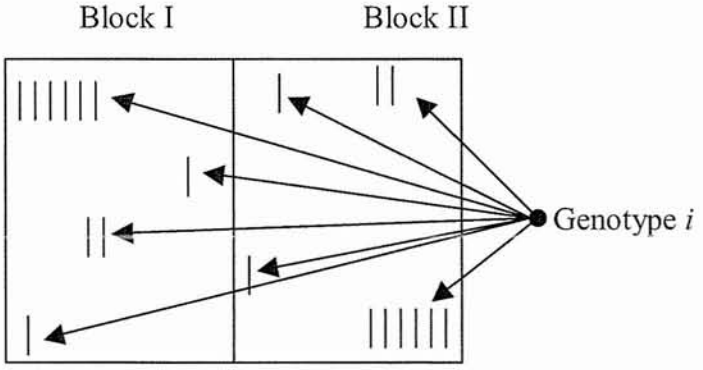

Figure 2.1: Outline of the experimental design used in the field trial that produced the data set A. Within each block, genotype *i* was planted once in six row and two row plots and twice in a one row plots. These plots were represented with one, two and six vertical lines

The crossed experimental design B x G, with G (genotype) being a random and B (block) being a fixed factor, can be represented with the following linear model:

$$y_{i_l r} = \mu + \beta_r + g_{i_l} + e_{i_l r} \tag{2.19}$$

where:

$y_{i_l r}$ - observed cane yield or CCS of the i_l^{th} genotype i being planted in the l^{th} plot size and the r^{th} block;

μ - sample mean;

β_r - effect of the r^{th} block, $j = 1,2$;

g_{i_l} - effect of the i_l^{th} genotype i being planted in the l^{th} plot size, $i = 1,2,3,...,k$, k the number of genotypes in the trial;

$e_{i_1 r}$ - interaction between the block effect β_r and the genotype effect g_{i_1} (error).

Variance components for genotype were estimated with the following expected mean squares:

Source of variation	Expected mean square
B	Not a random effect
G	$\sigma_e^2 + 2\sigma_g^2$
Error	σ_e^2

Multivariate analysis of variance (MANOVA) was used to provide the correlation estimates (Johnson and Wichern, 1982). An example of how MANOVA outputs were used to calculate correlation are given in Appendix A. The analyses can be represented by the linear system:

$$\begin{pmatrix} y_{i_1} \\ y_{i_2} \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} + \begin{pmatrix} g_{i_1} \\ g_{i_2} \end{pmatrix} + \begin{pmatrix} e_{i_1} \\ e_{i_2} \end{pmatrix} \tag{2.20}$$

y_{i_1}, y_{i_2} - observed expression;

μ_1, μ_2 - population means;

g_{i_1}, g_{i_2} - genetic effect;

e_{i_1}, e_{i_2} - error effect for a trait of the i^{th} genotype planted in the plot size 1 and plot size 2 respectively; $\varepsilon \sim N(0, \Sigma)$, where

$$\Sigma = \begin{pmatrix} \sigma_{e_1}^2 & \text{cov}(e_1, e_2) \\ \text{cov}(e_1, e_2) & \sigma_{e_2}^2 \end{pmatrix} \tag{2.21}$$

The first analysis performed on the data set A was to estimate the genetic variance σ_g^2 and error variance components σ_e^2. Point and interval estimates (95% confidence

interval for variance) for the error variances in three plot sizes together with the corresponding genotype variances are given in the Table 2.2 for both CCS and cane yield.

The genotype variance estimate $\hat{\sigma}_g^2$ for cane yield, decreased from a mean of 1789.24 in the small plots to 531.63 in the large plot (Table 2.2). This indicates that genetic expression of genotype in smaller plot sizes is largely inflated by competition between neighbouring genotypes. However, $\hat{\sigma}_g^2$ for CCS was about 4 irrespective of the plot size (Table 2.2), indicating thus that CCS measurement in sugarcanes is less affected by the competition between neighbouring genotypes. This confirms the reports of Skinner (1961), and Jackson and McRae (2001, 1998) that competition has greater affect on cane yield than on CCS. Note that the only genotype variance estimate $\hat{\sigma}_g^2$ of interest for this study is that in large plots since it gives an indication of its true magnitude when it is unaffected by the competition.

Table 2.2

Summary of the point and 95% confidence interval (in parenthesis) estimates of the genotype variance σ_g^2 and error variance respectively, for CCS (percentage of sucrose in cane) and cane yield (TCH - tonnes of cane per hectare), in three plot sizes, the small plot size was repeated twice in the trial

	CCS		TCH	
	$\hat{\sigma}_g^2$	$\hat{\sigma}_e^2$	$\hat{\sigma}_g^2$	$\hat{\sigma}_e^2$
Small plot rep 1	4.26 (3.00,4.94)	2.89 (2.04,3.35)	1347.92 (949.63,1564.27)	1512.30 (1065.43,1775.03)
Small plot rep 2	4.00 (2.82,4.64)	3.95 (2.78,4.58)	2230.55 (1571.45,2588.57)	1109.29 (781.51,1287.34)
Small plot mean	4.13 (2.91,4.79)	3.42 (2.41,3.69)	1789.24 (1260.54,2076.43)	1310.80 (923.47,1521.19)
Medium plot	4.74 (3.34,5.50)	1.17 (0.82,1.36)	1057.20 (744.81,1226.89)	488.37 (344.06,566.76)
Large plot	4.66 (3.28,5.43)	0.70 (0.49,0.81)	531.63 (374.54,616.96)	342.66 (241.41,397.66)

In cane yield, error variance estimate $\hat{\sigma}_e^2$ decreased from a mean of 1310.80 for small plots to 342.66 in large plot (Table 2.2), whereas for CCS it decreased from an average 3.42 in small plots to 0.70 in large plot (Table 2.2). Therefore, in small plots the magnitude of the error variance estimates $\hat{\sigma}_e^2$ for cane yield is similar or even greater to that of the genetic variance $\hat{\sigma}_g^2$, which further shows how unreliable cane yield measurements are in small plots. On the other hand, the magnitude of the error variance

estimates $\hat{\sigma}_e^2$ for CCS is the half of the genetic variance $\hat{\sigma}_g^2$, confirming the findings of Jackson and McRae (1998, 2001) that when small plots are used selection should be relying more on CCS then cane yield.

Table 2.3 presents the point and interval estimates (95% confidence interval for correlation) of the genetic correlation, $\rho_{g,g+c}$ between the three plot sizes. The mean genetic correlation for CCS between the small and large plots of 0.88 and between the medium and large plots of 0.93, confirm that CCS in small plots may provide an effective prediction of CCS in pure stand (Jackson and McRae; 2001, 1998). By contrast, the mean of genetic correlations for cane yield between small and large plots of 0.52; and between medium and large plots of 0.61 show that cane yield is a less affective criterion for selection in small plots. Thus, selecting heavily for cane yield in small plots could lead to discarding many promising genotypes early in selection. Note that since the large plot size was assumed to represent a pure stand, the genetic correlation between the pure stand and large plot $\hat{\rho}_{g,g+c}$ was assumed to equal one.

Table 2.3

Summary of the point and 95% confidence interval estimates (in parenthesis) for the genetic correlation $\rho_{g,g+c}$ between small (two replicates), medium and large plots in CCS and cane yield (TCH)

		$\hat{\rho}_{g,g+c}$			
		Small plot rep 1	Small plot rep 2	Small plot mean	Medium plot
Large plot	CCS	0.86 (0.77,0.92)	0.89 (0.82,0.94)	0.88 (0.80,0.93)	0.93 (0.88,0.96)
Large plot	TCH	0.53 (0.30,0.69)	0.51 (0.28,0.69)	0.52 (0.29,0.69)	0.61 (0.41,0.76)

Table 2.4 presents point and interval estimates (95% confidence interval for correlation) of the correlation, $\rho_{g,c}$ between the genotypic value of a genotype and its competitiveness. Again, because a large plot was assumed to represent a pure stand, the correlation between genotype and competition effects $\hat{\rho}_{g,c}$ in large plots was not applicable.

The correlation between genetic value and the competition $\rho_{g,c}$ in cane yield is a not significant, being a mean of 0.07 for small plots and 0.12 for medium plots (Table 2.4).

Similarly, for CCS, the correlation $\rho_{g,c}$ is not significant for either plot sizes, for medium plots it is –0.16 and in small plots the mean is –0.36 (Table 2.4).

Table 2.4

Summary of the point and 95% confidence interval estimates (in parenthesis) of the correlation $\rho_{g,c}$ between genotype and competition for CCS and cane yield (TCH) in small (two replicates), medium, and large plots

	$\hat{\rho}_{g,c}$				
	Small plot rep 1	Small plot rep 2	Small plot mean	Medium plot	Average
CCS	-0.34 (-0.56,-0.08)	-0.38 (-0.59,-0.12)	-0.36 (-0.57,0.09)	-0.16 (-0.41,0.11)	-0.26
TCH	-0.12 (-0.38,0.16)	0.08 (-0.26,0.28)	0.07 (-0.26,0.28)	0.12 (0.16,038)	0.10

The analysis confirmed Pillai and Ethirajan (1993) findings that there is a significant positive phenotypic correlation $\rho_{CCS,TCH}$ between CCS and cane yield. However, the findings are confirmed only for small and medium plots having the mean value for the correlation $\rho_{CCS,TCH}$ of 0.46 and 0.36 respectively (Table 2.5). For the large plot size, the correlation $\rho_{CCS,TCH}$ was not significant at –0.05 (Table 2.5).

Table 2.5

Summary of the point and 95% confidence interval estimates (in parenthesis) of the phenotypic correlation between CCS and cane yield (TCH), $\rho_{CCS,TCH}$ in small, medium and large plots

$\hat{\rho}_{CCS,TCH}$				
Small plot rep 1	Small plot rep 2	Small plot mean	Medium plot	Large plot
0.33 (0.07,0.55)	0.58 (0.37,0.74)	0.46 (0.22,0.65)	0.36 (0.09,0.57)	-0.05 (-0.28,0.27)

2.4.2 Experimental design, analysis and summary of results for the advanced stage variety trials – data set B

Two series of final stage selection trials were conducted in each of three years, providing six data sets available for analysis. These were coded 1995-1 for the first trial series in 1995 and 1995-2 for the second trial series in 1995 and similarly for 1996 and 1997.

Within each series of trials, four sites (S) within the region were chosen. Sites were split into two blocks (B). Each genotype (G) selected to go into the final stage was planted in four row plots (10 meter long) and was represented once in each block and each site (Figure 2.2). There were 17 genotypes planted in 1995-1 trial, 12 in 1995-2, 65 in 1996-1, 94 in 1996-2, 48 in 1997-1, and 94 in 1997-2. The two middle rows of the crop data were used to measure CCS and cane yield, as they were not exposed to competition with neighbouring genotypes.

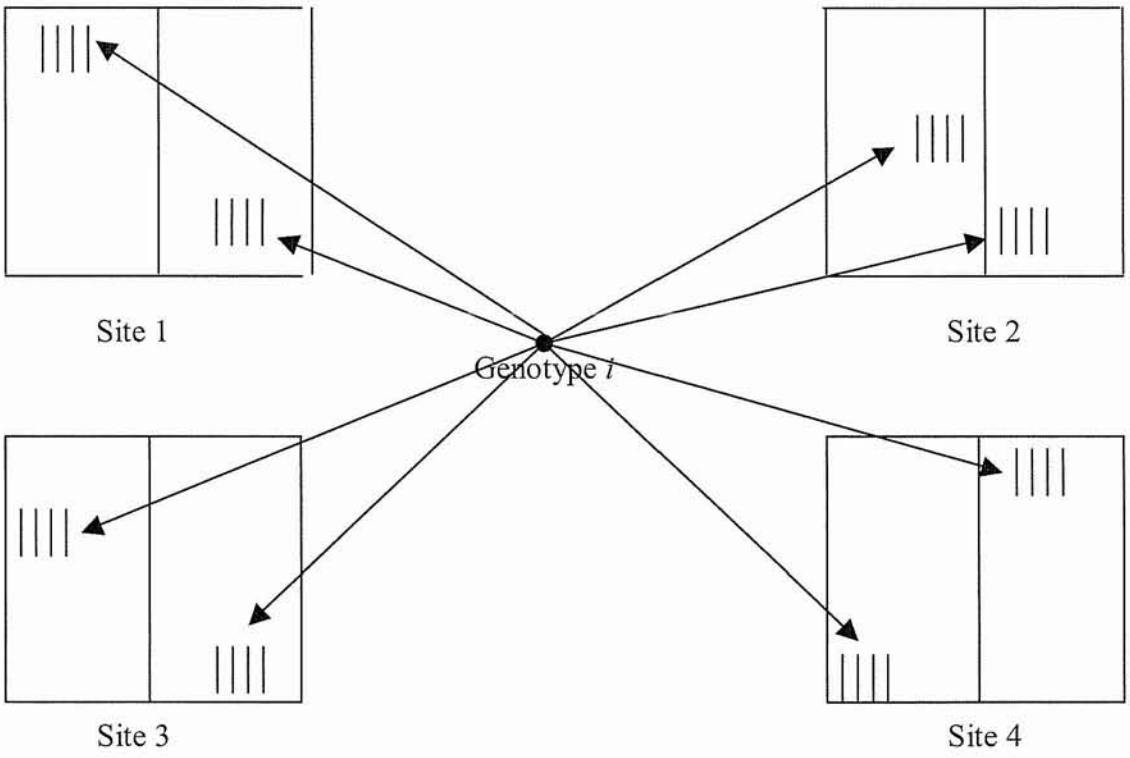

Figure 2.2: Outline of the planting design for the final selection stage the Burdekin region. Genotype *i* was planted in four row plots (represented by four vertical lines), at each of four sites, and within each of two blocks at each site

The experimental design $\begin{pmatrix} S \\ | \\ B \end{pmatrix} xG$ (blocks B, within sites S, crossed with genotypes G), where sites (S) and blocks (B) were fixed factors, and genotypes (G) were random factors, can be represented by the following linear model for each crop-year:

$$y_{irs} = \mu + \varsigma_s + \beta_{rs} + g_i + x_{is} + e_{irs} \qquad (2.22)$$

y_{irs} - observed trait cane yield or CCS of the i^{th} genotype in the s^{th} site and r^{th} block;

μ - sample mean;

g_i - effect of the i^{th} genotype, $j = 1,2,3,...,k$;

ς_s - effect of the s^{th} site, $s = 1,2,3,4$;

β_{rs} - effect of the r^{th} block within the s^{th} site, $r = 1,2$;

x_{is} - interaction between the i^{th} genotype and the s^{th} site;

e_{irs} - interaction between the i^{th} genotype and the r^{th} block within s^{th} site (error);

where k was the number of genotypes in the trial. Variance components for genotype (g), σ_g^2 and genotype by environment interaction σ_χ^2, were estimated with the following expected mean squares:

Source of variation	Expected mean square
B(S)	not a random effect
G	$\sigma_e^2 + 8\sigma_g^2$
G x S	$\sigma_e^2 + 2\sigma_x^2$
Error	σ_e^2

All the factors that contributed significantly to the variance were kept in the model while others were removed and considered as "noise". Such a correction allowed a more precise estimate of required variance parameters.

Table 2.6 summarises the point and interval estimates (95% confidence interval) for the genetic variance σ_g^2 in selected populations, genotype by environment interaction variance σ_x^2 and error variance σ_e^2. An estimate of the genetic variance σ_g^2 in selected populations was obtained as the average estimate of the genetic variances σ_g^2 from the six selection trials (Table 2.6) and was needed to verify the performance of the SSSM. In contrast to the genetic variances σ_g^2 of 4.66 and 531.63 for the unselected populations (Table 2.2), those for selected populations of 0.8 and 88.89 for CCS and cane yield respectively, were consistently lower as the result of the selection that eliminated inferior genetic material.

The average estimates for the genotype by environment interaction variance σ_x^2 were 0.46, 61.42 (Table 2.6) for CCS, cane yield respectively. The estimates for the genetic variance σ_g^2 were used together with the estimates for the genotype by environment interaction variance σ_x^2 for the six selection trials to get the estimates for the σ_x^2/σ_g^2 ratio of 0.58 ($\sigma_x^2/\sigma_g^2 = 0.46/0.80$ Table 2.6) and 0.69 ($\sigma_x^2/\sigma_g^2 = 61.42/88.89$ Table 2.6) for CCS and cane yield respectively. Being more than half the genetic variation, genotype by environment interaction σ_x^2 is a significant source of variation, confirming thus findings of Rattey and Kimbeng (2001), who estimated the relevant σ_x^2/σ_g^2 ratio to be equal to 0.30 and 0.26 respectively for CCS and cane yield in the Burdekin region.

Table 2.6

Summary of the point and 95% confidence interval estimates (in parenthesis) for genotype variance σ_g^2, genotype by environment interaction variance σ_x^2 and, error variance σ_e^2, for CCS and cane yield (TCH) in each of the six trials independently, together with the average values across trials

	CCS			TCH		
	$\hat{\sigma}_g^2$	σ_x^2	$\hat{\sigma}_e^2$	$\hat{\sigma}_g^2$	σ_x^2	$\hat{\sigma}_e^2$
1995-1	0.76 (0.42,1.02)	0.23 (0.16,0.27)	0.76 (0.56,0.87)	36.77 (20.40,49.39)	0.00 NA	460.65 (336.16,526.60)
1995-2	0.69 (0.35,1.01)	0.00 NA	1.43 (0.91,1.76)	106.92 (53.66,155.08)	0.00 NA	406.56 (257.87,500.29)
1996-1	0.87 (0.64,0.99)	0.83 (0.70,0.88)	1.15 (1.01,1.20)	105.42 (77.18,119.97)	88.88 (75.08,94.77)	214.57 (151.1,304.69)
1996-2	0.63 (0.28,0.69)	0.19 (0.74,0.21)	1.11 (0.97,1.16)	101.19 (79.32,111.48)	123.86 (99.95,134.65)	198.78 (173.70,208.89)
1997-1	0.77 (0.53,0.89)	0.77 (0.62,0.83)	1.78 (1.50,1.89)	78.68 (54.52,92.10)	0.00 NA	594.21 (499.18,634.73)
1997-2	1.06 (0.79,1.19)	0.72 (0.56,0.79)	2.31 (1.95,2.45)	104.38 (78.31,117.37)	155.76 (121.32,172.08)	374.78 (316.72,399.33)
Mean	**0.80**	**0.46**	**1.42**	**88.89**	**61.42**	**374.92**

The estimates for the error variance σ_e^2 were also obtained for the six data sets, and can be used compare it to the previous findings (Table 2.2). The average error variance estimate $\hat{\sigma}_e^2$ of 374.92 (Table 2.6) for cane yield was as expected; considering that eight replicates of four-row plots were used of each genotype; between the estimate for the

medium – two row plot of 488.37 (Table 2.2) and that for the large – six row plot of 342.66 (Table 2.2). On the other hand the average $\hat{\sigma}_e^2$ for CCS is somewhat larger at 1.42 (Table 2.6) than expected, between 0.7 for large plot and 1.17 for medium plot (Table 2.2). However, looking at the separate estimates of the error variance σ_e^2 for the six data sets, three (1995-1, 1996-1 and 1996-2) were as between the above two values.

To test the model assumptions, a graph of predicted values versus residuals was drawn for both CCS and cane yield for each trial and based on the linear model (2.1). All the graphs produced were similar to one another and thus as an illustration, only one of them is given below (Figure 2.3). Figure 2.3 depicts the predicted values versus the residuals for the CCS 1995-1 field trial. The equal spread of points around zero indicates that the residuals are independent of the mean, which shows the correctness of the model assumptions, which in turn indicates that the linear model (2.1) was appropriate for the data.

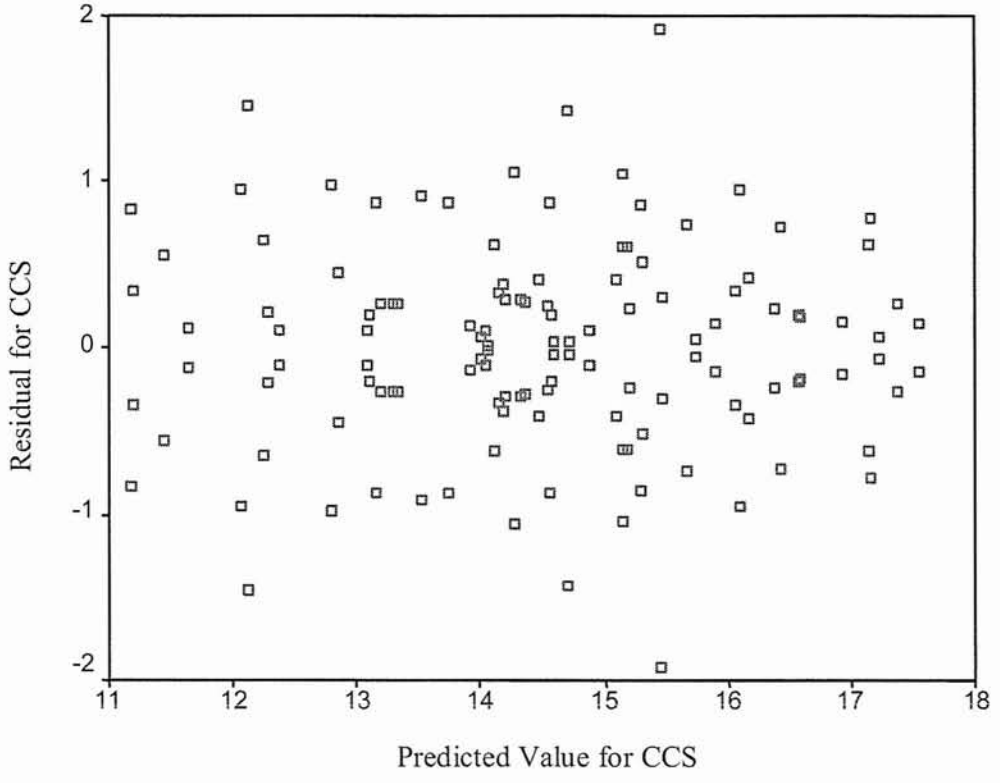

Figure 2.3: Predicted values of CCS versus residuals in the trial 1995-1 indicating that the model assumptions were suitable to the data

To determine whether the distribution of a variable matches a test distribution, in this case a normal distribution, a Q-Q plot was drawn for both CCS and cane yield for each trial. It is designated Q-Q since it plots the quantiles of a variable's distribution against those of test distributions. Again the data for CCS from the field data 1995-1 was chosen at random as an illustration, as all field data produced similar graphs to one another. If the selected variable matches the test distribution, the points cluster around a straight line, as in Figure 2.4 where the vicinity of the points to the $y = x$ line indicates that the data is approximately normally distributed.

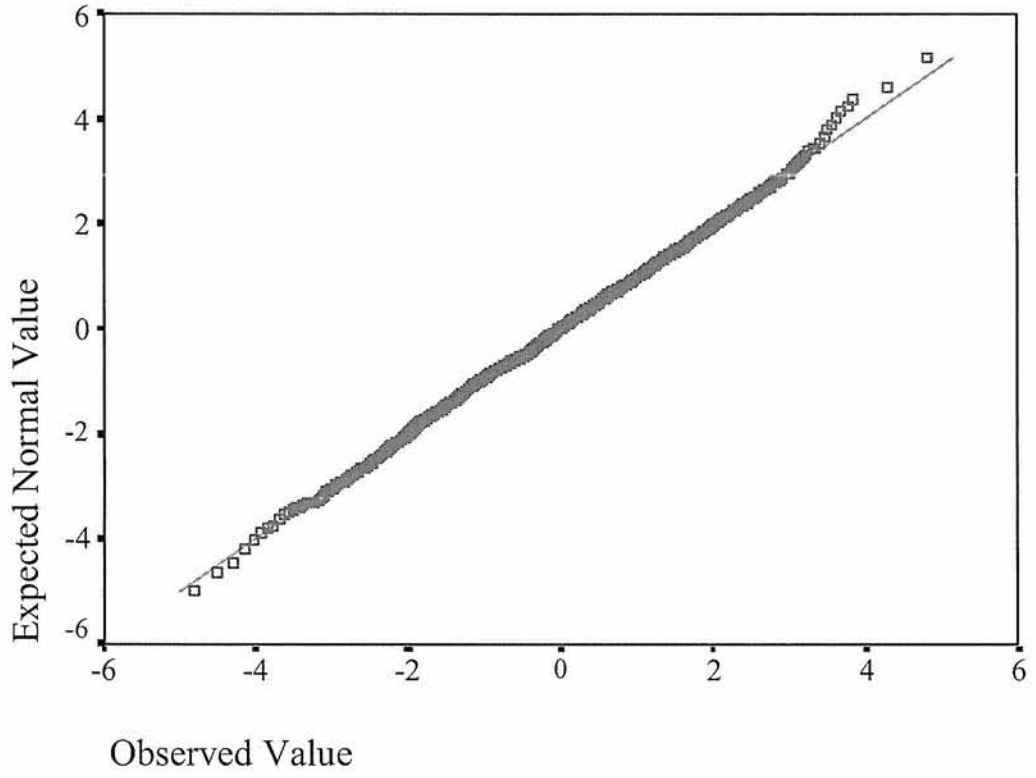

Figure 2.4: Q-Q plot for CCS from the 1995-1 trial indicating the normality of the data

2.4.3 Breeders' judgement of the required estimates and the estimates used in the simulation model SSSM

Estimates obtained from analyses of data (Section 2.4.1 and Section 2.4.2) were compared with "judgements" from two experienced sugarcane breeders: Phillip Jackson (CSIRO) and Mike Cox (BSES). Values proposed by breeders are not based on a single field experiment data but rather on many field trials conducted in the region. Thus, they represent breeders' "feeling" on what each of these components is expected to be in the region and as such are of more importance then the results obtained according a single field trial (Section 2.4.1 and Section 2.4.2). "Breeders" values are given in tables Table 2.7 and Table 2.8 alongside the estimates obtained from Section 2.4.1 and Section 2.4.2. Note that data needed to estimate, within this study, the total variability attributable to differences between families δ was unavailable at the time of research. Subsequently, the breeders' value of 30% was the estimate available and thus, used in the following sections.

Table 2.7

Comparison of the breeders' estimates for the genetic variance σ_g^2, genotype by environment interaction variance σ_x^2, correlation between genotype and competition $\rho_{g,c}$, and the proportion of variability among versus within families δ with those resulting statistical analyses (Sections 2.4.1 and 2.4.2) for CCS and cane yield (TCH)

	CCS		TCH	
	Breeders	Sections 2.4.1 &2.4.2	Breeders	Section 2.4.1 & 2.4.2
$\hat{\sigma}_g^2$	1.44	4.66 (Table 2.2)	506.25	531.63 (Table 2.2)
$\hat{\sigma}_x^2$	0.43	0.46 (Table 2.6)	151.88	61.42 (Table 2.6)
$\hat{\rho}_{g,c}$	0.20	-0.26 (Table 2.4)	0.2	0.10 (Table 2.4)
$\hat{\delta}$	30.00%	NA	30.00%	NA

The estimate for the genetic variance σ_g^2 for cane yield of 531.63 (Table 2.2) was quite close to the breeders' judgement of 506.25 (Table 2.7) for the trait. On the other hand the estimate for the genetic variance σ_g^2 for CCS of 4.66 (Table 2.2) was much higher than the value of 1.44 (Table 2.7) assigned by breeders.

The estimates for GE interaction variance σ_x^2 of 0.46 (Table 2.6) for CCS was close to breeders' judgements of 0.43 (Table 2.7). However, the estimate of σ_x^2 for cane yield was less then half at 61.42 (Table 2.6) of the estimated by breeders to be 151.88. This inconsistency could be due to the fact that statistical analyses (Section 2.4.1) were performed on data from a small number of trials conducted on four sites whereas plant breeders' judgement was based on their experiences on trials throughout the region and stretching many years.

Plant breeders provided an estimate for the total variability attributable to differences between families δ of thirty percent for both CCS and cane yield (Table 2.7). This was consistent with findings by Jackson and McRae (1998), Brown *et al* (1968) and Skinner *et al* (1987) (Section 2.3.1). The breeders' estimate for the correlation between genotype and competition $\rho_{g,c}$ was 0.2 across plot sizes, for both CCS and cane yield (Table 2.7), while the estimates based on the data gathered from the Burdekin region (Section 2.4.1) gave averages for the parameter of -0.26 and 0.10 for CCS and cane yield respectively (Table 2.4). This inconsistency can be interpreted by the fact that a single data set was used to obtain the averages (Table 2.4) whereas plant breeders judgement is based on many years of field trials.

Table 2.8 presents the comparison between breeders' judgement and estimates obtained in Section 2.4.1 for the variance parameters that depend on the plot size. Note that the data needed to estimate variance components in single seedling and four-row plots, both often used in the region, were not available at the time of research. The error variance estimate, $\hat{\sigma}_e^2$ in one and two row plot equals 3.42 and 1.17 respectively and the plant breeders' judgement for this statistic was 2.08 and 1.44 (Table 2.8). Similarly, the genetic correlation between pure stand and one-row plot and pure stand and two-row plot $\hat{\rho}_{g,g+c}$ for cane yield is estimated to be 0.52 (small i.e. one row plot mean) and 0.61 (medium i.e. two row plot mean) (Table 2.3) respectively, which is close to breeders' estimates of 0.50 and 0.60 (Table 2.8).

Generally, most results obtained for one and two row plots obtained from the data

gathered from the Burdekin region (Section 2.4.1) closely correspond to the breeders' estimates (Table 2.8). The only exception was the error variance estimate $\hat{\sigma}_e^2$ for cane yield in two row plot being almost a half at 488.37 (Table 2.2) of the breeders' estimate for the same parameter at 900.00 (Table 2.8).

Table 2.8

Comparison of the plant breeders' estimates for the error variance σ_e^2 and genetic correlation between pure stand and plot size $\rho_{g,g+c}$ for CCS and cane yield (TCH) to those resulting statistical analyses (Sections 2.4.1)

	CCS				TCH			
	$\hat{\sigma}_e^2$		$\hat{\rho}_{g,g+c}$		$\hat{\sigma}_e^2$		$\hat{\rho}_{g,g+c}$	
Plot size	Breeders	Section 2.4.1	Breeders	Section 2.4.1	Breeders	Section 2.4.1	Breeders	Section 2.4.1
Single seedling	3.24	NA	0.80	NA	2025.00	NA	0.40	NA
One row plot	2.08	3.42	0.90	0.88	1406.25	1310.80	0.50	0.52
Two row plot	1.44	1.17	0.95	0.93	900.00	488.37	0.60	0.61
Four row plot	0.52	NA	1.00	NA	225.00	NA	1.00	NA

Table 2.7 and Table 2.8 were used in developing the initial set of parameters (Table 2.9) for the SSSM. Although the statistical analysis performed during the course of this study (Section 2.4.1 and Section 2.4.2) was invaluable in establishing magnitudes of variance components necessary to define populations of phenotypic effects using the linear model (2.1), nevertheless it was based on the data obtained from only two experiments out of many conducted in the Burdekin region. Because of that it was decided that between the two sets of estimates, greater emphasis should be given to those based on years of breeders' experience. The paragraphs to follow detail the rationale used when deciding between the two sets of estimates, with Table 2.9 summarising the set of parameters initially used in the simulation model SSSM when applied to the Burdekin region.

Because of the closeness of the two estimates; breeders' judgement and statistical analysis (Section 2.41 and Section 2.4.2); for the genetic variance estimate $\hat{\sigma}_g^2$ for cane yield the breeders' estimate of 506.25 (Table 2.7) was adopted. However, for CCS a middle point of 3.24 between 1.44 and 4.66 was adopted (Table 2.9). The plant breeders' estimates for GE interaction variance σ_x^2 for both CCS and cane yield of 0.43 and 151.88 (Table 2.7) respectively were taken as initial estimates.

Table 2.9

Summary of initial parameter estimates used for both CCS and cane yield (TCH), for the genetic variance σ_g^2, correlation between genotype and competition $\rho_{g,c}$, proportion of variation between families δ, genotype by environment interaction variance σ_x^2, error variance σ_e^2 and genetic correlation between pure stand and plots $\rho_{g,g+c}$ for four plot sizes

	$\hat{\sigma}_g^2$	$\hat{\rho}_{g,c}$	$\hat{\delta}$	$\hat{\sigma}_x^2$
CCS	3.24	0.2	30%	0.43
TCH	506.25	0.2	30%	151.88
Plot size	$\hat{\sigma}_e^2$		$\hat{\rho}_{g,g+c}$	
	CCS	TCH	CCS	TCH
Single seedling	3.24	2025.00	0.8	0.4
One row plot	2.43	1406.25	0.88	0.52
Two row plot	0.92	729.00	0.93	0.61
Four row plot	0.59	324.00	1	1

Because of the lack of data to estimate the total variability attributable to differences between families δ, δ of thirty percent for both CCS and cane yield (Table 2.7) was adopted (Jackson and McRae, 1998; Brown *et al,* 1968 and Skinner *et al,* 1987) (Section 2.3.1). Similarly, because of the lack of data the correlation between genotype and competition $\rho_{g,c}$, plant breeders' estimate of this variable of 0.2 across plot sizes, for both CCS and cane yield (Table 2.7) was adopted.

As shown in Table 2.9, for the error variance estimate, $\hat{\sigma}_e^2$ in one and two row plot a middle point of 2.43 between 3.42 and 2.08 and a lower value of 0.92 that the two estimates 1.17 and 1.44 (Table 2.9) was agreed upon respectively. For the single seedling and the four row plot the plant breeders' estimate of 3.24 and 0.59 were adopted. In the case of the genetic correlation between a pure stand and one-row plot and that between pure stand and two-row plot $\hat{\rho}_{g,g+c}$ for both CCS and cane yield, on the other hand, statistical estimates were taken as initial parameters; being 0.88 and 0.93, and 0.52 and 0.61 respectively; and for the single seedling and four row plot breeders' suggestions were taken; being 0.8 and 1, and 0.4 and 1 respectively.

In cases when breeders' judgements and statistical analyses coincide the estimates could be used with greater confidence. However, some parameters where level of confidence is lacking and/or where further designed field trials are needed to ensure the accurate estimate include:

- error variance σ_e^2 both in CCS and cane yield in single seedling and four row plot
- correlation between genotype and competition $\rho_{g,c}$ for both CCS and cane yield
- genetic correlation between pure stand and single seedling and between pure stand and four row plot $\rho_{g,g+c}$ for both CCS and cane yield
- proportion of variability between families δ
- genotype by site interaction variation σ_x^2 for unselected populations

2.4.4 An illustration of the parameter computations for a selection stage

Table 2.10 outlines the set of variance components relevant for stage one of a typical selection system from the Burdekin region (Section 1.1.1). To compute these variance components the initial statistical parameters (Table 2.9) estimated in Sections 2.4.1 and 2.4.2 were used together with the quantitative genetics theory (Section 2.2 and Section 2.3).

Table 2.10

Summary of the calculated estimates based on the initial set of estimates (Table 2.9) relevant for stage one of the selection system in the Burdekin region (Section 1.1.1): the between family genetic variance σ_f^2, competition variance σ_c^2, genotype by environment interaction variance σ_x^2, error variance σ_e^2, correlation between genotype and competition $\rho_{g,c}$

CCS					TCH				
$\hat{\sigma}_f^2$	$\hat{\sigma}_c^2$	$\hat{\sigma}_x^2$	$\hat{\sigma}_e^2$	$\hat{\rho}_{g,c}$	$\hat{\sigma}_f^2$	$\hat{\sigma}_c^2$	$\hat{\sigma}_x^2$	$\hat{\sigma}_e^2$	$\hat{\rho}_{g,c}$
0.97	0.26	0.13	0.61	0.20	151.90	239.40	45.56	351.56	0.20

Because family selection was used in stage one (Section 1.1.1), the initial genetic variance estimates $\hat{\sigma}_g^2$ of 3.24 and 506.25 (Table 2.9) for CCS and cane yield

respectively were used to calculate the between families variance component $\hat{\sigma}_f^2$. According to the proportion of variability between families estimate $\hat{\delta}$ of 30% (Table 2.9), the between families variance component $\hat{\sigma}_f^2$ relevant for the starting population of genetic effects were 0.97 and 151.90 for CCS and cane yield respectively (Table 2.10).

Initial competition variance estimates $\hat{\sigma}_c^2$ in one row plot of 0.94 and 1016.45 for CCS and cane yield respectively were calculated by substituting the following parameter values into formula 2.20 (Section 2.3.10).

(i) genetic correlations estimates $\hat{\rho}_{g,g+c}$ between pure stand and one row plot of 0.88 and 0.52 respectively for CCS and cane yield (Table 2.9);

(ii) genetic variance estimates $\hat{\sigma}_g^2$ of 3.24 and 506.25 (Table 2.9) for CCS and cane yield respectively;

(iii) correlation between genotype and competition estimates $\hat{\rho}_{g,c}$ of 0.20 for both CCS and cane yield (Table 2.9).

Again, since family selection is practiced in stage one the proportion of variability between families estimate $\hat{\delta}$ of 30% (Table 2.9) was used to give 0.26 and 239.40 for CCS and cane yield respectively (Table 2.10).

The GE interaction variance estimate $\hat{\sigma}_x^2$ of 0.43, 151.88 respectively for CCS, cane yield (Table 2.9) was used to calculate the variability attributable for the family selection 0.13 and 45.56 respectively for CCS and cane yield (Table 2.10). The magnitude of $\hat{\sigma}_x^2$ on a family mean basis was determined from $\hat{\sigma}_x^2/s$, where s is the number of environments ie sites used in evaluating families.

Since one row plots are currently used in sugarcane breeding programs for family selection, error variance σ_e^2 of 3.24 and 2025.00 were initially assumed for CCS and cane yield respectively (Table 2.9). On a family (genotype) basis, error variance estimates were determined from these variances divided by the total number of

replicates (product of number of sites by number of replicates per site) used in calculating them to calculate 0.61 and 351.56 for CCS and cane yield respectively. The correlation between genotype and competition estimates $\hat{\rho}_{g,c}$ for both stages equal to 0.20 for both CCS and cane yield (Table 2.9).

Because the magnitude of each relevant parameter (Table 2.9) changes depending on the input selection variables (Section 1.1.2), to reflect the difference onto the selection results the computations similar to those detailed above were performed prior to the simulation of each stage.

Now that the way in which selection variables (Section 1.1.2) impact on the observed expressions of CCS and cane yield have been elucidated, assumptions of the model tested and an initial set of variance components estimated for the region of interest, the selection system from the Burdekin region can be mathematically modelled.

Chapter 3

Simulation of selection systems

3.1 Introduction

In Section 2.2 it was explained how observed values for CCS and cane yield could be expressed as a linear model (equation 2.1). The magnitude of the effects that constitute the phenotypic value y_{ijk}: the genetic effect g_i, competition effect c_{ik}, GE interaction effect x_{ij}, error effect e_{ijk}; depend on the plot size p_z, the number of sites s_z and the number of replicates r_z per site used at a particular stage z. Section 2.4.4 detailed the procedure used to compute the variance components to be used in stage one of the selection system defined in Section 1.1.1. As a genotype passes through different selection stages, the magnitude of each generated effect, apart from the genetic effect g_i, changes for each of the two traits CCS and cane yield, and this affects the observed values for these traits. In addition to the performance for these traits, the particular genotypes that are selected at a stage z also depend on the selection index d_z (a function of CCS and cane yield) and selection intensity t_z used in stage z. Gains made in any stage of the selection are affected by all the above variables. Therefore, the performance of a selection system depends on the combination of design variables (Section 1.1.2) used through the system.

The interactions between selection variables (Section 1.1.2) and the way in which they impact on the magnitude of the phenotypic expressions for CCS and cane yield are not easily expressible by a set of mathematical formulas. The factors that influence the

performance of selection system are thus too complicated to use a deterministic model and a stochastic model is required. Furthermore, the simulation of phenotypic values for CCS and cane yield involve generation of random variables, which indicate that rather than a single outcome, a distribution of solutions are obtained. Thus, the appropriate simulation model that can capture the "real life" selection system is a Monte Carlo stochastic simulation model (Winston, 1994).

Simulation models have been successfully applied in a wide range of situations, such as assessing the cost effectiveness of asthma management strategies (Price and Briggs, 2002), evaluating the operation of wastewater treatment plants (Makinia *et al*, 2002); testing different measles vaccination schedules and their dosages (Zekri and Clerc, 2002), determining peanut farm net returns (Parman *et al*, 1996), and testing different harvesting strategies (Semenzato, 1995, and Sorensen and Gilheany, 1970).

Section 3.2 develops a framework for the simulation model SSSM (Sugarcane Selection Simulation Model), summarising the way important effects of the sugarcane phenotypic expression were generated (eg. linear model 2.1). Section 3.3 documents the method used for comparing selection systems in terms of cost and the gain. Section 3.4 provides verification of the accuracy of the simulation model SSSM as well as documenting the sensitivity analysis of selection gain to change in the statistical parameters.

3.2 Overall concepts and structure of the Sugarcane Selection Simulation Model (SSSM)

In this section the overall structure of the SSSM model is described (Section 3.2.1), with Section 3.2.2 giving a short application manual of the model. Section 3.2.3 then explains how the effects of the linear model (2.1) are generated to give the magnitude of phenotypic values for CCS and cane yield, which define genotypes. Selection variables (Section 1.1.2) that affect the magnitude of the phenotypic values for CCS and cane yield for a given genotype, y_{ijk} at a particular stage z are the plot size p_z, number of sites s_z and number of replicates per site r_z used at that stage. Section 3.2.4 explains the selection of the genotypes so generated. Following the generation of phenotypes,

these are ranked according to the selection index, d_z, and then the highest ranking ones selected to go forward to the next stage. The proportion selected is given by the selection intensity, t_z.

3.2.1 SSSM flow chart

A flow chart of the simulation model SSSM is given in Figure 3.1. Selection variables (Section 1.1.2) need to be entered into to the simulation model by the user. The SSSM first checks whether selection in stage 1 is family or individual, based on inputted selection variables.

When family selection is chosen at stage one, the selection variables necessary to select families in the first year of stage one as well as the variables necessary to select genotypes from within selected families are collected. Families are then generated according to the defined plot size, site and replication numbers. A user-defined percentage of families are selected, based on a defined selection index. Then for each selected family a specified number of genotypes are generated, and a specified percentage of genotypes from within each family are selected based on the phenotype of CCS and cane yield, and a defined selection index, which is a function of these two traits. For all following stages, individual clone selection is used.

When individual selection is practiced at any stage a similar procedure is followed. The number of genotypes is first generated according to the plot size; site number and the number of replicates defined by the user. A defined percentage of the generated genotypes are selected based on the phenotype of CCS and cane yield, and the selection index defined by the user. The size of the population generated depends on the stage at which individual selection of genotypes is used, and it could either be user-specified number, in the case it is stage one; or the number of genotypes selected from the previous stage, for all other stages.

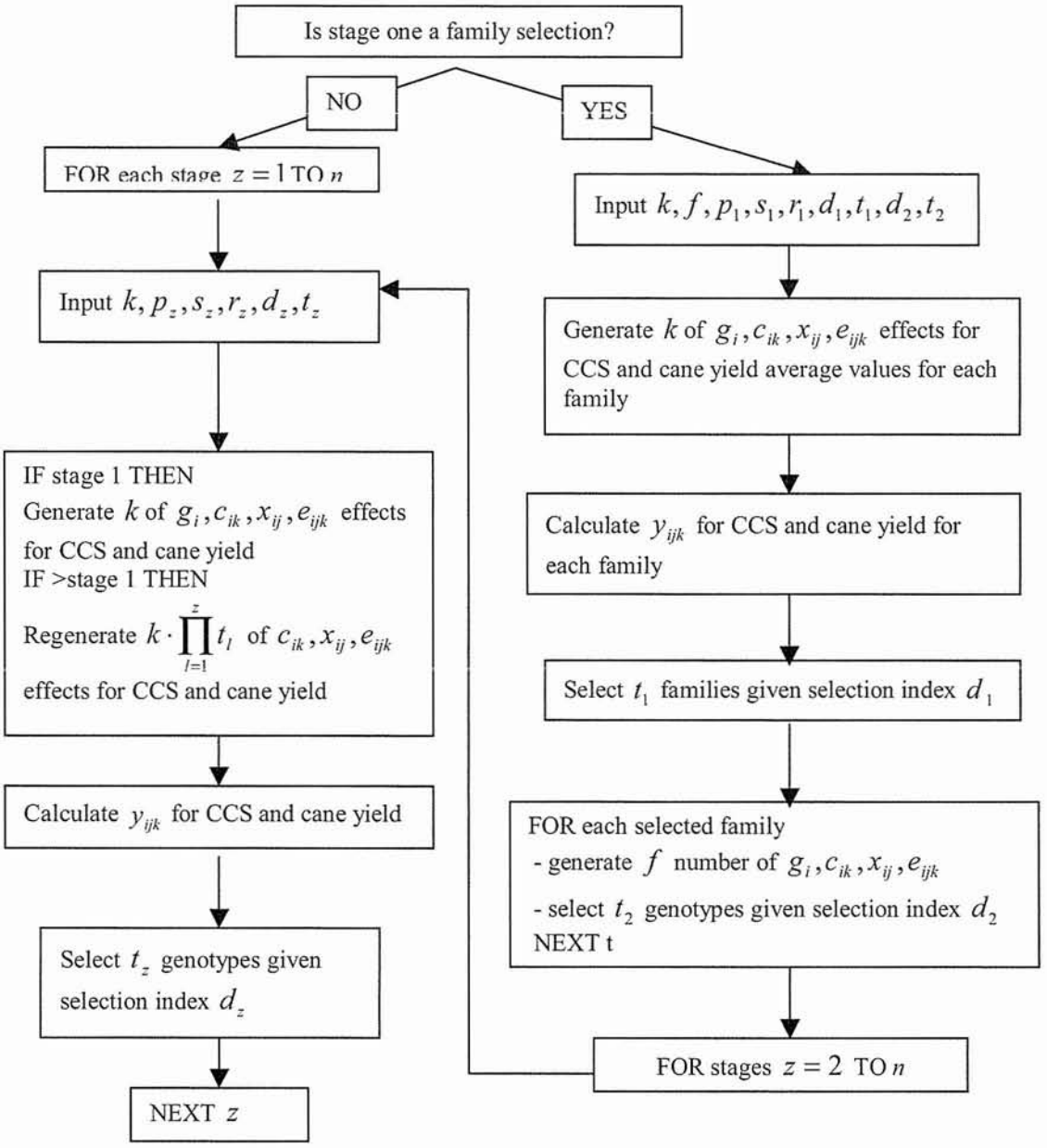

Figure 3.1: The flow-chart of the SSSM, where: g_i is the genotype effect of genotype i; c_{ik} is the competition effect of genotype i being planted in plot size k; x_{ij} is the genotype by environment interaction effect of genotype i being planted in environment j; e_{ijk} is the error effect; n the number of stages; k the number of families; f the number of genotypes per family; p_z plot size, s_z number of sites, r_z number of replicates, d_z selection index, t_z selection intensity used at a stage z

3.2.2 SSSM application manual

The simulation model SSSM application was developed in Microsoft Access® using the Visual Basic® (Fox, 1999) code. The application consists of a number of user interfaces, some of which are given in Figures 3.2 and Figure 3.3.

A selection system to be simulated is defined by entering selection variables (Section 1.1.2). Figure 3.2 shows the SSSM user interface that allows the user to design a new selection system. In the case of family selection being used in stage one, the family option check-box needs to be ticked and all the variables that define the selection of genotypes from within families need to be entered (Figure 3.2). Note that there are three options given for the selection index at this stage: sugar yield, user defined weighing between CCS and cane yield, or visual evaluation of genotypes; whereas at all other stages, two first options are available, as visual evaluation is only applicable to the second part of stage one when genotypes are selected from within each selected family (Section 2.3.5).

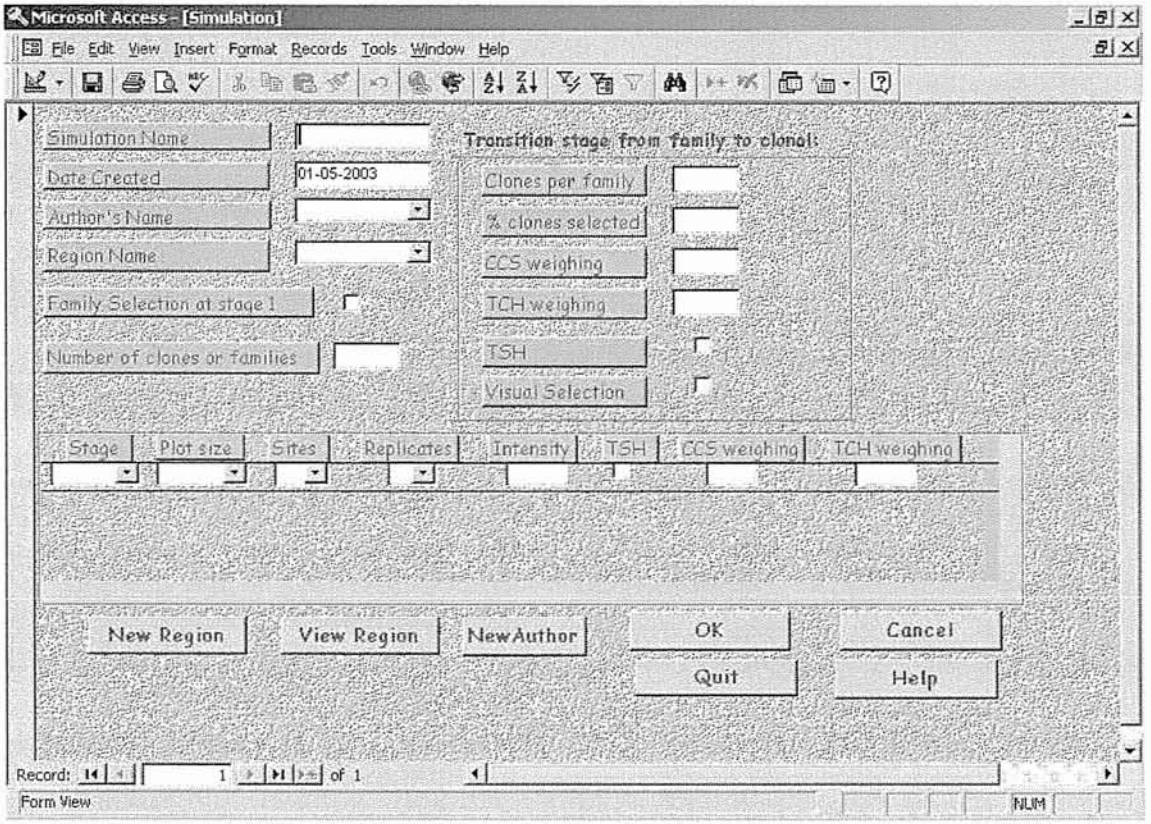

Figure 3.2: The SSSM interface that allows the definition of a new selection system

This interface (Figure 3.2) gives further options to define a new region through entering all the variance parameters (Table 2.9) or to view and change an existing region. To be able to identify and later retrieve the information regarding any selection system simulated, each is given a unique name.

Because of variation in genetic gain due to random effects, the simulation model may be run a number of times to obtain a precise estimate of the performance of particular selection system designs and therefore accurately compare different designs. The user specifies how many times the simulation should be run. The result given for genetic gain is the mean of a number of runs specified. The final results (Figure 3.3) interface gives detailed information regarding the performance of the selection system. There are further options available in order to view the performance of each selection stage separately. The main result is the "Pure Economic gain" and cost of the selection system chosen, both of which are explained later in this chapter.

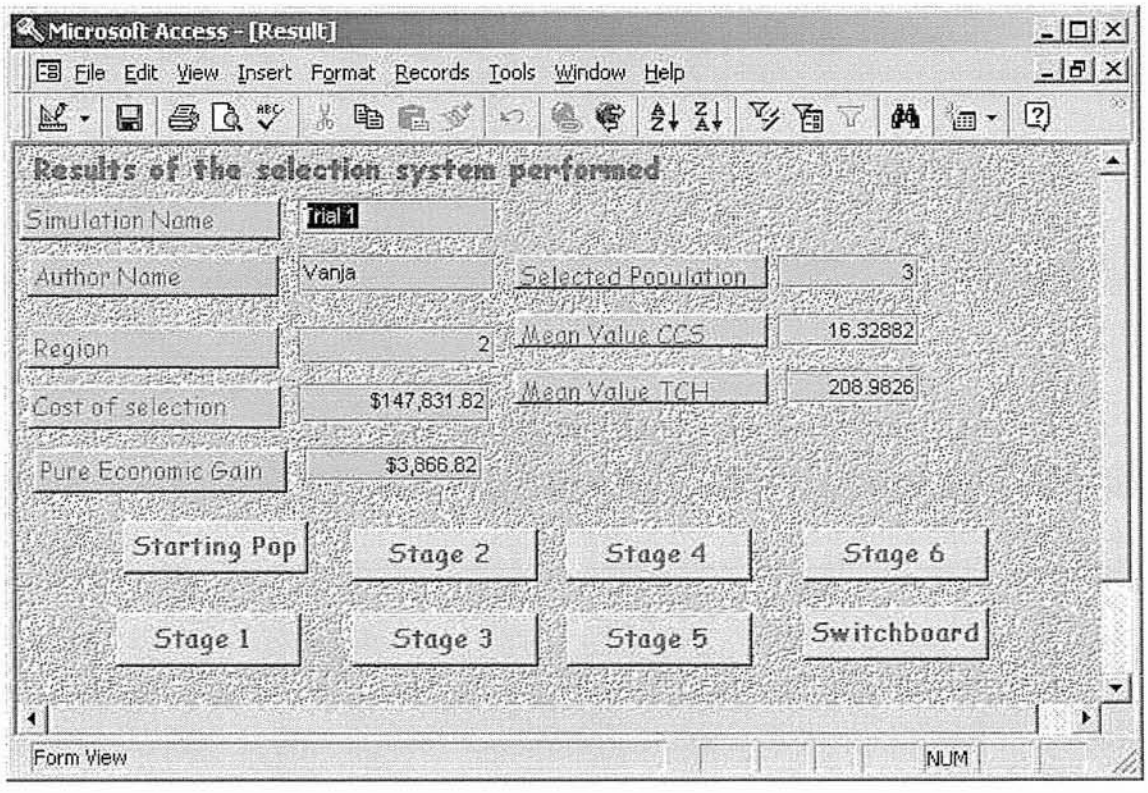

Figure 3.3: The SSSM interface that gives main selection simulation results

3.2.3 Generation of phenotypic effects

Although all effects of the linear model (2.1) are sampled from normally distributed populations, not all are generated in the same way. The genotype g_i and competition c_{ik} effects are in some circumstances correlated with each other (Section 2.3.2) and therefore need to be generated to maintain the relationship between them. Furthermore, the genetic effect g_i stays unchanged throughout selection as it does not depend on planting environment in which the genotype has been grown, whereas all other effects change. The error effect e_{ijk} (Section 2.3.4) and the genotype by environment interaction effect x_{ij} (Section 2.3.3) depend on the experimental design used and therefore need to be regenerated at each new stage of selection. The competition effect c_{ik} changes as the plot size changes, but potential competitiveness of a genotype is defined by its genotype, which stays unchanged. Therefore, c_{ik} needs to be re-scaled according to the plot size with each stage of selection, rather then re-generated anew.

3.2.3.1 Generation of the error effect and genotype by environment interaction effect

The error effect e_{ijk} and GE interaction effect x_{ij} that are not correlated to any other effect in the linear model (2.1), are generated independently from the area under the bell-shaped curve similar to the one in Figure 3.5 defined by the univariate probability density function (Walpole, 1990):

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\left[\left(\frac{x}{\sigma}\right)^2 / 2\right]} \tag{3.1}$$

where σ^2 is replaced by σ_e^2 for the error effect and with σ_x^2 for the genotype by environment interaction effect. The Visual Basic®'s standard function *Rnd* randomly generates a number ξ from the [0,1] interval. Thus, the numbers for within each increment within the [0,1] interval are generated with the equal probability using the Rnd function. Let $\xi \in [0,1]$ represent the area under the standard normal curve (Figure

3.4) to the left of a standard score, z_0. There is only one standard score z_0 such that the area to its left equals ξ. Normal tables (Walpole, 1990) give z-scores and the corresponding areas under the normal curve between zero and the z-score.

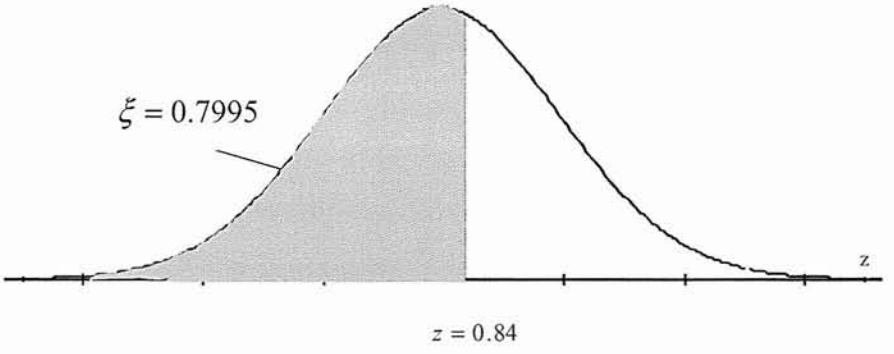

Figure 3.4: An example of a z-score $z = 0.84$ and the corresponding area $\xi = 0.7995$ under the standard normal curve

The formula for standard score z :

$$z = \frac{x - \mu}{\sigma} \tag{3.2}$$

where μ and σ are respectively mean and standard deviation of the population of interest, is then used to calculate the magnitude of the effect x: $x = \mu + z \cdot \sigma$, ie $x = z \cdot \sigma$ because $\mu = 0$.

For example, if 0.7995 is the randomly generated number it corresponds to the area of $\xi = 0.7995$ under the standard normal curve (Figure 3.4), which corresponds to the standard score $z = 0.84$ from the Normal tables. Given the standard deviation σ of the desired population and the standard score, the formula $x = z \cdot \sigma$ would generate an effect x from the desired population.

3.2.3.2 Generation of the genotype and the competition effects

Previous studies have indicated that genetic effects and competition effects may be correlated (Section 2.3.2). Therefore, the two effects can not be generated independently as can GE interaction and error effects (Section 3.2.3.1), but rather they come from a two-dimensional normal distribution. The equi-probability contour of the region from which the two correlated effects to be generated can be computed from the following bivariate probability density function is (Deák, 1990):

$$f(g_i, c_{ij}) = \frac{1}{2\pi|\Sigma|^{1/2}} e^{-(X-\mu)'\Sigma^{-1}(X-\mu)/2} \tag{3.3}$$

where g_i is genetic effect and c_{ik} competition effect of genotype i, Σ is the covariance matrix of the genotype and competition effects,

$$\Sigma = \begin{bmatrix} \sigma_g^2 & \mathrm{cov}(g,c) \\ \mathrm{cov}(g,c) & \sigma_c^2 \end{bmatrix} \tag{3.4}$$

The contour of constant density for the bivariate normal distribution is an ellipse (Figure 3.6) defined by all $X = (g,c)'$ such that

$$(X-\mu)\Sigma^{-1}(X-\mu)' = \kappa^2 \tag{3.5}$$

where $\kappa^2 = \chi_2^2(\alpha)$ is the upper $(100\cdot\alpha)^{\text{th}}$ percentile of a chi-squared distribution with two degrees of freedom (Johnson and Wichern, 1982). The ellipse is centred at μ and have axes $\pm\kappa\sqrt{\lambda_1}\tilde{e}_1$ and $\pm\kappa\sqrt{\lambda_2}\tilde{e}_2$, where $\lambda_1, \tilde{e}_1$ and $\lambda_2, \tilde{e}_2$ are the two eigenvalue-eigenvector pairs of the covariance matrix Σ. Note that for the genotype and competition effects, the ellipse is centred in the origin $(0,0)$ since the mean value μ for both effects equals zero (Figure 3.5).

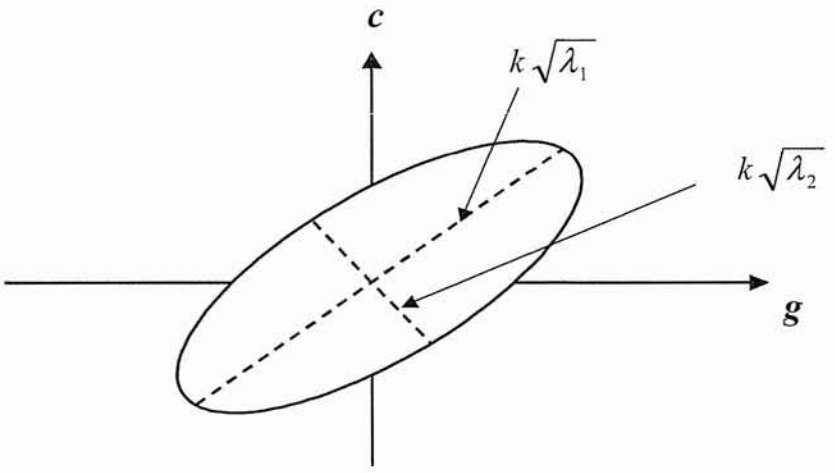

Figure 3.5: The population of correlated pairs (g_i, c_{ik}) was generated from within the ellipse (cross section) with major and minor axis being defined by $\kappa\sqrt{\lambda_1}$ and $\kappa\sqrt{\lambda_2}$ respectively

One way to compute the correlated pair (g_i, c_{ik}) is by solving the following system of simultaneous equations:

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} \tilde{e}_1 & \tilde{e}_2 \end{pmatrix}' \begin{pmatrix} g_i \\ c_{ik} \end{pmatrix} \tag{3.6}$$

where y_1, y_2 are two randomly generated numbers from normally distributed populations with mean zero and variances equal to $\kappa^2\lambda_1$ and $\kappa^2\lambda_2$ respectively.

Alternatively, if y_1, y_2 are two randomly generated numbers from a normally distributed populations with mean zero and variances one, then the two correlated effects g_i, c_{ik} could be computed by the following set of formulas (Kleijnen and Van Groenendaal, 1992):

$$\begin{aligned} g_i &= y_1 \cdot \sigma_g \\ c_{ik} &= \sigma_c \left(\rho_{g,c} \cdot y_1 + \sqrt{1 - \rho_{g,c}^2} \cdot y_2 \right) \end{aligned} \tag{3.7}$$

where $\rho_{g,c}$ is the correlation between genotype and competition effects. Due to its simplicity and accuracy the latter approach was adopted in the simulation model SSSM to generate the two correlated effects g_i, c_{ik}.

3.2.3.3 Re-calculation of the competition effect

Competition effect c_{ik} represents a response of a genetic component to the change in the plot size. Therefore, it cannot be regenerated like other components, but rather needs to be reduced or enlarged depending on plot size used (Section 2.3.2).

If (g_i, c_{ik}) is the original correlated genetic and competition pair from the starting population as defined in Section 3.2.3.2, the first component g_i of the pair stays the same throughout selection. The competition effect c_{ik} is enlarged or reduced using the transformation c'_{ik}:

$$c'_{ik} = \frac{(c_{ik} - \bar{x})}{\sigma_o}\sigma_n \tag{3.8}$$

where σ_o^2 and σ_n^2 are the generated population competition variance – 'old' and the desired competition variance – 'new' respectively and $\bar{x}$ is the mean value of c_{ik}'s. To ensure that the mean equals zero before the transformation, the average $\bar{x}$ is taken away from the original competition.

3.2.3.4 An illustration of the generation of populations of effects

Firstly, the set of variance components that defines each population of effect need to be computed from the initial statistical parameters (Table 2.9). Section 2.4.4 illustrated the computation of the relevant variance components for stage one of the selection system as detailed in Section 1.1.1.

Next, a desired number of genotype effects g_i, competition effects c_{ik}, GE interaction effect x_{ij} and error effects e_{ijk} are generated, as described in sections 3.2.3.1 and 3.2.3.2. Table 3.1 illustrates the generation of effects for fifteen genotypes. Once all the effects have been generated the phenotypic values y_{ijk} for CCS and cane yield are calculated according the linear model (2.1) $y_{ijk} = \mu + g_i + c_{ik} + x_{ij} + e_{ijk}$. Note that the mean values μ for CCS and cane yield were proposed by two experienced sugarcane breeders: Phillip Jackson (CSIRO) and Mike Cox (BSES) to be 12% and 150 tonnes per hectare respectively.

Table 3.1

An illustration of the generation of populations of effects that make the phenotypic value $y_{ijk} = \mu + g_i + c_{ik} + x_{ij} + e_{ijk}$ for CCS and cane yield (TCH), where g_i is the genetic effect, c_{ik} competition effect, x_{ij} GE interaction effect, and e_{ijk} error effect for a genotype i being planted in environment j and plot size k

i	CCS						TCH					
	μ	g_i	c_{ik}	x_{ij}	e_{ijk}	y_{ijk}	μ	g_i	c_{ik}	x_{ij}	e_{ijk}	y_{ijk}
1	12	0.565	0.132	0.049	1.755	14.501	150	-22.846	3.479	3.582	16.313	150.528
2	12	0.289	0.143	0.670	0.640	13.744	150	-3.095	-15.253	1.653	9.375	142.680
3	12	-0.638	-0.780	-0.031	0.140	10.691	150	16.762	-11.634	-2.039	-8.813	144.276
4	12	-0.931	-0.633	-0.154	-0.913	9.369	150	10.565	-10.175	-3.031	-13.500	133.859
5	12	-0.389	-0.660	0.216	1.139	12.306	150	-8.206	-0.999	-3.582	3.938	141.151
6	12	-1.898	-0.187	0.397	1.225	11.537	150	2.813	-30.697	-4.741	-15.188	102.487
7	12	-0.805	0.179	-0.097	1.708	12.985	150	-5.478	-10.132	2.263	-21.375	115.278
8	12	-1.186	-0.120	-0.785	0.640	10.549	150	-21.334	-9.302	-7.110	12.188	124.442
9	12	0.797	-0.091	0.146	-1.115	11.737	150	19.162	-5.116	5.236	-18.375	150.907
10	12	0.512	-0.485	0.304	0.920	13.251	150	2.916	-21.535	0.000	14.438	145.781
11	12	0.294	0.845	0.459	1.123	14.721	150	-22.620	-19.627	-0.661	6.375	113.467
12	12	0.830	-0.383	-0.375	-0.601	11.471	150	6.913	-5.602	-1.102	-27.938	122.271
13	12	-0.779	-0.448	0.093	-1.498	9.368	150	-12.796	-5.178	-2.039	8.813	138.800
14	12	1.056	0.015	0.604	-0.390	13.285	150	12.724	23.018	0.772	-9.375	177.139
15	12	0.175	0.330	0.009	-1.069	11.445	150	-9.283	-15.656	3.858	-14.063	114.856

3.2.4 Selection of genotypes

Following generation of effects as described above, a selection index is calculated for each genotype based on phenotypes for CCS and cane yield. Genotypes are then ranked according to this selection index and the top ranked genotypes are then selected, with the actual number being selected dependent on the intensity of selection. Details of these steps are given below.

(i) Selection index d_z

The selection index d_z is a function of CCS and cane yield and is defined for each stage z of the selection. In the model used in this study, the selection index used comprised one of the following three types:

1. Sugar yield (TSH) calculated as $TCH \times CCS$

2. Visual rating, a selection criterion correlated to cane yield (TCH)

3. A function, $d = \alpha \cdot CCS + \beta \cdot TCH$, where α and β were two real numbers chosen by the user. Assigning for example $\alpha = 0$ and $\beta = 1$ creates a selection index based on cane yield only.

In practice, for selection systems in Australia, the visual estimate of cane yield and the general appearance of the cane visual rating is used only for the second part of stage one in selection systems when genotypes from within selected families may be selected to enter the subsequent stage. However, since visual rating is subject to breeders' personal evaluation of the physical appearance of a particular genotype, it is not possible to simulate it using a computer program. However, it is assumed that visual appearance rating relates predominantly to cane yield and when this is chosen the SSSM uses cane yield as the selection criterion, and treats it with a level of error variation considered appropriate by the user for this method of estimating cane yield.

(ii) Selection intensity t_z

The selection intensity t_z defined at each stage z is the percentage (between 1% and 100%) of genotypes, or families in the case of family selection, to be transferred to the next stage $z+1$ of selection.

Note that in the actual selection system, genotypes selected during the last stage of selection are further tested for disease resistance and ratoon performance, so that finally only a small set of genotypes would be released for industry adoption and evaluation. In order to select enough genotypes to allow for these testings and to make selection systems comparable to one another it was necessary to limit the maximal number of genotypes to result from a selection system. For the purposes of this study, thus, it was adopted that at most ten genotypes could be selected through selection systems simulated using the SSSM.

3.3 Defining genetic gains and determining costs

The effectiveness of selection systems can be determined from (i) genetic gain achieved, and (ii) total cost. In practice, breeders are usually faced with a limited budget and their task is to maximise genetic gains within these budget limits. The framework used in this study to measure genetic gains and costs of alternative selection systems is described below.

3.3.1 Genetic gain for economic value

To allow a comparison between different selection designs, a new measure of the selection system performance, called the genetic gain for economic value ($\widetilde{G}$) is defined. The determination of this is described below.

The economic value E_i of a genotype i is defined as the difference between the return R_i from sugar produced and the costs C_i associated with producing sugar from genotype i commercially. The return R_i from sugar produced and the costs C_i

associated with growing genotype i are defined below, using the same assumptions as in Jackson and McRae (2001):

$$\begin{aligned} R_i &= \$350 \cdot TSH_i \\ C_i &= \$5.64 \cdot TCH_i + \$10.53 \cdot TCH_i + \$0.64 \cdot TSH_i + \$1013 \end{aligned} \tag{3.9}$$

where \$350 is assumed to equal the long term price of one tonne of raw sugar, \$5.46 is assumed to equal harvesting costs for one tonne of cane, \$10.53 is assumed to equal cane transportation and milling costs for one tonne of cane, \$0.64 is assumed to equal sugar transportation costs for tonne of sugar, and \$1013 assumed to equal growing costs for one hectare of cane. While these assumptions of costs and prices will change over time, the values used are expected to be adequate for obtaining accurate rankings of genotypes with varying CCS and cane yield.

The genetic gain for economic value ($\widetilde{G}$) is the difference (gain) between the average genetic effect for economic value of the starting population ($\overline{G}_{start}$) and the average genetic effect for economic value of the final selected population ($\overline{G}_{sel}$). The genetic effect for economic value G_i of a genotype i is calculated from $R_i - C_i$ as above, using the genetic effects, g_i, for CCS and cane yield.

To illustrate how genetic gain for economic value was computed, a starting population of six generated genotypes is given in Table 3.2, represented only by their genetic values g_i for CCS and cane yield. The average of the genetic effects for economic value $\overline{G} = \frac{1}{6}\sum_{i=1}^{6} G_i$ of the starting population equals \$2857.22.

Table 3.2

A sample of the computer generated population of genotype values for CCS and cane yield (TCH), expressed as the sum of $\mu + g_i$, where g_i is the genotype effect; their sugar yield (TSH) value; return R_i; costs C_i for each genotype i and genetic effect for economic value

$$G_i = R_i - C_i$$

i	CCS			TCH			TSH	R_i	C_i	G_i
	μ	g_i	$\mu + g_i$	μ	g_i	$\mu + g_i$				
1	12	-1.495	10.505	150	7.292	157.292	16.524	$5,783.40	$3,566.99	$2,216.41
2	12	-1.023	10.977	150	-16.088	133.912	14.701	$5,145.00	$3,187.77	$1,957.23
3	12	0.342	12.346	150	20.147	170.147	20.996	$7,348.60	$3,777.71	$3,570.89
4	12	0.796	12.796	150	-2.916	147.084	18.821	$6,587.35	$3,403.39	$3,183.96
5	12	0.455	12.452	150	-17.143	132.857	16.541	$5,789.35	$3,171.88	$2,617.47
6	12	0.873	12.873	150	10.064	160.064	20.605	$7,211.75	$3,614.42	$3,597.33

Select for example genotypes 1, 2, and 3. The average value of the genetic effect for economic value gain G_i of selected genotypes is $2581.51. The genetic gain for economic value $\widetilde{G}$ achieved by selecting these three genotypes equal $-$275.71 ($2581.51 - $2857.22) per hectare. This indicates that, from an economic point of view, the final population of genotypes is worse than the starting population. Having a selection stage that selected those three particular genotypes from such a starting population (Table 3.2) would result in an ineffective selection, since it discarded the economically valuable genotypes 4 and 6 and it selects genotypes 1 and 2 that are economically inferior to them.

Note that the three genotypes used above were just used to illustrate the procedure of the calculation of the genetic gain for economic value $\widetilde{G}$. In this study, the genetic gain for economic value $\widetilde{G}$ given to characterise a selection system was based on the difference in genetic values for economic value, between the ten best performing genotypes selected from the selection system and the average of the starting population.

3.3.2 Costs of selection systems

At the time of writing this thesis, there was no published information available on the costs associated with selection trials. Therefore, estimates were obtained from personal consultation with sugarcane breeders and technical staff: Dr Mike Cox (BSES), Mr Terry Morgan (CSR) and Dr Phillip Jackson (CSIRO).

Selection trial costs were classified as either (i) fixed costs that are independent of the number of plots or varieties tested in a trial or (ii) variable costs that are proportional to the numbers of plots planted and the measurements made.

(i) Fixed costs: the following fixed costs were assumed:

- finding a trial site, planning all operations associated with the trial, and liaising with growers or farm management staff: total cost of $960. This was based on an assumption of six days for a technical staff person at $160 per day ($20 per hour)

- analysis of data obtained from a trial and associated administration: total cost of $480, based on three days for a technical staff member at $160 per day ($20 per hour) rate

If the ratoon crops were grown and additional data obtained, then an additional $480 per ratoon crop would be incurred. This was the cost of analysing data and administration only, since there was no need to find trial sites and other operational costs associated with planting a new trial.

(ii) Variable costs: it was assumed that genotypes may be planted as single seedlings in the field, in one row plots 5 meters or 10 meters long, or in plots that were discrete multiples of the 10 meter long one row plot. The following costs were assumed:

- planting a ten meter long one row plot costs $10 and planting a single seedling costs $2.20;
- growing a ten meter long one row plot costs $2 and growing a single seedling costs 13c;

- measuring CCS costs \$4 regardless of whether a single seedling or one row plot is measured;
- weighing cane of a ten meter long one row plot costs \$2 and weighing a single seedling costs 13c;
- harvesting of a ten meter long one row plot costs \$1.02 and harvesting a single seedling costs 7c;
- compensation to farmers for a ten meter long one row plot costs \$1.50 and for a single seedling it costs 9c;
- visual evaluation of a single seedling costs 30c.

Every time a genotype was harvested and weighed it needed to be regrown to produce sufficient planting material for planting the next stage of selection. Thus, there is also the cost associated with ratooning or propagation of the crop. The costs incurred are those for growing the crop and compensating farmers the lost income.

Apart from variables defined below, all selection variables used in the following cost functions are as assigned in Section 1.1.2. Thus, if individual genotype selection is practiced at stage 1, the cost of that stage is calculated using the following formula:

$$C_1 = \$1{,}440 + (\$2.44 + \$4 \cdot CCS_1 + \$0.13 \cdot TCH_1) \cdot k \tag{3.10}$$

where CCS_1 ie TCH_1 is a boolean function that equals 0 or 1 depending whether CCS ie TCH needs to be measured at the stage. If, on the other hand, family selection is practiced there are two options to consider. If families are grown in 2 row plots, one juice sample is extracted per row to measure CCS. Similarly, if 4 row plots are practiced two middle rows are used, again one juice sample per row, to measure CCS. On the other hand, if one row plots are used there is only one row available to do the same. Accordingly, the cost function is either given with:

$$\begin{aligned} C_1 = {} & \$1{,}920 + \$10 \cdot f \cdot p_1 \cdot r_1 \cdot s_1 + [(\$8.14 + \$2 \cdot TCH_1) \cdot p_1 + \$4 \cdot CCS_1] \cdot r_1 \cdot s_1 \\ & + \$0.30 \cdot k \cdot f \cdot t_1 \end{aligned} \tag{3.11}$$

for 2 and more rows plots being used, and:

$$C_1 = \$1{,}920 + \$10 \cdot f \cdot p_1 \cdot r_1 \cdot s_1 + [(\$8.14 + \$2 \cdot TCH_1) \cdot p_1 + \$2 \cdot CCS_1] \cdot r_1 \cdot s_1 + \$0.30 \cdot k \cdot f \cdot t_1 \quad (3.12)$$

for one row plots.

Assuming that at least 2 row plots were used, the cost for stages $z \geq 2$ is calculated using the following function:

$$C_z = \$960 + \$10 \cdot k_z \cdot p_z \cdot r_z \cdot s_z + \{\$480 + [(\$4.07 + \$2 \cdot TCH_z) \cdot p_z + \$4 \cdot CCS_z] \cdot r_z \cdot s_z\} \cdot (1 + m_z) \quad (3.13)$$

where k_z is the number of genotypes entering stage z, CCS_z ie TCH_z is a boolean function that equals 0 or 1 depending whether CCS ie TCH needs to be measured at stage z, and m_z is the number of ratoon crops grown at stage z.

3.4 Examination of some basic results from the SSSM when applied to the Burdekin region

A selection system (Table 3.3) similar to that which that has been routinely carried out in the past by BSES and CSR (Section 1.1.1) was simulated. This selection system was defined with the following selection variables:

- The starting population comprised 200 families ie $f = 200$, and 60 genotypes per family ie $k = 60$.
- In stage $z = 1$ plot size $p_1 = 1$, number of sites $s_1 = 1$, number of replicates per site $r_1 = 4$, selection index $d_1 = TSH$, and selection intensity $t_1 = 30\%$.
- In stage $z = 2$, the crop is regrown in the ratoon crop to allow for the selection of individual genotypes from within each selected family thus, selection index $d_2 = visual.selection$ and selection intensity $t_2 = 50\%$ defines the stage.

- In stage $z = 3$, plot size $p_3 = 1$, number of sites $s_3 = 1$, number of replicates per site $r_3 = 2$, selection index $d_3 = TSH$, and selection intensity $t_3 = 5\%$.
- In the final stage $z = 4$, plot size $p_4 = 4$, number of sites $s_4 = 4$, number of replicates per site $r_4 = 2$, selection index $d_1 = TSH$, and selection intensity $t_4 = 20\%$.

Each effect of the linear model (2.1) that determined the phenotypic expression of CCS and cane yield was randomly generated, so the genotypes generated in the simulation exhibit some level of random variability. A statistical analysis of model output was performed to gauge the level of variability in genetic gains arising. After ensuring that the model is generating populations with variances as specified, the performance of the SSSM was tested by ensuring that the SSSM obtained genetic variance of the population of genotypes selected matched the genetic variances observed in real breeding programs, as estimated in Section 2.3.2.

The simulation was repeated one hundred times, and the data from each simulation was collected. A sample from one of these simulations is given in Appendix B.

Table 3.3

Description of the selection system simulated on SSSM, where f is the starting number of families, k is the number of genotypes per family to start selection, z is the stage number, p_z the plot size used at stage z, s_z number of sites and r_z number of replicates per site used at stage z, and d_z and t_z selection index and intensity respectively used at stage z

	$f = 200$			$k = 60$	
z	p_z	s_z	r_z	d_z	t_z
1	1	1	4	Sugar yield	30%
2				Visual selection	50%
3	1	1	2	Sugar yield	5%
4	4	4	2	Sugar yield	20%

3.4.1 Variation between simulations

To establish how many times each selection system needs to be simulated to allow a decision of whether one selection system is superior to another to be made, the genetic gains for economic value $\widetilde{G}$ were also gathered from the one hundred simulations of the selection system described by Table 3.3. The average genetic gain for economic value $\widetilde{G}$ obtained equals \$3372.86 with a standard deviation of \$374.01. These were used to calculate the expected standard errors for one, two, five, ten, one thousand simulations. Figure 3.6 shows how the standard error for genetic gain in economic value varies for different number of simulations. The magnitude of the standard error decreases sharply from one to ten simulations and then continues decreasing at a much slower rate. It was concluded that for the purposes of identifying an optimal selection system in this study, running the simulation model one hundred times would provide an appropriate balance between computational efficiency and desired accuracy to detect differences between alternative selection systems.

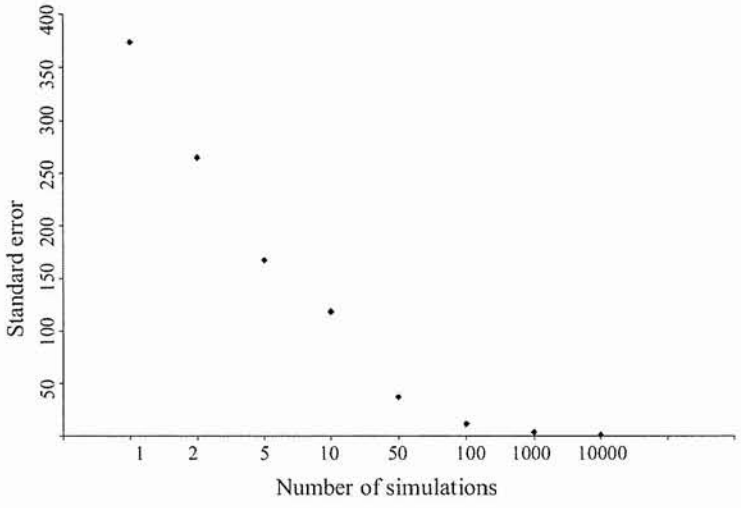

Figure 3.6. The change in the expected standard errors for the genetic gain for economic value $\widetilde{G}$ with the change in the number of simulations

To detect the minimal difference in the genetic gain for economic value of ΔG between two selection systems the following formula was used:

$$N \geq \frac{2s^2}{\Delta G^2}\left(t_{\alpha,\phi} + t_{\beta(1),\phi}\right)^2 \qquad (3.14)$$

where N is the sample size ie the number of simulations to be run, s^2 is sample variance which is assumed to be the population variance where the population comprises the set of 100 genetic gains for economic value simulated, $t_{\alpha,\phi}$ and $t_{\beta(1),\phi}$ are the critical values for the t-test with the significance level α and the power of the test $1-\beta$ and ϕ degrees of freedom (Cochran and Cox, 1975). Note that to calculate the degrees of freedom ϕ the sample size N is required. This is estimated using the above formula.

The above formula could be used to establish the sample size that will allow a 90% chance of detecting a true difference of at least \$50 ($\Delta G = \$50$) between gains of two selection systems at 0.05 significance level. The standard deviation of the one hundred simulations equals \$374.01. To find the critical values $t_{\alpha,\phi}$ and $t_{\beta(1),\phi}$ the sample size of 100 was first assumed. Then $\phi = 2(n-1) = 198$, $t_{0.05(2),198} = 1.972$, $t_{0.1(1),198} = 1.286$, and $N \geq 1281.51 \cong 1282$. Use now $N = 1282$ to get $\phi = 2(n-1) = 2562$, $t_{0.05(2),2562} = 1.960$, $t_{0.1(1),2562} = 1.282$, and $N \geq 1268.96 \cong 1269$. Therefore, each selection system should be simulated at least 1269 times to have a 90% chance of declaring two selection systems that differ by \$50 in the genetic gain for economic value, at $P < 0.05$. Assuming that one simulation of an average selection system lasts only one minute, it would take over 21 hours to give an estimate of the $\widetilde{G}$ value that would detect this minimal difference. However, simulating each selection system 1269 times or more would decrease the computer time efficiency of the simulation model SSSM.

Alternatively, the above formula (3.9) could be manipulated to give an expression for the minimal difference ΔG given the number of simulations N:

$$\Delta G \geq \sqrt{\frac{2}{N}}s\left(t_{\alpha,\phi} + t_{\beta(1),\phi}\right) \qquad (3.15)$$

Using the above formula, and the critical values $t_{0.05(2),198} = 1.972$ and $t_{0.1(1),198} = 1.286$ for $N = 100$, it was determined that when one hundred simulations were run, the minimal detectable difference ΔG of \$178.99 could be achieved. To achieve therefore, approximately a third of the above minimal detectable difference twelve times more computer time needs to be used. Consequently, each selection system was simulated one hundred times and the $\widetilde{G}$ value that gives an indication of its performance was taken to be the average of these simulations.

3.4.2 Sensitivity analysis of the SSSM

The variance components (Table 2.9) used throughout the study were decided upon based on two different sources: breeders' views and the estimates from two data sets (Section 2.4). For those variance components for which the two sources agree there is more certainty about their actual values than for others. However, by identifying variance components that contribute significantly to the magnitude of $\widetilde{G}$ would determine the subset of variance components that needs to be estimated more accurately in future.

Thus, to establish the robustness of the simulation model SSSM to any change in the initial set of parameters (Table 2.9) a screening design was applied. Taking a reduced and enlarged value for each of the initial values (Table 2.9), and then changing a parameter at the time the effect such change could have on the SSSM results could be observed. To get a complete picture of any possible sensitivity all possible combinations of the three values needs to be tested, ie the full factorial design needs to be performed, which would also identify any possible interaction between different factors.

However, the number of experiments in a full factorial design increases exponentially with the number of factors present in the design. In this case it would not be feasible to run a full factorial design, since it would require comparing 3^{24} selection system designs, twenty-four factors (Table 2.9), at three levels. Thus, to reduce the amount of experiments required a fractional factorial experimental design (Box *et al*, 1978) was performed.

Fractional factorial design eliminates redundant combinations of parameters ie factors by using only a fraction (eg. half or quarter) of the full factorial designs. The choice of designs to be included in a fractional factorial design is not random but are carefully chosen so that they capture all relevant information. The experimental software NEMROD-W version 9901 developed by Mathieu, D., Nonnu, J and Phan-Tan-Luu, R from the Laboratoire de Méthodologie de la Recherche Expérimental de l'Universite d'Aix-Marseille (LPRAI), Marseille, France, was used to obtain the experimental matrix as well as to analyse results.

To provide a judgement about the relative confidence intervals (Table 3.4) of the initial statistical parameters (Table 2.9), two plant breeders (Phillip Jackson and Mike Cox) involved in conducting sugarcane selection programs were contacted. They suggested the following parameters are associated with a high level of uncertainty: the genetic variance of the starting population σ_g^2, the correlation between genetic effect and competition effect $\rho_{g,c}$, the proportion of genetic variation retained between (rather than within) families δ, GE interaction variance σ_x^2 expressed through the σ_x^2/σ_g^2 ratio; error variance σ_e^2 in single seedling and one row plot; and genetic correlation between performance in pure stand and performance in plots $\rho_{g,g+c}$ for plot sizes comprising single seedlings, one row and two rows, and they nominated likely lower (1) and upper (3) limits for each of these parameters (Table 3.4). It was expected that the genetic correlation $\rho_{g,g+c}$ estimates for all three plots would co-vary in a similar way. Therefore, the genetic correlation $\rho_{g,g+c}$ between performances in small plots and that in pure stands, which is a reflection of the importance of competition variance, could be reasonably considered as a single factor across all plot sizes rather than an independent variable for each plot size.

The factorial experimental matrix for fourteen factors at three levels obtained from NEMROD-W software is given in Appendix C. The selection system defined in Section 1.1.1 was simulated one hundred times (Section 3.4.1) for each of the eighty-one (Appendix C) combinations of factors.

Table 3.4

Summary of the factors relevant for the screening design for CCS and cane yield (TCH), at three levels (1, 2 and 3), 2 being the initial original value (Table 2.9), used in the factorial experimental designs: the genetic variance σ_g^2, correlation between genotypic value and competition $\rho_{g,c}$, proportion of variation between families δ, σ_x^2/σ_g^2 ratio; error variance σ_e^2 in single seedling and one row plot; and genetic correlation between plots $\rho_{g,g+c}$ in the three plot sizes considered as a single factor

	CCS								
							$\rho_{g,g+c}$		
Level	σ_g^2	$\rho_{g,c}$	δ	σ_x^2/σ_g^2	σ_e^2 seedling	σ_e^2 one row	$\rho_{g,g+c}^s$ seedling	$\rho_{g,g+c}^1$ one row	$\rho_{g,g+c}^2$ two row
1	1.44	0.00	0.15	0.20	1.44	1.44	0.70	0.78	0.83
2	3.24	0.20	0.30	0.30	3.24	2.43	0.80	0.88	0.93
3	5.76	0.40	0.45	0.80	12.96	5.76	0.90	0.98	1.00
	TCH								
							$\rho_{g,g+c}$		
Level	σ_g^2	$\rho_{g,c}$	δ	σ_x^2/σ_g^2	σ_e^2 seedling	σ_e^2 one row	$\rho_{g,g+c}^s$ seedling	$\rho_{g,g+c}^1$ one row	$\rho_{g,g+c}^2$ two row
1	225.00	0.00	0.15	0.20	900.00	506.25	0.30	0.42	0.51
2	506.25	0.20	0.30	0.30	2025.00	1406.25	0.40	0.52	0.61
3	900.00	0.40	0.45	1.00	3600.00	2916.00	0.60	0.72	0.81

Table 3.5 summarises the results of the analysis of the experimental data given in Table 3.4. For each of the analysed factors two coefficients and associated significance levels are given: (1) one measuring the effect a change from level 3 to level 1 in the magnitude of factors has on the SSSM model; and other (2) measuring that of change from level 3 to level 2. The coefficient of a factor measures the average increase or decrease in the genetic gain for economic value $\widetilde{G}$ of the selection system when the magnitude of the particular factor goes from one level to another. To identify factors and their levels which change would significantly affect the magnitude of $\widetilde{G}$, the observed significance level p based on test statistics was used. Those effects for which there is enough evidence to suggest that the change in their magnitude from one level to another affects the $\widetilde{G}$, have their observed significance level p represented with stars, where three

starts indicate that claim could be accepted with $\alpha \leq 0.01$, two stars indicate $0.05 \leq \alpha < 0.01$ and one star indicate $0.1 \leq \alpha < 0.05$.

There was thus, enough evidence to suggest that any change in the genetic variance σ_g^2 in both CCS and cane yield, regardless of whether it changed from 5.76 to 1.44 or from 5.76 to 3.24 (Table 3.4), significantly impacted the gain $\widetilde{G}$ with $\alpha \leq 0.01$ (Table 3.5). When genetic variance σ_g^2 in CCS changed from 5.76 to 1.44 (Table 3.4) the $\widetilde{G}$ decreased from the average of \$5,368.42 for \$1,971.46, whereas the change in σ_g^2 from 5.76 to 3.24 (Table 3.4) decreased the $\widetilde{G}$ for \$1012.77 (Table 3.5). For cane yield (TCH) the effect of similar change was halved, so the decrease in σ_g^2 from 900.00 to 225.00 (Table 3.4) decreased the $\widetilde{G}$ on average \$1026.12, whereas the decrease from 900.00 to 506.25 (Table 3.4) resulted in a decrease in $\widetilde{G}$ of an average \$487.59 (Table 3.5). Both situations for CCS and cane yield were expected, as more genetic variability in starting population would bring higher chances of potentially superior genotypes being contained within such population and thus would increase expected yield $\widetilde{G}$.

There was also enough evidence to suggest that the change in magnitude of the correlation between genotypic value and competition $\rho_{g,c}$ in cane yield significantly affected the gain $\widetilde{G}$. When $\rho_{g,c}$ in cane yield changed from 0.4 to 0 (Table 3.4) the significance level of the impact on $\widetilde{G}$ was $\alpha \leq 0.01$ (Table 3.5), and when it changed from 0.4 to 0.2 (Table 3.4) its impact was as expected marginally smaller with $0.1 \leq \alpha < 0.05$ (Table 3.5). This indicates that the greatest impact was whether there is any correlation between genotype and competition $\rho_{g,c}$ in cane yield, with its impact decreasing with the increase in the correlation.

Table 3.5

Summary of the results of the factorial experimental design for each of the fourteen factors each represented with two measurements: one measuring the change from level 3 to 1 and other from level 3 to 2: the genetic variance σ_g^2, correlation between genotypic value and competition $\rho_{g,c}$, proportion of variation between families δ, error variance σ_e^2 in single seedling and one row plot, σ_x^2/σ_g^2 ratio; and genetic correlation between plots $\rho_{g,g+c}$ in the three plot sizes considered as a single factor

Factor		Coefficient	t.exp.	P	Coefficient	t.exp.	P
		5368.42	75.25	***			
			CCS			TCH	
σ_g^2	3-1	-1971.46	-60.75	***	-1026.12	-31.62	***
	3-2	-1012.77	-31.21	***	-487.59	-15.03	***
$\rho_{g,c}$	3-1	-26.49	-0.82	42.30%	-250.87	-7.73	***
	3-2	-42.24	-1.30	19.60%	-83.65	-2.58	*
δ	3-1	-37.05	-1.14	25.80%	47.86	1.48	14.20%
	3-2	-9.84	-0.30	76.10%	31.56	0.97	33.70%
σ_e^2	3-1	-5.10	-0.16	87.00%	7.36	0.23	81.60%
Seedling	3-2	-14.86	-0.46	65.30%	-7.81	-0.24	80.60%
σ_e^2	3-1	-9.09	-0.28	22.30%	19.26	0.59	43.70%
1 row plot	3-2	36.28	1.12	73.20%	20.18	0.62	45.60%
σ_x^2/σ_g^2	3-1	94.76	2.92	**	18.97	0.58	43.10%
	3-2	77.83	2.40	*	-12.04	-0.37	28.70%
$\rho_{g,g+c}$	3-1	-76.17	-2.35	*	-227.28	-7.00	***
	3-2	-41.25	-1.27	79.30%	-115.81	-3.57	***

The occurrence and the change in the magnitude of the same correlation in CCS however, had no significant impact on the genetic gain for economic value (Table 3.5). In fact the change from 0.40 to 0.00 (Table 3.4) for the correlation between genotype and competition effects, $\rho_{g,c}$ in CCS brought a decrease in $\widetilde{G}$ of an average \$26.49, whereas the decrease in $\rho_{g,c}$ from 0.40 to 0.20 (Table 3.4) brought a decrease in $\widetilde{G}$ of an average \$42.24 (Table 3.5). For cane yield the effect the change in $\rho_{g,c}$ had on the magnitude of $\widetilde{G}$ was even greater, so the change from 0.40 to 0.00 (Table 3.4) brought a decrease in $\widetilde{G}$ of an average \$250.87, and the change from 0.40 to 0.20 (Table 3.4), \$83.65 (Table 3.5).

The magnitude of $\widetilde{G}$ is sensitive to the change in the genetic correlation between plots $\rho_{g,g+c}$ in cane yield with the significance level of $\alpha \leq 0.01$ (Table 3.5) regardless of whether the change is from the level 3 to the level 1 or from the level 3 to the level 2 (Table 3.4). However, for CCS sensitivity is restricted to that from the level 3 to the level 1 (Table 3.4) with $0.1 \leq \alpha < 0.05$ (Table 3.5). For CCS, which already has a relatively high genetic correlation $\rho_{g,g+c}$ (0.8 or 0.9) between pure stand and smaller plot sizes (Table 2.9), the change in the correlation $\rho_{g,g+c}$ does not affects the gain $\widetilde{G}$ as dramatically as it does for cane yield for which the genetic correlation $\rho_{g,g+c}$ is marginally smaller (0.4 and 0.5) (Table 2.9). Thus the possibility of a greater predictability of the cane yield performance in pure stand inevitably influences the overall selection performance through the change in the gain $\widetilde{G}$. Generally, when the genetic correlation between performance in a pure stand and performance in a small plot, $\rho_{g,g+c}$ for cane yield, changed from level 3 (which is 0.60 between single seedling and pure stand, 0.72 between one row plot and pure stand, and 0.81 between two row plot and pure stand) to level 1, (defined by 0.30, 0.42 and 0.51), the $\widetilde{G}$ decreased on average by \$227.275 (Table 3.5). Similarly, when the genetic correlation between plots $\rho_{g,g+c}$ for cane yield in three plot sizes changed from level 3 (defined by 0.60, 0.72 and 0.81) to level 2 (defined by 0.40, 0.52 and 0.61) the $\widetilde{G}$ decreased on average by \$115.813 (Table 3.5). In CCS, on the other hand the average decrease in $\widetilde{G}$ was smaller at \$76.17 when $\rho_{g,g+c}$ decreased from 0.90, 0.98 and 1 to 0.70, 0.78 and 0.83 (Table 3.4) for the three plot sizes; and decreased by \$41.25 (Table 3.5) when $\rho_{g,g+c}$ decreased from 0.90, 0.98 and 1 to 0.80, 0.88 and 0.93 (Table 3.4).

The change in the σ_x^2/σ_g^2 ratio from 0.8 to 0.2 (Table 3.4) for CCS significantly affected the gain $\widetilde{G}$ with $0.05 \leq \alpha < 0.01$ (Table 3.5); and from 0.8 to 0.3 (Table 3.4) with $0.1 \leq \alpha < 0.05$ (Table 3.5), whereas for cane yield the change in σ_x^2/σ_g^2 did not affect the gain $\widetilde{G}$ significantly. The decrease in the σ_x^2/σ_g^2 ratio from 0.80 to 0.20 (Table 3.4) in CCS brought an increase in $\widetilde{G}$ of \$94.76, and similarly the change from 0.80 to 0.30 (Table 3.4) brought an increase of \$77.83 (Table 3.5). In cane yield on the other hand, the effect was much smaller as the change from 1.00 to 0.20 (Table 3.4)

brought an increase of $18.97, and the change from 1.00 to 0.30 (Table 3.4) brought an increase of $12.04 (Table 3.5).

3.5 Application and limitations of the simulation model

The simulation model SSSM may be used independently to evaluate different selection systems as designed by its user, but is limited as an optimisation technique. The model however may be used to compare particular selection designs to determine how a change in any selection variable (Section 1.1.2) and/or variance components (Section 2.2) could impact the magnitude of the $\widetilde{G}$. It allows analysis of each stage separately as well as analysis of all stages together as part of a coherent system of selection. Furthermore, because a single value $\widetilde{G}$ may be assigned to each selection system design, as a measure of its effectiveness, the simulation model SSSM provides a basis for accurate comparison of alternative selection systems and subsequently, optimisation of all design variables. The simulation model SSSM could be easily updated to allow selection system prediction in other regions in the case of sugarcane or to other crop species, by changing the parameters to be representative of the changed situation.

On the other hand, the limitations of the SSSM are that it does not allow testing of the ratoon crops performance neither does it allow selecting for traits other than CCS and cane yield. Another limitation of the SSSM is that it does not allow testing in 5-meter long plots, the plot length often used in sugarcane selection trials, as row plots for which variance components have been estimated are all 10 meters long: one 10-meter long plot, two 10-meter long plots, and four 10-meter long plots.