2009

Statistical methods and models for analyzing sugarcane (Saccharum species hybrids) plant breeding data

Marvellous Mabeza Zhou
Louisiana State University and Agricultural and Mechanical College, marvzhou@yahoo.com

STATISTICAL METHODS AND MODELS FOR ANALYZING SUGARCANE (*SACCHARUM* SPECIES HYBRIDS) PLANT BREEDING DATA

A Dissertation

Submitted to the Graduate Faculty of the
Louisiana State University and
Agricultural and Mechanical College
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

In

The School of Plant, Environmental and Soil Sciences

by
Marvellous Mabeza Zhou
B.Sc. Agriculture Honors (Crop Science), University of Zimbabwe, 1989
M.Sc. Agriculture, University of Natal, South Africa, 2003
Masters of Applied Statistics, Louisiana State University, 2008

August, 2009

DEDICATION

This Dissertation is dedicated to my son, **Tapiwanashe**.

ACKNOWLEDGEMENTS

I wish to express my gratitude and appreciation to my major professor, Dr Collins Kimbeng, my mentor whose advice, guidance, willingness to help and the freedom he gave me to learn was a great inspiration. I would like to acknowledge the help from my graduate committee Dr Kenneth Gravois, Dr Kevin McCarter and Dr Gerald Myers who spared time from their busy schedules to help me. The comments from the Graduate School representatives Dr Michael Saska (general examination) and Dr Roberto Barbosa (final examination) are acknowledged. I am grateful to Dr Bill White, USDA-ARS, for allowing me to use his data for my dissertation, for his advice, comments and help with chapter 6 and Dr Tom Tew for allowing me to collect data from his trial and help with data collection for chapters 2, 3 and 4. Help with data collection from staff at the USDA-ARS, Houma and LSU AgCenter, St. Gabriel is greatly appreciated.

Special thank you goes to Dr Freddie Martin (Director, School of Plant, Environment and Soil sciences) who guaranteed the assistantship that supported my studies and linked me with Dr Kimbeng. Profound gratitude goes to Mr. Karl Nuss (Head: SASRI Plant Breeding), who encouraged me to pursue a PhD and helped establish contact with Dr Martin. The support Dr Muntubani Nzima (ZSAES Director, who encouraged me to pursue a PhD in the USA) and the ZSAES board of directors (granted study leave and financed travel and settling costs) are acknowledged. The American Sugar Cane League is thanked for providing the fellowship that supported my stipend.

I would like to acknowledge the help from my fellow graduate assistants in the sugarcane lab (Sreedhar Alwala and Suman Andru) for their help in getting me settled and teaching me the molecular lab techniques. Special thanks go to Samuel Ordonez Jr. for being a great friend and for just listening during those stressful times. Thanks also go Nkosinathi Dhlamini, for always

dragging me to the gym and jogging sessions, and being my most wonderful roommate and being always positive and encouraging. I would like to thank my internet friends for their encouragement and making me realize it was a matter of time. I would like to thank my colleagues and staff at ZSAES and the Zimbabwe sugar industry for their goodwill and encouragement.

I would like to thank my sister and my bother in law (Mr. and Mrs. Dube) for taking in my son, Tapiwanashe as their own during the course of my studies. It is their unconditional love for my son that I will cherish forever. I would like to acknowledge the support of my mother (Mrs. Tendai Zhou) for always having me in her prayers and always encouraging me to aim for the best. Special tribute goes to my late father, whose passion for education was a great inspiration, and wherever he is, he must be proud of this achievement. I would like to thank my brothers and sisters, my brother in laws and sister in laws, and my nieces and nephews for their support.

Special thanks go to my son, Tapiwanashe for his unconditional love and never, complaining at least directly to me for my absence. His maturity and always asking how my studies were going was my greatest inspiration, I could never afford to fail. It is to Tapiwanashe that I dedicate this work and I hope this work inspires Tapiwanashe to realize that the sky should always be the only limit.

Finally and most importantly, I would like to acknowledge and thank the Lord God Almighty, without whom I would not have been able to embark on or even complete these studies. If there be any glory arising from this work, let it be directed to thee Him, the Creator and Sustainer of all things.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Early generation selection of sugarcane families using means is inadequate while visual seedling selection is subjective and inefficient. Data from advanced variety trials (yield, quality and agronomic traits) are collected over several crop-years to determine yield potential and ratooning ability of genotypes follow a multivariate repeated measures structure. In Louisiana, the sugarcane borer and recently the Mexican rice borer are major insect pests of sugarcane. Both borers have similar feeding habits, providing an opportunity for investigating if genotypes resistant to one species would provide resistance to the other (cross-resistance). The objectives of the study were to identify statistical methods to evaluate family yield potential and repeatability, enhance seedling selection for yield, analyze advanced variety trials data and prove cross resistance between the sugarcane borer and the Mexican rice borer.

Random coefficient models (RCM) identified elite families with higher cane yield potential and higher repeatability between seedlings and clones. These elite families comprised a larger proportion of higher yield seedlings that produced high yielding clones. Logistic regression models (LRM) provided an objective statistical decision support tool for selecting high yielding seedlings and were more flexible at adjusting the number of seedlings to advance than visual selection. The LRM can be used to identify important traits in breeding populations as well as directionally shifting population trait values during selection. Neural network models can be used to automate the LRM. The multivariate repeated measures analysis (MRM) reduced Type I errors associated with univariate analysis by including covariance to compute experimental errors. The MRM showed greater statistical differences among genotypes for yield traits than univariate analysis. Cross resistance between the sugarcane and Mexican rice borer

was proved using log linear models, and using a population with known sugarcane borer resistance status.

Using RCM will significantly increase the efficiency of early generation selection by identifying families with high yield potential and repeatability while LRM will increase efficiency of identifying high yielding seedlings from these elite families. MRM will increase the accuracy of evaluating genotypes for yield and ratooning ability. Cross-resistance will allow breeders to take advantage of parents from the sugarcane borer recurrent selection program.

CHAPTER 1: GENERAL INTRODUCTION

Sugarcane improvement through plant breeding started around 1888 after the observation in 1858 of viable seeds (Stevenson, 1965). Up until then, sugarcane was cultivated vegetatively from noble canes (*Saccharum officinarum*) (James, 2004). The *S. officinarum* varieties, the noble canes, were highly susceptible to diseases and therefore plant breeding started as an attempt to develop resistant varieties (Heinz, 1987).

Sugarcane (*Saccharum* spp. hybrid) is a crop for which interspecific hybridization has provided a major breakthrough in its improvement (Berding *et al*., 2004). Modern sugarcane cultivars were derived from the interspecific hybridization between two major *Saccharum* species, namely *S. officinarum* and *S. spontaneum,* in the early 1900s (Price, 1963). *S. officinarum* was the primary source of genes for sucrose accumulation whereas *S. spontaneum* contributed genes for general adaptability and high biomass, but also contributed unfavorable attributes relating to sugar quality (Roach, 1986).

Significant achievements towards increasing cane and sugar yield (Hogarth and Berding, 2006; Milligan *et al*., 1994; Nuss, 2001; SASRI, 2007a, b; Zhou, 1996, 2004), disease resistance (Bailey, 2004; Walker, 1987; Zhou, 1996, 2004), insect resistance (Leslie, 2004; White *et al*., 1996), and stress tolerance (Moore, 1987) have occurred across the world through sugarcane breeding. Recently, there have been reports of a sugarcane yield plateau in sugarcane in Australia, South Africa and other sugarcane breeding programs (Garside *et al*., 1997). Horgath and Berding (2006), and Butterfield and Ulian (2006) have advocated new innovations and approaches to break the yield plateau and create opportunities for further advances in sugar yield.

This study focused on introducing and demonstrating statistical methods and models for improving early generation selection, analyzing data from advanced variety trials and

determining if cross resistance exist between the sugarcane borer and the Mexican rice borer in sugarcane breeding populations.

1.1 Early Generation Selection

Early generation selection involves identifying sugarcane families (elite families) comprising a greater proportion of high cane yielding seedlings followed by selecting individual seedlings with the potential to produce high cane yield from these elite families. Genotype-by-environment (GE) interaction effects and the competition among closely spaced seedlings as well as clones planted in small plots (Jackson and McRae, 2001; Tovey *et al*., 1973) are known to influence the precision of individual seedling selection during early generation selection (Skinner *et al*., 1987). GE interaction effects are known to be particularly important for traits controlled by quantitative genes such as cane yield (Jackson and Horgath, 1992; Jackson and McRae, 1998; Falconer and Mackay, 1996; Kang and Miller, 1984; Kimbeng *et al*., 2002, 2009; Mirzawan *et al*., 1993). Reducing GE interaction effects in early generation selection through replicating plots may not be possible in all cases because of limited planting material, shortage of land and the cost of planting large numbers of genotypes in multiple plots. Therefore, an investigation of other approaches to reduce the impact of GE interaction effects such as using statistical methods and models that account for the effects and also using decision support tools that reduce the subjectivity in selection is warranted.

Marker assisted selection is unlikely to have an impact on selecting for traits controlled by quantitative genes such as cane yield in the immediate future (Bernado, 2008; Xu and Crouch, 2008; Heffner *et al*., 2009). In the short to medium term, improving the current selection methods remains one of the most promising options available to breeders for increasing selection

efficiencies (Hogarth and Berding, 2006). Statistical methods that offer easy computations can be used as decision support tools provide the greatest potential for improving family evaluation and individual seedling selection in sugarcane breeding programs.

1.1.1 Family Evaluation

Research done in Australia has proven that family selection was superior to seedling selection (Hogarth, 1971) particularly for traits with low heritability such as cane yield (Jackson and McRae, 1998; Falconer and Mackay, 1996). Family selection occurs when all the seedlings in a family are selected or rejected based on their family means (Falconer and Mackay, 1996). The selected families are expected to produce a higher proportion of seedlings producing high cane yield (Kimbeng and Cox, 2003; Cox and Hogarth, 1993). The proven cross system is a family evaluation method that uses the proportion of advanced seedlings and the performance of varieties from each cross to define the value of a family (Skinner *et al*, 1987). It has been widely used in Australia, South Africa (Heinz and Tew, 1987; Skinner *et al*., 1987) and several other sugarcane breeding programs. The proven cross system focuses on old crosses to the exclusion of new ones. Family evaluations using means and the proven cross system have proved to be inadequate. The number of high yielding seedlings recovered from some of the elite families were found to deviate significantly from expectations based on family means and the proven cross system advancement numbers (Kimbeng *et al*., 2000; Skinner *et al*., 1987). Therefore the investigation of other approaches to improving family evaluation and selection is required.

The first stage of a sugarcane selection program involves the evaluation of clones as single plants grown from true seed (Jackson and McRae, 2001). Subsequent stages and the commercial crops are planted from vegetative material, creating a potential confounding due to

the seed type between the seedling and clonal stages. Repeatability between seedlings and clones has been shown to be significant indicating that seedlings can be used to predict the performance of clones (Hogarth, 1971; Cesnik and Venkovsky, 1974; Ladd *et al*., 1974; Marriotti, 1974, 1977; James and Miller, 1975; Miller and James, 1975; Kang *et al*., 1983; Bressiani *et al*., 2003). To date, repeatability between seedlings and clones is not directly used in most breeding programs and continues to be overlooked during family evaluation. One of the reasons for not using the repeatability between seedlings and clones in family selection could be the unavailability of appropriate statistical methods that are adapted to the use of repeatability to evaluate families.

Random coefficient models (RCM) are statistical methods developed from the analysis of covariance (ANCOVA) (Bryk and Raudenbush, 1992). The RCM analysis offers the potential for evaluating families for both their yield potential and repeatability. Repeatability of each family would be approximated by the slope of the association between the cane yield of the seedlings and clones. We hypothesize that there could be variability for repeatability among families and this variability can be used to select for those families that produce higher repeatability and also have higher yield potential. If this hypothesis is true, then the families could be compared for their repeatability, and the repeatability comparisons could be used as a proxy to predict and compare the trends in the distribution of yield between the seedlings and clones.

1.1.2 Seedling Selection

In sugarcane seedling selection, one of the greatest challenges facing sugarcane breeders is the correct identification of seedlings with the potential to produce high cane yield. Competition

effects are known to be large for seedlings planted in small plots (Jackson and McRae, 2001; Milligan *et al*., 2007) yet because of lack of resources and space, close spacing continues to be used. Visual selection, the primary method that is currently used for individual seedling selection is largely subjective (Cox and Stringer, 1998) and has proved to be inefficient (Hogarth and Berding, 2006). The confounding influence on seedling performance caused by the effects of genotype by environment interaction and the competition among seedlings planted in close spacing further reduces the precision of visual selection.

Path coefficient analysis studies in sugarcane proved that there was strong and significant influence of stalk number, stalk height and stalk diameter (the cane yield components) on cane yield (De Sousa-Vieira and Milligan, 2005; Kang *et al*., 1983, 1989; Milligan *et al*., 1990). Yield components are rarely measured in most sugarcane breeding programs because of the cost and labor limitations required to make these measurements. The yield components are also probably considered too costly to measure partly because of the unavailability of appropriate statistical methods and models that would generate quick selection decisions using these yield components. If statistical methods that use the yield components for seedling selection are made available, and these methods produce significant gains in selecting seedlings that produce higher cane yield, then there would be an incentive to measure the yield components. One study investigated the utility of the logistic regression model as a potential decision support statistical tool for aiding individual seedling selection where the stalk number, stalk height and stalk diameter would be used as the predictor variables. A second study investigated the utility of the artificial neural network (ANN) model as a decision support tool for enhancing individual seedling selection.

1.2 Multivariate Repeated Measures Analysis of Data from Advanced Variety Trials

Data from sugarcane breeding advanced variety trials include several variables measured from each experimental unit (plot) every year for several years. These data are used to evaluate the potential of genotypes to produce high yield, high quality, desirable agronomic traits, and high ratooning ability (Gauch *et al*., 2008). Data for several variables measured from the same plot resemble a multivariate structure (Johnson and Wichern, 2002). Values of variables measured from the same plot are likely not independent because they are influenced by the same factors existing in the plot. Data of a variable measured from each plot over several years resemble repeated measures (Littell *et al*., 2002, 2005). These measurements are likely not independent because the sequential crop-years cannot be randomized to the experimental units. Therefore the analysis of data from advanced variety trial should account for the within plot correlation of the multiple variables measured (multivariate structure) and the correlation from one crop-year to another (repeated measures).

Currently, the univariate analysis method is used. Univariate analysis assumes split-plot in time as the experimental design. The univariate analysis approach also assumes independence among the variables measured from each plot and also assumes independence of measurements derived from each plot across years (Freund and Wilson, 2003). The multiple variables derived from each plot are likely to be correlated, thereby invalidating the assumption of independence. The measurements of a variable from each plot over several years are also likely to be correlated because the years are always sequential. The univariate analysis could therefore result in the likely violation of the assumption of independence. The violation of the assumption of independence may result in the underestimation or overestimation of the experimental errors used to test the effects. Significantly underestimating or overestimating experimental errors

would result in inaccurate tests and incorrect interpretation of the data. The ideal analysis should combine multivariate and repeated measures, to create multivariate repeated measures analysis. Multivariate repeated measures would account for both the correlation between variables and the correlations between crop-years.

1.3 Cross Resistance Between the Sugarcane Borer and the Mexican Rice Borer

Moth (Lepidoptera) stem borers are major pests of sugarcane (Smith *et al.,* 1993). In North America, two important stem borers are the crambids, *Diatraea saccharalis* (F.) (= sugarcane borer) and *Eoreuma loftini* (Dyar) (= Mexican rice borer). The sugarcane borer has been the dominant stem borer of sugarcane in the U.S.; however, in 1980, the Mexican rice borer became established in the Lower Rio Grande Valley of Texas (Johnson and van Leerdam, 1981) and subsequently supplanted the sugarcane borer as the dominant insect pest of that industry (Johnson 1984). In December 2008, the Mexican rice borer was identified in Louisiana as predicted by Reay-Jones *et al.* (2007).

The Mexican rice borer and the sugarcane borer are taxonomically closely related species, and share the same hosts but differ in their oviposition behavior. Once the first instar larvae eclose from the egg, the larvae of both species share similar feeding habits. The larvae move to the green leaf sheaths and begin feeding (Ring *et al*., 1991; White, 1993). They bore into the young, developing internodes. Larvae enter the stalk more quickly in susceptible varieties than resistant ones (White *et al*., 1996). This study hypothesized that, due to the similarities in larval behavior, particularly the feeding habits of the two borer species, selecting for resistance for one species will obtain resistance to the other, that is, cross resistance.

1.4 Objectives of the Study

The objectives of the dissertation research projects were:

1. To investigate the use of random coefficient models (RCM) as a tool for evaluating sugarcane families for cane yield potential and repeatability.

2. To investigate the potential of using the SAS enterprise miner (SAS Institute, 2007) artificial neural network (ANN) model as a decision support tool for identifying individual seedlings with high cane yield potential at the seedling stage of a sugarcane breeding program.

3. To demonstrate the use of the logistic regression models as a decision support statistical tool for individual seedling selection in sugarcane using the cane yield components namely, stalk number, stalk height, and stalk diameter as the independent or predictor variables. The study also evaluated the utility of the confidence intervals in enhancing decision making during the seedling selection process.

4. To demonstrate the use of multivariate repeated measures analysis of the linear mixed model as a tool for analyzing sugarcane breeding data from advanced variety trials. Specifically we determined the multivariate effects, and the appropriate covariance structure for analyzing the ratooning effects. The study also compared the univariate and multivariate repeated measures analyses for model fit and the ability to discriminate between the experimental and control genotypes.

5. To determine if cross resistance exist among sugarcane genotypes between two sugarcane pests, namely the Mexican rice borer and the sugarcane borer.

1.5 References

Bailey, R.A. (2004). Diseases. In: G. James (editor): Sugarcane: Second Edition. Blackwell Publishing: 54 – 77.

Berding, N., Hogarth, M. and Cox, M. (2004). Plant improvement of sugarcane. In: G. James (editor): Sugarcane: Second Edition. Blackwell Publishing: 20 – 53.

Bernardo, R. (2008). Molecular markers and selection for complex traits in plants: Learning from the last 20 years. *Crop Science* 48: 1649 – 1664.

Bressiani, J.A., Vencovsky, R. and da Silva, J.A.G. (2003). Repeatability within and between selection stages in a sugarcane breeding program. *Journal of the American Society of Sugarcane Technologists* 23: 40 – 47.

Butterfield, M.K. and Ulian, E.C. (2006). Breeding for a better industry: New breeding techniques. *Sugarcane International* 24 *Number* 2: 26 – 31.

Bryk, A.S. and Raudenbush, S.W. (1992). Hierarchical Linear Models. Newbury Park, CA, Sage Publications Inc.

Cesnik, R. and Venkovsky, R. (1974). Expected response to selection, heritability, genetic correlations and response to selection of some characters in sugarcane. *Proceedings of the International Society of Sugar Cane Technologists 15*: 96 – 101.

Cox, M.C. and Hogarth, D.M. (1993). Progress and changes in the South Queensland Variety Development Program. *Proceedings of the International Society of Sugarcane Technologists* 15: 251 – 255.

Cox, M.C. and Stringer, J.K. (1998). Efficacy of early generation selection in a sugarcane improvement program. *Proceedings of the Australian Society of Sugarcane Technologist* 20: 148 – 153.

De Sousa-Vieira, O. and Milligan, S. B. (2005). Interrelationships of cane yield components and their utility in sugarcane family selection: Path coefficient analysis. *Interciencia* Volume 30 number 2: 93 – 96.

Falconer, D.S. and Mackay, T.F.C. (1996). Introduction to Quantitative Genetics. Fourth Edition. Longman Group Ltd, UK.

Freund, R.J. and Wilson, W.J. (2003). Statistical Methods. Third Edition. Academic Press, New York.

Garside, A.L., Smith, M.A., Chapman, L.S., Hurney, A.P. and Magarey, R.C. (1997). The yield plateau in the Australian Sugar industry: 1970 – 1990. In Keating, B.A. and Wilson, J.R. (editors). Intensive Sugarcane Production: Meeting the Challenges Beyond 2000. CAB International, Wallingford, United Kingdom: 103 – 124.

Gauch, H.G., Piepho, H.P. and Annicchiarico, P. (2008). Statistical analysis of yield trials by AMMI and GGE: Further consideration. *Crop Science* 48: 866 – 889.

Heffner, E.L., Sorrells, M.E. and Jannink, J.L. (2009). Genomic selection for crop improvement. *Crop Science* 49: 1 – 12.

Heinz, D.J. (1987). Introduction. In: Heinz, D.J. (editor). Sugarcane Improvement through Breeding. Elsevier. New York.

Heinz, D.J. and Tew, T.L. (1987). Hybridization procedures. In: Heinz, D.J. (editor). Sugarcane Improvement through Breeding. Elsevier. New York.

Hogarth, D.M. (1971). Quantitative inheritance studies in sugarcane. II. Correlations and predicted responses to selection. *Australian Journal of Agricultural Research* 22: 103 – 109.

Horgath, D.M. and Berding, N. (2006). Breeding for a better industry: Conventional breeding. *Sugarcane International* 24 Number 2: 26 – 31.

Jackson, P.A., and Hogarth, D.M. (1992). Genotype x environment interactions in sugarcane. I. Patterns of response across sites and crop-years in North Queensland. *Australian Journal of Agricultural Research* 43: 1447 – 1459.

Jackson, P.A. and McRae, T.A. (1998). Gains from selection of broadly adapted and specifically adapted sugarcane families. *Field Crops Research* 59: 151 – 162.

Jackson, P.A. and McRae, T.A. (2001). Selection of sugarcane clones in small plots: Effects of plot size and selection criteria. *Crop Science* 41: 315 – 322.

James, G.L. (2004). An introduction to sugarcane. In: G. James (editor): Sugarcane: Second Edition. Blackwell Publishing: 1 – 19.

James, N.I. and Miller, J.D. (1975). Selection in six crops of sugarcane. II. Efficiency and optimum selection intensities. *Crop Science* 15: 37 – 40.

Johnson, H.J.R. (1984). Identification of *Eoreuma loftini* (Dyar) (Lepidoptera: Pyralidae) in Texas, 1980: forerunner for other sugarcane boring pest immigrants from Mexico? *Bulletin of the American Society of America* 30: 47 – 52.

Johnson, K.J.R. and van Leerdam, M.B. (1981). Range extension of *Acigona loftini* into the Lower Rio Grande Valley of Texas. *Sugar y Azucar* 76: 34.

Johnson, R.A. and Wichern, D.W. (2002). Applied Multivariate Statistical Analysis. Prentice Hall, New Jersey, USA.

Kang, M.S., Miller, J.D. and Tail, P.Y.P. (1983). Genetic and phenotypic path analysis and heritability in sugarcane. *Crop Science* 23: 643 – 647.

Kang, M.S., and Miller, J.D. (1984). Genotype x environment interactions for cane and sugar yield and their implications in sugarcane breeding. *Crop Science* 24: 435 – 440.

Kang, M.S., Sosa, O. and Miller, J.D. (1989). Path analysis for percent fiber, and cane and sugar yield in sugarcane. *Crop Science* 29: 1481 – 1483.

Kimbeng, C.A., McRae, T.A. and Stringer, J.K. (2000). Grains from family and visual selection in sugarcane, particularly for heavily lodged crops in the Burdekin region. *Proceedings of the Australian Society of Sugarcane Technologists* 22: 163 – 169.

Kimbeng C.A., A. R. Rattey, and M. Hetherington. 2002. Interpretation and implications of genotype by environment interactions in advanced stage sugarcane selection trials in central Queensland. *Australian Journal of Agricultural Research* 53:1035 – 1045.

Kimbeng, C.A. and Cox, M.C. (2003). Early generation selection of sugarcane families and clones in Australia: A review. *Journal of the American Society of Sugar Cane Technologists* 23: 20 – 39.

Kimbeng, C.A., Zhou, M.M. and da Silva, J.A. (2009). Genotype by environment interactions and resource allocation in sugarcane yield trials in the Rio Grande valley region of Texas. *Journal of the American Society of Sugarcane Technologists (Accepted/In press).*

Ladd, S.L., Heinz, D.J., Meyer, H.K. and Nishimoto, B.K. (1974). Selection studies in sugarcane (Saccharum spp. Hybrids). I. Repeatability between selection stages. *Proceedings of the International Society of Sugar Cane Technologists* 14: 102 – 105.

Leslie, G. (2004). Pests of sugarcane. In: G. James (editor): Sugarcane: Second Edition. Blackwell Publishing: 78 – 100.

Littell, R.C., Stroup, W.W. and Freund, R.J. (2002). SAS for Linear Models. Fourth Edition. SAS Institute Inc., Cary, NC, USA.

Littell, R.C., Milliken, G.A., Stroup, W.W. and Wolfinger, R.D. (2005). SAS System for Mixed Models. 7th edition. SAS Institute Inc., Cary, NC, USA.

Mariotti, J.A. (1974). The effect of environments on the effectiveness of clonal selection in sugarcane. *Proceedings of the International Society of Sugar Cane Technologists* 14: 89 – 95.

Mariotti, J.A. (1977). Sugarcane clonal selection research in Argentina: A review of experimental results. *Proceedings of the International Society of Sugar Cane Technologists* 14: 121 – 136.

Milligan, S.B., Balzarini, B., Gravois, K.A. and Bischoff, K.P. (2007). Early stage sugarcane selection using different plot sizes. *Crop Science* 47: 1859 – 1864.

Milligan, S.B., Gravois, K.A., Bischoff, K.P. and Martin, F.A. (1990). Crop effects on genetic relationships among sugarcane traits. Crop Science 30: 927 – 931.

Milligan, SB, Martin, FA, Bischoff, KP, Quebedeaux, JP, Dufrene, EO, Quebedeaux, KL, Hoy, JW, Reagan, TE, Legendre, BL and Miller, JD (1994). Registration of 'LCP85-384' sugarcane. *Crop Science* 34: 819 – 820.

Miller, J.D. and James N.I. (1975). Selection in six sugarcane crops. I. Repeatability of three characters. *Crop Science* 15: 23 – 25.

Mirzawan, P.D.N., Cooper, M. and Hogarth, D.M. (1993). The impact of genotype by environment interactions for sugar yield on the use of indirect selection in southern Queensland. *Australian Journal of Experimental Agriculture* 33: 629 – 638.

Moore, P.H. (1987). Breeding for stress resistance. 503 – 542. In DJ Heinz (editor). Sugarcane Improvement Through Breeding. Elsevier.

Nuss, K.J. (2001). The contribution of variety NCo376 to sugar production in South Africa from 1955 to 200 and its value as a parent in the breeding program. *Proceedings of the South African Sugar Technologists Association* 75: 54 – 59.

Price, S. (1963). Cytogenetics of modern sugarcanes. *Economic Botany* 17:97 – 105.

Reay-Jones, F.P.F., Wilson, L.T., Reagan, T.E., Legendre, B.L. and Way, M.O. (2007). Predicting economic losses from the continued spread of the Mexican rice borer (*Lepidoptera: Crambidae*). *Journal of Economic Entomology*, Volume 101 number 2: 237 – 250.

Ring, D.R., Browning, H.W., Johnson, K.J.R., Smith, Jr., J.W., and Gates, C.E. 1991. Age-specific susceptibility of sugarcane internodes to attack by the Mexican rice borer (Lepidoptera: Pyralidae). *Journal of Economic Entomology,* Volume 84 number 3: 1001-1009.

Roach, B. T. (1986). Evaluation and use of sugarcane germplasm. *Proceedings of the International Society of Sugar Cane Technologist* 17: 492 – 503.

SASRI Information Sheet Number 13.4 (2007a). Variety N14. 2 pages. http://www.sugar.org.za/sasri/varietyN14.pdf.

SASRI Information Sheet Number 13.18 (2007a). Variety N31. 2 pages. http://www.sugar.org.za/sasri/varietyN31.pdf**.**

Skinner, JC, Hogarth, DM and Wu, KK (1987). Selection methods, criteria and indices. 409 – 453. In DJ Heinz (editor). Sugarcane Improvement Through Breeding. Elsevier.

Smith, J.W., Jr., Wiedenmann, R.N. and Overholt, W.A. (1993). Parasites of lepidopteran stemborers of tropical gramineous plants. ICIPE Science Press, Nairobi, Kenya, 89 pages.

Stevenson, G.C. (1965). Genetics and Breeding of Sugar Cane. Longmans, London, 284 pages.

Tovey, D.A., Glasziou, K.T., Farquhar, R.H. and Bull, T.A. (1973). Variability in radiation received by small plots of sugarcane due to differences in canopy heights. Crop Science 13: 240 – 242.

Walker, D.I.T. (1987). Breeding for disease resistance. 455 – 502. In DJ Heinz (editor). Sugarcane Improvement Through Breeding. Elsevier.

White, W. H. 1993. Cluster analysis for assessing sugarcane borer resistance in sugarcane line trials. *Field Crops Res*. 33: 159-168.

White, W.H., Legendre, B.L., and Miller, J.D. 1996. Progress in breeding for sugarcane borer resistance. *Sugar Cane* 5: 3 – 7.

Xu, Y and Crouch, J.H. (2008). Marker assisted selection in plant breeding: From publications to practice. *Crop Science* 48: 391 – 407.

Zhou, M.M. (1996). The potential of new sugarcane varieties released by the ZSA Experiment Station in South East Lowveld of Zimbabwe. *Proceedings of the South African Sugarcane Technologists Association* 70: 111-113.

Zhou, M.M. (2004). Performance of varieties N14 and NCo376 in the South East Lowveld of Zimbabwe. *Proceedings of the South African Sugarcane Technologists Association* 78: 153 – 160.

CHAPTER 2: EVALUATING SUGARCANE FAMILIES FOR YIELD POTENTIAL AND REPEATABILITY USING RANDOM COEFFICIENT MODELS

2.1 Introduction

Selection is the cornerstone of plant breeding and is done across all stages of a sugarcane breeding program (Skinner *et al*., 1987). Although sugarcane is clonally propagated, the first stage of selection in sugarcane breeding programs involves the evaluation of clones as seedlings that are planted from true seed. Referred to as the Seedling Stage or Stage I, this is the only stage to be established from true seed and the seedlings are appraised either as individual seedlings or in family plots. The second stage of selection, Stage II, occurs when individual seedlings selected in Stage I are clonally (vegetatively) propagated and evaluated in clonal plots.

The seedlings grown from true seed are subjected to individual seedling selection for cane yield via its components and this selection process aims to predict the cane yield of clones that are grown from vegetative material harvested from the selected seedlings. This scenario creates confounding for seed type between the seedlings (Stage I) and clones (Stage II). Confounding could negatively impact the early selection stages. If confounding exists, some of the seedlings selected as high cane yield in Stage I may produce lower cane yield when planted as clones in Stage II. In sugarcane, confounding could occur when the seed type affects differently the genotype cane yield between the seedlings and clones. In crops established from true seed such as cotton, this confounding is less important. The effect of seed type could be caused by seedlings grown from true seed yielding differently compared to the same genotypes planted from vegetative material as clones. The confounding could also be a reflection of the effect of plot size between the seedlings and clones. The seedlings in stage I can only be represented by one stool transplanted as one seedling plant grown from one true seed. The clones in stage II are grown from vegetative material harvested from the seedlings. At maturity, each

seedling, also called a stool, would have produced several stalks due to the tillering developmental process, facilitating the planting of clones in larger plots. The tiller development process is known to differ significantly between genotypes and is significantly influenced by plot sizes and spacing (Zhou *et al*., 2003). All these factors could play an important role by contributing to the confounding that could exist between the yield of seedlings and clones.

Family selection involves the selection or rejection of whole families of seedlings based on information derived from family plots (Falconer and Mackay, 1996). Family selection is now widely practiced in the Seedling Stage by sugarcane breeding programs all over the world including Australia (Hogarth *et al*., 1990; Cox and Stringer, 1998; Kimbeng *et al*., 2000; Jackson *et al*., 1995a, b), the USA (Milligan and Legendre, 1990; Chang and Milligan, 1992a, b), India (Shanthi *et al*., 2008) and Brazil (de Resende and Barbosa, 2006). Family selection is followed by individual seedling selection which is restricted to the selected families. Some of the advantages of family selection stem from the facts that, families can be evaluated in replicated family plots across locations and the plots can be harvested mechanically and weighed. This cannot be achieved with individual seedlings because of the lack of planting material but more importantly, because of the large number of seedlings involved at this stage of the program. The ability to replicate families across time and space would account for genotype by environment interaction effects and increase gains to selection particularly for traits controlled by quantitative genes such cane yield. This aspect is important because cane yield is the primary trait that is selected at the early selection stages in most breeding programs.

Prior to family selection, sugarcane breeders relied on the proven cross status to assess the potential of a family or cross to produce elite progeny (Heinz and Tew, 1987). The proven cross system defined elite families using the proportion of seedlings advanced to later stages of the

program. It was widely used in Australia and South Africa (Heinz and Tew, 1987; Skinner *et al*., 1987) and several other breeding programs. The value of the proven cross system was questioned by Walker (1963) because larger numbers of seedlings were planted from elite families at the expense of new crosses creating a bias against the new families. Another disadvantage of the proven cross system was the lack of statistical tests to compare the families. The proven cross system also took several years to evaluate the family potential because the breeder had to wait for advancements of clones to later stages to quantify the value of the individual families.

The availability of objectively measured data (e.g., cane yield and sucrose content) from family plots has prompted sugarcane breeders to rely increasingly upon information obtained from family appraisals to make selection decisions that impact several other important aspects of the breeding program. Decisions relating to the breeding value of parents (Balzarini, 2000; Cox and Stringer, 1998; Stringer *et al*., 1996; Chang and Milligan, 1992a, b) to retain for future crossing, which cross combinations to make, and the number of crosses and seedlings per cross to plant and ultimately select from are all guided by information derived during family selection. It is therefore, vitally important for breeding programs to apply the most appropriate methods to collect, analyze and interpret data from family appraisals.

Sugarcane breeders have customarily relied on differences between the family mean values as determined by the analysis of variance (ANOVA) to select elite families during family selection (Hogarth *et al*., 1990; Cox and Stringer, 1998; Kimbeng *et al*., 2000; Jackson *et al*., 1995a, b; Milligan and Legendre, 1990; Chang and Milligan, 1992a, b). Despite its success in improving genetic gain, relative to individual seedling selection alone, these gains are not optimal because clones have been found to perform better or worse than expected on the basis of their family performance in seedling trials (Kimbeng *et al*., 2000; Hogarth *et al*., 1990). Even when within family variances were taken into consideration, Skinner *et al*. (1987) found that

families with similar means and variances produced different proportions of elite clones indicating the need to explore other statistical methods that could be used to characterize sugarcane populations in the early stages of the program. The deficiency of family means was attributed to the failure to determine and account for the distribution patterns for cane yield of seedlings within the families (Skinner *et al*., 1987) yet it is generally acknowledged that the objective of selection is to alter the distribution patterns of the cane yield of seedlings within the families. Evaluating families using clones was also found to be correlated to family evaluation using seedlings by Chang and Milligan (1992a, b) further pointing to the need to investigate other alternative approaches to family evaluation.

Although an overall increase in family mean is desirable, the ultimate goal for sugarcane breeders is to select seedlings that lead to the best-yielding clones (Kimbeng and Cox, 2001). Thus, the repeatability between seedling and clonal performance should be an important aspect of selection in sugarcane (Bressiani *et al*., 2003; Ladd *et al*., 1974; Miller and James, 1974). Despite this acknowledgement, studies that have evaluated performance in the seedling and clonal stages have relied upon statistics such as the ranks, means, BLUPS and correlations (Chang and Milligan, 1992a, b; Cox and Stringer, 1998; Kimbeng *et al*., 2000) between the two stages to draw inferences. No studies have evaluated and modeled the potential variation for repeatability that could exist among families between the seedling and clonal plots. Several studies looking at repeatability in sugarcane have been confined to between the seedling and clonal stages (Hogarth, 1971; Cesnik and Venkovsky, 1974; Ladd *et al*., 1974; Marriotti, 1974, 1977; James and Miller, 1975; Miller and James, 1975; Kang *et al*., 1983; De Sousa-Vieira *et al*., 2005; Bressiani *et al*., 2003), and no studies have investigated the variability in repeatability among families. To date, repeatability between seedlings and clones is not directly used in most

breeding programs and continues to be overlooked as a parameter for use in family evaluation. One of the reasons for not using the repeatability between seedlings and clones in family selection could be the unavailability of appropriate statistical methods adapted to use repeatability for family evaluation.

Accounting for repeatability as a parameter for family evaluation and selection would allow comparison of the family distribution patterns and trends between the seedlings and clones. In this study, we hypothesized that variation for repeatability among families exists causing some families to produce larger correlations between seedlings and clones than others. If this hypothesis is true, then families can be evaluated for repeatability between seedlings and clones for cane yield using a novel statistical tool known as random coefficient models (RCM). With RCM the clonal cane yield could be modeled as the response variable and seedling cane yield the predictor variable. The intercept would be used to measure yield potential while the slopes would measure the repeatability of the families.

The objective of this study was to demonstrate the use of RCM analysis to evaluate sugarcane families for yield potential and repeatability between the seedling and clonal stages. The ability of the RCM analysis, analysis of covariance (ANCOVA) and family means derived from ANOVA to identify elite families was compared.

2.2 Materials and Methods

2.2.1 Experimental Materials and Data Collection

2.2.1.1 Families

The 17 sugarcane families (crosses) used in this study were a random sample from the 2000 (HB00 series) and 2001 (HB01 series) crossing program at the United States Department of Agriculture, Agricultural Research Service (USDA, ARS) Sugarcane Research Station at

Houma, Louisiana, USA (Table 2.1). The seedlings from these families were germinated and grown in the greenhouse, and later transplanted to the field as individual plants in the spring of 2002. The transplanted seedlings were harvested in the fall and left to over winter.

2.2.1.2 Stage I Trial (Seedlings)

The stage I trial refers to seedling stools that were initially established from true seed. One set of the unselected individual seedlings was planted in the first replication and the other set in the second replication. Seedlings from each family were planted to 2-row plots with about 16 seedlings per row. Families but not seedlings were replicated. The seedlings were harvested and left to overwinter in 2003 (Table 2.1). From the seedlings that survived the winter (Table 2.1), eight seedlings (four seedlings from each row) were randomly chosen per family per replication (plot). At harvest, the number of stalks produced by each of the chosen seedlings was counted. The stalk height was measured as the height from the base of the tallest stalk to the top most visible dewlap. The stalk diameter of three random stalks per seedling was measured at the center of the stalk (without reference to the node) using a caliper and the mean diameter together with stalk number and stalk height were used to estimate the seedling cane yield (Equation 2.1). The seedling cane yield was calculated using the formula used by De Sousa-Vieira and Milligan (1999) (Equation 2.1). Their calculation assumed the sugarcane stalk was a perfect cylinder with specific gravity of 1.00 g cm^3 (Miller and James, 1974; Gravois *et al*., 1991; Chang and Milligan, 1992b).

$$\text{Seedling cane yield (g)} = ndπr^2L, \qquad \text{Equation 2.1}$$

where, n = number of stalks, d = density at 1.00 g cm^{-3}, r = stalk radius (cm) (radius was calculated from the diameter divided by 2), and L = stalk length (cm).

Table 2.1: The crossing series, family female and male parents, and number of seedlings that survived winter in replications 1 and 2 of the 17 sugarcane families used in family appraisal trials.

Series	Family	Female Parent	Male Parent	Rep 1	Rep2
HB00	306	Ho94-856	HoCP96-540	18	17
HB01	3055	HoCP00-945	HoCP99-866	18	17
HB01	3074	HoCP00-950	HoCP96-540	16	19
HB01	3093	HoCP00-945	HoCP96-540	17	18
HB01	3101	HoCP99-866	HoCP96-540	17	16
HB01	3107	HoCP00-950	LCP85-384	18	16
HB01	3111	HoCP99-866	LCP85-384	18	18
HB01	3174	HoCP00-945	LCP85-384	18	18
HB01	3249	N27	LCP85-384	23	11
HB01	3255	HoCP00-945	Ho94-856	18	18
HB01	3256	HoCP00-950	Ho94-856	17	18
HB01	3257	L98-207	Ho94-856	18	18
HB01	3276	TUCCP77-42	HoCP99-866	17	17
HB01	3322	TUCCP77-42	L98-207	18	18
HB01	3328	HoCP91-555	LCP85-384	18	20
HB01	3345	HoCP91-555	L98-207	17	17
HB01	3417	HoCP91-555	TUCCP77-42	18	18

2.2.1.3 Stage II Trial (Clones)

In stage II, the 17 families were planted in a trial with two replications where each family was randomized to a plot. Each family was planted to two rows per plot. Each row of a plot was planted to four clones, making up a total of eight clones per plot. Each clone was planted to a sub-plot that was one row by 1.2 meters long within the main plot. Therefore the families and not the clones were replicated. Each family was represented by sixteen individual clones derived from the 16 chosen seedlings from Stage I. The identity of the seedlings was maintained in the

clonal plots in stage II. At harvest, the number of stalks per sub-plot was counted. From each sub-plot, five random stalks were manually cut and weighed. The five-stalk sample weights were used to calculate the average stalk weight (in kilograms) for each clone. The number of stalks per sub-plot was multiplied by the average stalk weight for that sub-plot to estimate the clonal cane yield. The data were measured in the plant (2004) and second ratoon (2006) crops. The first ratoon crop (2005) was severely lodged after hurricane Katrina and no data were collected.

2.2.2 Statistical Considerations and Data Analysis Using Random Coefficient Models

The RCM or conditional hierarchical linear models were developed from ANCOVA (Bryk and Raudenbush, 1992). In ANCOVA, the families are treated as fixed populations and produce fixed intercepts and slopes. In sugarcane breeding, the RCM analysis assumes a hierarchical or multilevel arrangement between the population and the sub-populations within the population (Goldstein, 1987; Bryk and Raudenbush, 1992). The individual families (sub-populations) are nested within the population of families grown every year in the breeding program. In RCM analysis, the families are independent and random subjects derived from the fixed population (Littell *et al*., 2005). The intercepts and slopes of each family that are derived from the regression of the cane yield of clones and seedlings are therefore treated as random parameters. Within each family, the intercept and slope are correlated because they come from the same subject. The family intercepts and slopes, like any random variable are described by their variances and covariance. The covariance defines the correlation between the intercept and slope of the families. With RCM analysis, the intercept and slope of each family is tested against the population intercept and slope (Longford, 1993). The RCM analysis was previously applied in animal breeding (Longford, 1993), education (Raudenbush, 1988), finance (Fieldsend *et al*., 1987), health (Lundbye-Christensen, 1991), and real estate (Harrison and Rubinfeld, 1978)

studies. For example, in real estate, the RCM analysis was used to compare the trends for house prices over time across states, cities and suburbs.

In sugarcane family appraisal, the objective is to identify the families that comprise the highest proportion of high cane yielding seedlings from a population of crosses made each year in a breeding program. The seedlings within each family represent a sub-population derived from the several seedlings from all the crosses. The individual family seedling population therefore represents a random sample from the entire breeding population. Family appraisal aims to identify those sub-populations of seedlings with high yield potential than the population and have a distribution pattern that show the high yielding seedlings are associated with high yielding clones (high repeatability). Individual families (sub-populations) are compared to the population and those families that produce higher yield potential and higher repeatability than the population are selected as the elite families. The statistical comparison and test of the family sub-populations against the entire seedling population planted in a breeding program provides an ideal application of the RCM analysis.

In this study, the population regression model was,

$$\boldsymbol{y_j} = \boldsymbol{\alpha} + \boldsymbol{\beta x_j} + \boldsymbol{\varepsilon}, \qquad \text{Equation 2.2}$$

where $\boldsymbol{y_j}$ is the cane yield of the j^{th} clone (j = 1, 2, …, s), $\boldsymbol{\alpha}$ is the population intercept, $\boldsymbol{\beta}$ is the population slope, $\boldsymbol{x_j}$ is the cane yield of the j^{th} seedling and $\boldsymbol{\varepsilon}$ is the residual error. The individual family model was,

$$\boldsymbol{y_{ij}} = \boldsymbol{a_i} + \boldsymbol{b_i x_{j(i)}} + \boldsymbol{e_{ij}}, \qquad \text{Equation 2.3}$$

where $\boldsymbol{y_{ij}}$ is the cane yield of the j^{th} clone nested within the i^{th} family (i = 1, 2, …, f), $\boldsymbol{a_i}$ is the i^{th} family intercept, $\boldsymbol{b_i}$ is the i^{th} family slope $\boldsymbol{e_i}$ is the residual error for the i^{th} family. The intercepts and slopes are not independent and follow a normal distribution where

$$\begin{pmatrix} \boldsymbol{a_i} \\ \boldsymbol{b_i} \end{pmatrix} \sim \boldsymbol{iid}\ \boldsymbol{N}\left[\begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{pmatrix},\quad \boldsymbol{\Psi}\right];\ \boldsymbol{\Psi} = \begin{pmatrix} \boldsymbol{\sigma_a^2} & \boldsymbol{\sigma_{ab}} \\ \boldsymbol{\sigma_{ab}} & \boldsymbol{\sigma_b^2} \end{pmatrix} \text{ and } \boldsymbol{e_{j(i)}} \sim \boldsymbol{iid}\ \mathbf{N(0, \sigma^2)},$$

and $\boldsymbol{\Psi}$ is their covariance matrice, $\boldsymbol{\sigma_a^2}$ is the variance for intercepts, $\boldsymbol{\sigma_b^2}$ is the variance for slopes and $\boldsymbol{\sigma_{ab}}$ is the covariance of slopes and intercepts, ***iid*** means the families have identical and independent distributions.

Equations 2.2 (population) and 2.3 (family) are combined to produce the effects model,

$$\boldsymbol{y_{ij}} = \boldsymbol{\alpha} + \boldsymbol{a_i^*} + \boldsymbol{\beta x_{j(i)}} + \boldsymbol{b_i^* x_{j(i)}} + \boldsymbol{e_i}, \qquad \text{Equation 2.4}$$

$$\text{where } \boldsymbol{a_i^*} = \boldsymbol{a_i} - \boldsymbol{\alpha};\ \boldsymbol{b_i^*} = \boldsymbol{b_i} - \boldsymbol{\beta} \text{ and } \begin{pmatrix} \boldsymbol{a_i^*} \\ \boldsymbol{b_i^*} \end{pmatrix} \sim \boldsymbol{iid}\ \boldsymbol{N}\left[\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \boldsymbol{\Psi}\right].$$

The random effect $\boldsymbol{a_i^*}$ is the deviation of the *i*th family intercept from the fixed population intercept ($\boldsymbol{\alpha}$)**,** and the random effect $\boldsymbol{b_i^*}$ is the deviation of the *i*th family slope from the fixed population slope (***β)***. The random effects $\boldsymbol{a_i^*}$ and $\boldsymbol{b_i^*}$ have a mean of zero and covariance matrice $\boldsymbol{\Psi}$. Equation 2.4 resembles a mixed model,

$$\boldsymbol{y_{ij}} = \boldsymbol{\alpha} + \boldsymbol{\beta x_{j(i)}} + \boldsymbol{a_i^*} + \boldsymbol{b_i^* x_{j(i)}} + \boldsymbol{e_{ij}}, \qquad \text{Equation 2.5}$$

where $\boldsymbol{\alpha} + \boldsymbol{\beta x_{j(i)}}$ is the fixed effects component of the model (population model), and,

$\boldsymbol{a_i^*} + \boldsymbol{b_i^* x_{j(i)}} + \boldsymbol{e_i}$ is the random effects component of the model. Equation 2.5 can be written as,

$$\boldsymbol{y_{ij}} = \boldsymbol{\alpha} + \boldsymbol{\beta x_{j(i)}} + \boldsymbol{e_{ij}^*} \qquad \text{Equation 2.6}$$

where $\boldsymbol{E(y_{ij}) = \alpha + \beta x_{ij}}$, produces the fixed effects component of the model, and,

$\boldsymbol{Var(y_{ij}) = [1 \ x_{j(i)}]\Psi \begin{bmatrix} 1 \\ x_{j(i)} \end{bmatrix} + \sigma_e^2}$, the variance used for testing the random effects. Equation 2.5 was used in the mixed procedure of SAS (SAS Institute, 2007) to perform the RCM analysis.

2.2.3 Data Analysis Using Simple Linear Regression, ANCOVA and ANOVA

Simple linear regression was performed using SAS mixed procedures to determine if the population intercept and slope were significant. This analysis would test if the population produced significant yield potential (intercept) and if there was significant repeatability (slope) between clones and seedlings for the population. Using SAS mixed procedures removed the variation associated with the random variables (crop-years, replications and clones within families) from the experimental error. The clonal cane yield was the response variable and the seedling cane yield was the independent variable. The cane yield of seedlings was the independent variable because it was used to predict the cane yield of clones, the response variable. The linear mixed model used was,

$$\boldsymbol{Y_{ijkm} = T_i + R(T)_{j(i)} + \alpha + FR(T)_{jk(i)} + \beta x_{m(ijk)} + C(F)_{m(k)} + \varepsilon_{ijkm}}, \text{ Equation 2.7}$$

where $\boldsymbol{Y_{ijkm}}$ is the clonal yield estimated from the *i*th crop-year (*i* = 1, 2), *j*th replication (*j* = 1, 2), *k*th family (*k* = 1, 2,…, 17) and *m*th clone (*m* = 1, 2,…, 16), $\boldsymbol{T_i}$ is the random effect of the *i*th crop-year, $\boldsymbol{R(T)_{j(i)}}$ is the random effect of the *j*th replication nested within the *i*th crop-year, $\boldsymbol{\alpha}$ is the population intercept, $\boldsymbol{FR(T)_{jk(i)}}$ is the random interaction effect of the interaction of the *j*th replication by the *k*th family nested in the *i*th crop-year, $\boldsymbol{C(F)_{m(k)}}$ is the random effect of the *m*th clonal effect nested within the *k*th family, $\boldsymbol{\beta}$ is the population slope, $\boldsymbol{x_{m(ijk)}}$ is the estimated

seedling cane yield from the *m*th clone nested within the *i*th crop-year, *j*th replication and *k*th family and $\boldsymbol{\varepsilon_{ijkm}}$ is the residual error.

The ANCOVA was performed using SAS mixed procedures and generated individual family intercepts and slopes. The clonal cane yield was the response variable and the seedling cane yield was the covariate. The linear mixed model used was,

$$\boldsymbol{Y_{ijkm} = T_i + R(T)_{j(i)} + a_k + FR(T)_{jk(i)} + b_k x_{m(ijk)} + C(F)_{m(k)} + \varepsilon_{ijkm},} \quad \text{Equation 2.8}$$

where, $\boldsymbol{a_k}$ is the *k*th family intercept, and $\boldsymbol{b_k}$ is the *k*th family slope. The correlation coefficient of the cane yield of the clones and seedlings of each family was also determined.

The ANOVA was performed for the seedlings and clones using the SAS mixed procedures. The linear mixed model used for the seedlings was,

$$\boldsymbol{Y_{jkm} = R_j + F_k + FR_{jk} + C(F)_{m(k)} + \varepsilon_{jkm},} \quad \text{Equation 2.9}$$

and the linear mixed model used for the clones was,

$$\boldsymbol{Y_{ijkm} = T_i + R(T)_{j(i)} + F_k + FR(T)_{jk(i)} + C(F)_{m(k)} + \varepsilon_{ijkm},} \quad \text{Equation 2.10}$$

where $\boldsymbol{F_k}$ is the fixed effect of the kth family. Family means for the seedlings and clones derived from ANOVA was used to select the elite families as is currently done. The mean cane yield of the seedlings and clones of the elite families that were selected using ANOVA, ANCOVA and RCM analysis was compared to determine the method that would consistently identify those families producing high yield in the seedlings and clonal stages.

2.3 Results

2.3.1 Population Parameters

Preliminary analysis of the cane yield data using ANOVA, ANCOVA and RCM produced similar trends for the plant and second ratoon crops. Therefore further analyses combined the data from the plant and second ratoon crops. The clonal cane yield (*y*-axis) was plotted against seedling cane yield (*x*-axis) (Figure 2.1) and was also analyzed using simple linear regression (Equation 2.7). The best fit trend of the clonal cane yield versus the seedling cane yield was fitted using least squares. A perfect association representing the model of predicting clonal cane yield from the seedling cane yield was plotted against the population trend. This perfect linear association (PLA) assumed an intercept of zero and a slope of 3.17 [(92-0)÷(29-0)), where 92 kg was the maximum clonal cane yield, 29 kg was the maximum seedling cane yield and 0 was the minimum cane yield for seedlings and clones. The slope of 3.17 means that for every one kilogram increase in seedling cane yield, the clonal cane yield was expected to increase by 3.17 kilograms. The slope of the population trend line (Figure 2.1) was highly significant ($r = 0.45$, $P < 0.01$) and smaller than the PLA indicating a less than perfect association between the cane yield of seedlings and clones. The significant slope indicated that the seedling cane yield was predicting the clonal cane yield and also indicated significant repeatability between seedlings and clones. The slope of 1.66 meant that for every kilogram increase in seedling cane yield, the clonal cane yield was expected to increase by 1.66 kilograms. The wide scatter (Figure 2.1) represented the variability in intercepts and slopes among the family trends. This variability suggested the potential of comparing and selecting families for intercepts (yield potential) and slope (repeatability).

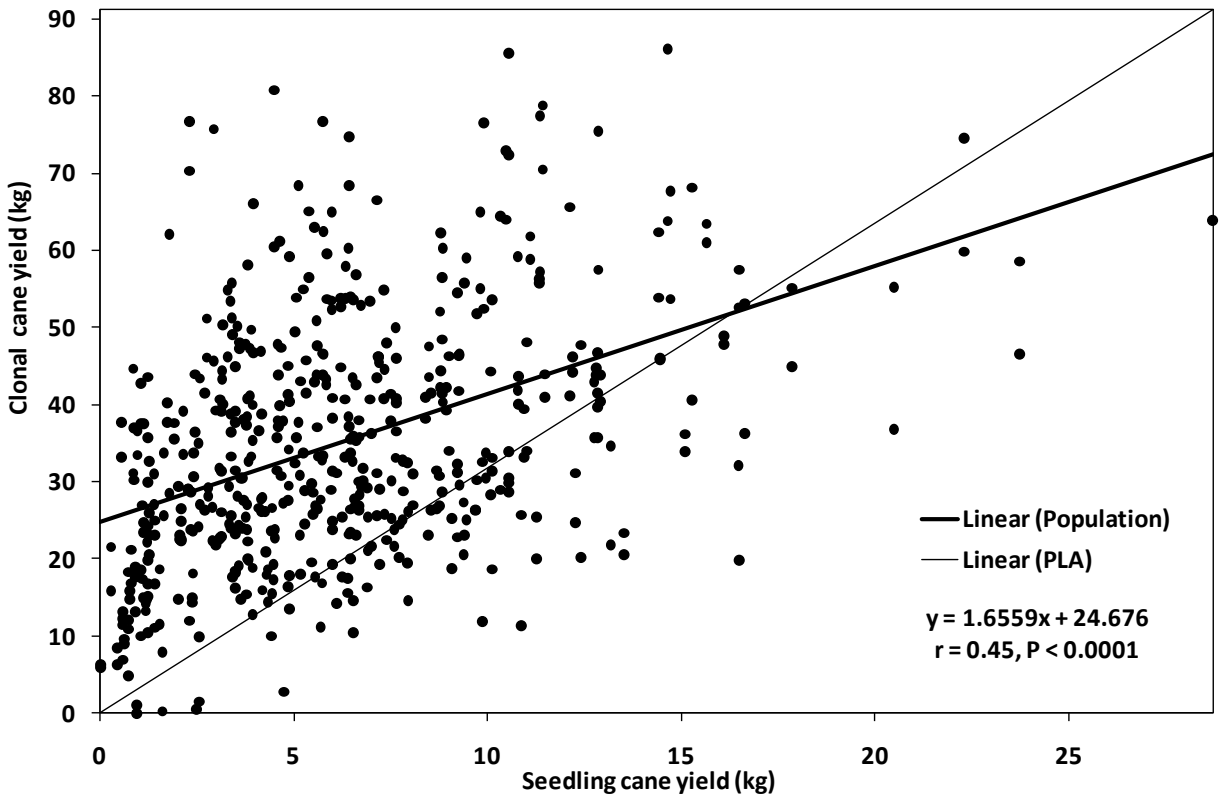

Figure 2.1: The clonal cane yield (*y*-axis) plotted against the seedling cane yield (*x*-axis) of the 17 families, their population trend and the perfect linear association (PLA). Each family comprised 16 entries.

2.3.2 Family Evaluation Using ANCOVA

The ANCOVA was performed using the clonal cane yield as the response variable and the seedling cane yield as the covariate. The analysis assumed that each family was a fixed sub-population. The overall fixed effects tests for the intercepts and slopes were highly significant ($P < 0.01$). Significant overall family intercepts indicate that at least one of the intercepts was significantly larger or smaller than zero. Similarly, significant overall family slopes indicate that at least one of the slopes was significantly larger or smaller than zero. Table 2.2 shows the estimates of the intercepts and slopes, their standard errors (S.E.), probability (P-value) of the

tests, and the correlation coefficient for each family. The estimates of the parameters (intercepts and slopes) divided by the standard errors produces a t-statistic. The P-value is the probability of obtaining a larger value of the t-statistic. All families except 3101 (P = 0.41), 3111 (P = 0.04) and 3174 (P = 0.03) produced significantly larger intercepts (P < 0.01) than zero. Families 3249 (44.08 kg) and 3257 (45.18 kg) produced the largest intercepts while 3101 (5.76 kg) and 3174 (11.74 kg) produced the smallest. Families 3055, 3093, 3101, 3107, 3111, 3174, 3255, 3276, 3322, 3345 and 3417 produced significantly larger slopes (P < 0.05) than zero indicating significant repeatability. Families 3101 (4.08) and 3174 (3.55) produced the largest slopes and 3249 (-0.42) and 3257 (-1.43) produced the smallest. Families with larger intercepts produced smaller slopes, whereas families with smaller intercepts produced larger slopes indicating that there could be negative correlations between the intercepts and slopes. Families with larger slopes (3101, 3174, 3276, and 3322) produced relatively higher correlations coefficients and families with smaller slopes (3249, 3257, and 3328) produced lower correlation coefficients between the cane yield of seedlings and clones (Table 2.2). From the ANCOVA, families 3093, 3101, 3111, 3174, 3255, 3276, 3322 and 3417 were selected as the elite families and 3249, 3257 and 3328 were rejected. Statistical comparison of families across trends is not possible with ANCOVA. With ANCOVA, when slopes of all the families are equal, larger intercepts would mean higher yield potential. Similarly, when intercepts of all the families are equal, then larger slopes mean higher repeatability. However, when the intercepts and slopes of all the families are different (Table 2.2), as is the case in this study, the comparison of the family parameters is more complex. In such a situation, the family intercepts and slopes can only be compared at a particular seedling cane yield using contrast statements. Such tests would provide limited insight into the differences in the distribution patterns for cane yield among the families.

Table 2.2: The estimates, standard errors (S.E.), and probability of a larger t-statistic (P-value) for the intercept and slope, and the correlation coefficient of the 17 families derived from the analysis of covariance with the clonal cane yield as the response variable and the seedling cane yield as the covariate.

Family	Intercept		Slope		Correlation coefficient
	Estimate ± S.E.	P-value	Estimate ± S.E.	P-value	
306	27.36 ± 5.38	0.01	1.01 ± 0.63	0.11	0.45
3055	18.73 ± 4.93	0.01	1.62 ± 0.47	0.01	0.46
3074	29.73 ± 6.00	0.01	0.97 ± 0.71	0.17	0.38
3093	18.15 ± 4.27	0.01	1.94 ± 0.34	0.01	0.71
3101	5.76 ± 6.86	0.41	4.08 ± 0.68	0.01	0.67
3107	21.24 ± 6.25	0.01	1.32 ± 0.62	0.03	0.33
3111	15.59 ± 7.48	0.04	2.03 ± 0.69	0.01	0.57
3174	11.74 ± 5.20	0.03	3.55 ± 0.56	0.01	0.62
3249	44.08 ± 5.16	0.01	-0.42 ± 0.63	0.50	0.01
3255	29.41 ± 5.97	0.01	1.84 ± 0.67	0.01	0.57
3256	33.60 ± 5.56	0.01	1.16 ± 0.76	0.13	0.38
3257	45.18 ± 5.87	0.01	-1.43 ± 0.85	0.09	-0.08
3276	25.56 ± 4.99	0.01	2.06 ± 0.44	0.01	0.61
3322	16.79 ± 4.88	0.01	2.88 ± 0.48	0.01	0.67
3328	32.15 ± 5.72	0.01	0.38 ± 0.90	0.68	0.07
3345	20.43 ± 5.59	0.01	1.22 ± 0.59	0.04	0.39
3417	23.67 ± 4.50	0.01	1.71 ± 0.42	0.01	0.56

2.3.3 Interrelationships Among the Family Parameters

The interrelationships among family intercepts, slopes, means and standard deviations were investigated graphically (Figure 2.2). A significant ($P < 0.01$) negative correlation was found between the slopes (*y*-axis) and intercepts (*x*-axis) (Figure 2.2a), a result suggested in Table 2.2. The means (*y*-axis) and intercepts (*x*-axis) showed significant ($P < 0.05$) and positive correlation (Figure 2.2b) suggesting that intercepts could indicate yield potential. The means (*y*-axis) and

slopes (*x*-axis) were not significantly (P > 0.05) correlated (Figure 2.2c) indicating that family means provided no insight into the repeatability between the cane yield of seedlings and clones. The slopes (*y*-axis) and standard deviations (*x*-axis) showed significant (P < 0.01) and positive correlation (Figure 2.2d) indicating that slopes could be used to infer within population variability in addition to measuring repeatability.

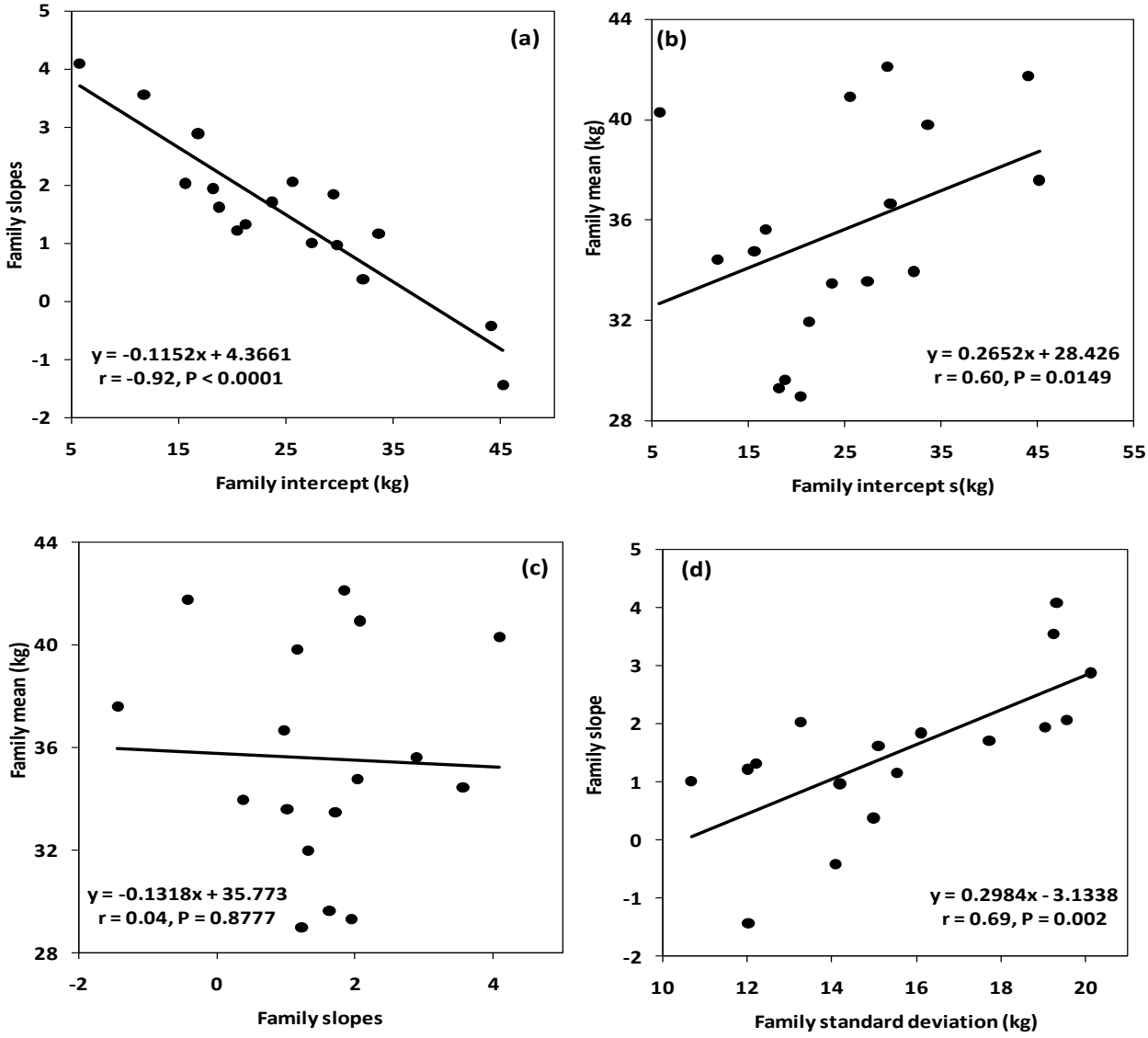

Figure 2.2: The plots of the family slopes (*y*-axis) versus the family intercepts (*x*-axis) (a), the family means (*y*-axis) versus the family intercepts (*x*-axis) (b), the family means (*y*-axis) versus the family slopes (*x*-axis) (c) and, the family slopes (*y*-axis) versus the family standard deviations (*x*-axis) (d) of the 17 families.

2.3.4 Covariance Parameter Estimates Derived From the Random Coefficient Models Analysis

The covariance parameters generated from RCM analysis describe the variation within, and association between the family intercepts and slopes. The ability to model the covariance structure creates more statistical power for the RCM analysis tests than those from ANCOVA. The covariance between the intercepts and slopes is included when computing the variances used to test the family intercepts and slopes against the population intercepts and slopes (Equation 2.6). The covariances account for the correlation between intercepts and slopes in the variances used for these tests. The covariance matrix was modelled using the unstructured structure (Appendix 1, Table 2.3). The variances for the family intercepts and slopes ($P < 0.05$) were significant indicating variability among the families. The covariance of the intercepts and slopes was negative and not significant ($P > 0.05$). The negative covariance confirmed the negative association reported in Figure 2.2a and suggested in Table 2.2.

Table 2.3: The estimates, standard errors, normal distribution statistic (Z-value) and the probability of obtaining a larger Z-value for the covariance parameters of the family intercepts and slopes

Parameters	Estimate	Standard Error	Z-value	Probability
Variance (intercepts)	39.58	21.68	1.83	0.03
Covariance (slopes, intercepts)	-3.67	2.39	-1.54	0.12
Variance (slopes)	0.53	0.31	1.69	0.05
Residual	186.70	11.76	15.87	<0.01

2.3.5 Family Evaluation Using Random Coefficient Models

The RCM analysis tested the family intercepts and family slopes against the population intercept and population slope, respectively. These tests provide a mechanism for testing the family yield

potential (intercept) and the family repeatability (slope), where families with larger values are desirable. The intercept effect was computed as the family intercept minus the population intercept. Similarly, the family slope effect was computed as the family slope minus the population slope. Positive effects indicated larger family intercepts or larger family slopes than the population intercept or population slope, respectively. Negative effects indicated smaller family intercepts or smaller family slopes than the population intercept or population slope, respectively. Equation 2.6 and the covariance parameters (Table 2.3) were used to compute the variances that were used for testing the family effects. The overall tests for the intercept and slope effects were highly significant ($P < 0.01$) (data not shown). The significant overall intercept effects indicated that the intercept effect of at least one family was significantly larger than zero. Similarly, significant overall slope effects indicated that the slope effect of at least one family was significantly larger than zero. Families 3249 and 3257 produced significant ($P < 0.05$) and positive intercept effects (Table 2.4) indicating higher yield potential than the entire population. Families 3101 and 3322 produced significant ($P < 0.10$) and positive slope effect indicating higher repeatability than the population while 3249 and 3257 produced significant ($P < 0.10$) and negative slope effects indicating lower repeatability. Using the RCM tests in Table 2.4, families 3101, 3174, 3255, 3256, 3276 and 3322 were selected as the elite families. These families produced positive slope effects, and a combination of both positive and negative intercept effects indicating higher repeatability and similar yield potential to the population. The rejected families (3249, 3257, and 3328) produced positive intercepts and negative slope effects, indicating lower repeatability than the population despite apparent greater yield potential.

Table 2.4: The effects (kg), standard errors (S.E) (kg) and the probability of obtaining a larger t-statistic (P-value) for the tests of the intercept and slope of the 17 families

Family	Test for the intercept		Test for the slope	
	Effect of intercept ± S.E.	P-value	Effect of slope ± S.E.	P-value
306	0.92 ± 3.82	0.81	-0.25 ± 0.49	0.61
3055	-3.37 ± 3.75	0.37	-0.12 ± 0.45	0.79
3074	1.53 ± 4.10	0.71	-0.10 ± 0.51	0.84
3093	-4.55 ± 3.23	0.16	0.25 ± 0.37	0.50
3101	-7.27 ± 4.60	0.11	1.07 ± 0.52	0.04
3107	-1.50 ± 4.48	0.74	-0.34 ± 0.51	0.51
3111	-3.79 ± 4.70	0.42	0.01 ± 0.49	0.98
3174	-5.14 ± 3.82	0.18	0.73 ± 0.47	0.13
3249	11.60 ± 3.78	0.01	-0.99 ± 0.50	0.05
3255	1.11 ± 4.17	0.79	0.48 ± 0.52	0.35
3256	4.93 ± 3.85	0.20	-0.05 ± 0.53	0.93
3257	8.65 ± 3.88	0.03	-0.95 ± 0.54	0.08
3276	0.68 ± 3.62	0.85	0.31 ± 0.40	0.44
3322	-4.42 ± 3.64	0.23	0.71 ± 0.44	0.10
3328	3.80 ± 3.85	0.32	-0.48 ± 0.57	0.40
3345	-2.85 ± 4.01	0.48	-0.30 ± 0.49	0.54
3417	-0.31 ± 3.38	0.93	0.01 ± 0.40	0.98

In addition to Table 2.4, the families were also evaluated graphically by plotting the seedling cane yield (*x*-axis) of each family against its clonal cane yield (*y*-axis) for all the families. The least square best fit lines were fitted for each family alongside the population best fit trend and the 1:1 line (Figure 2.3). Four clusters emerged from Figure 2.3. Families 3101, 3174, 3255, 3256, 3276, and 3322 made up cluster 1. The families in cluster 1 produced larger slopes or larger intercepts or both than the population, and were categorized as the elite families. Families 306, 3074, 3093, and 3417 (cluster 2) produced similar intercepts and similar slopes to

the population and were categorized as the average families. The families in cluster 3 were 3055, 3107, 3111, and 3345, and produced smaller intercepts, smaller slopes or both compared to the population. These families were categorized as below average. The families 3249, 3257, and 3328 (cluster 4) produced larger intercepts and smaller slopes than the population and were categorized as families to discard. The families in cluster 3 (low yield potential) and cluster 4 (lack of repeatability) could be rejected because they are likely to comprise a significantly low proportion of superior yielding genotypes. The families in clusters 1 and 2 are expected to yield more high cane yielding seedlings and clones and would be subjected to individual seedling selection.

2.3.6 Random Coefficient Models Analysis of Four Classified Family Groups

A new data set was created with the four family groups (elite, average, below average and discard) derived from the groupings defined in Figure 2.3, as the random subjects. The data set was subjected to RCM analysis to determine the RCM effects of the family groups. The elite families produced significant ($P < 0.10$) and positive slope effects while the discarded families produced significant ($P < 0.05$) and negative slope effects (Table 2.5). The discarded families produced significant ($P < 0.01$) and positive intercept effects. The elite, average and below average families produced non-significant negative intercept effects. The elite family trend was consistently larger than that of the population and the PLA, indicating greater yield potential and higher repeatability than the population. The average families were similar to the population while the below average families produced lower yield potential and marginally lower repeatability than the population. The discarded family trend showed zero repeatability and produced the largest intercept effect.

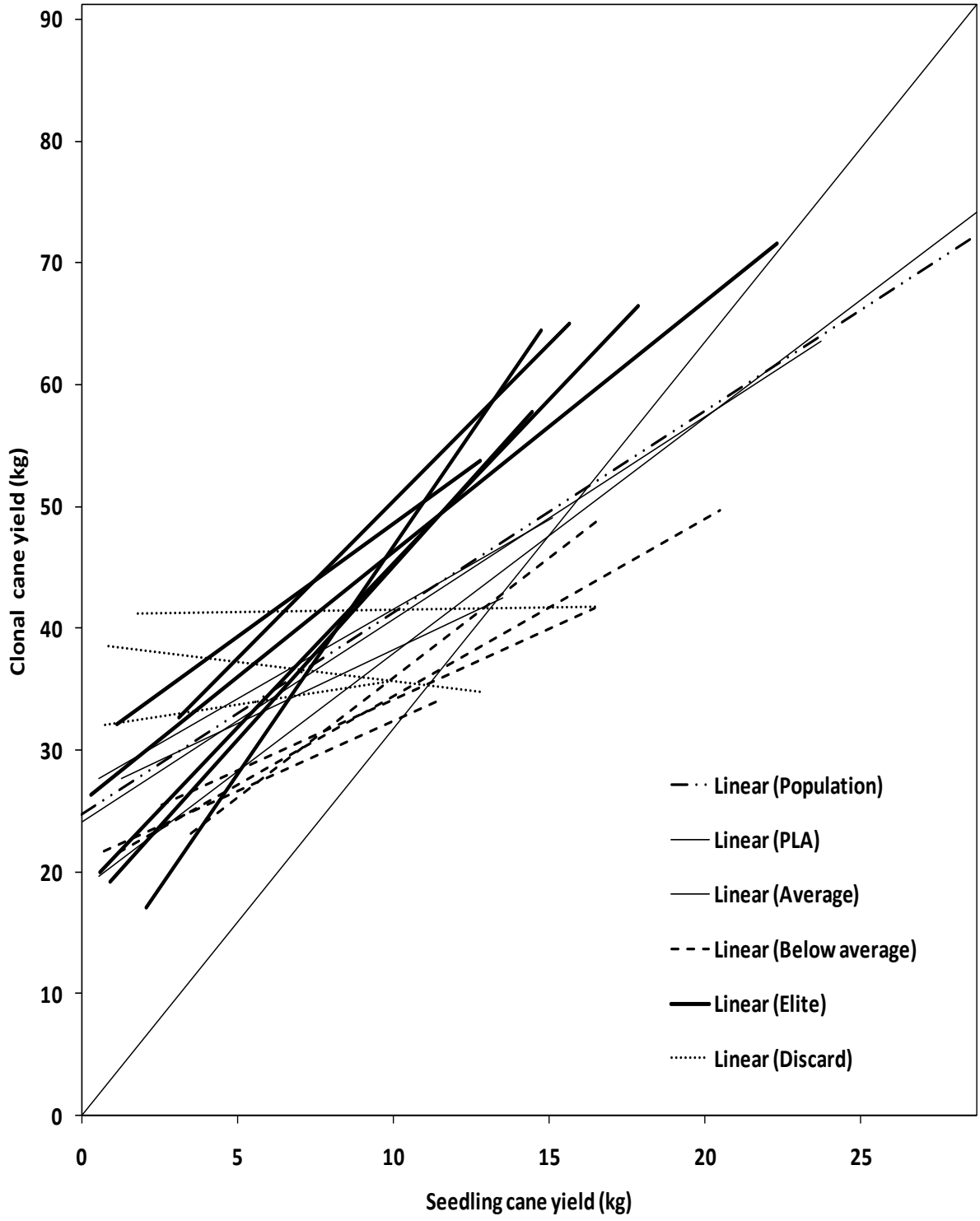

Figure 2.3: The best fit trend lines for the population, perfect linear association (PLA) and the 17 families (classified into elite, average, below average and discard) derived from the plot of clonal cane yield (*y*-axis) versus seedling cane yield (*x*-axis).

Table 2.5: The effects ± standard errors (S.E.) (kg) and the probability of obtaining a larger t-statistic (P-value) for the test of the intercepts and slopes of the elite, average, below average and discard family groups

Family Group	Test for the intercept		Test for the slope	
	Effect of intercept ± S.E.	P-value	Effect of slope ± S.E.	P-value
Elite	-3.17 ± 3.99	0.43	0.93 ± 0.50	0.06
Average	-2.23 ± 4.00	0.58	0.22 ± 0.50	0.66
Below Average	-4.87 ± 4.17	0.24	-0.03 ± 0.51	0.95
Discard	10.26 ± 4.19	0.01	-1.13 ± 0.55	0.04

2.3.7 Family Group Parameters

The group parameters for the elite, average, below average and discard families were evaluated. The group intercepts, slopes, means and standard deviations were compared to determine the most discriminating parameters among families. The elite, average and below average families produced similar intercepts while that of the discard families was double that of other groups (Table 2.6), indicating that the intercept could not discriminate between the elite, average and below average families. The group slopes decreased from the elite (highest) to the discarded family (lowest). The elite families produced a 49 % larger slope than that of the average families. The elite and discarded family groups produced similar means while the average and below average families also produced similar means. The elite families produced the largest standard deviations while the below average and discarded families produced the smallest. The slopes followed by the standard deviation were the most discriminating parameters while the intercept and family means (yield parameters) were the least discriminating. The slopes and standard deviations were significantly correlated (Figure 2.2d).

A mock seedling selection was done, targeting seedlings that produced ≥10 kg (1.5 times the population mean) cane yield. The group means of clones derived from seedlings with ≥10 kg

(1.5 times the population mean) decreased consistently from the elite (highest) to discarded (lowest) families. The seedlings selected from the elite families produced more clonal cane yield than the average (16 %), below average (47 %) and discarded families (62 %). Evaluation of the family groups further justified that the below average and discarded families could be discarded because of low within family variability and the seedlings selected from these families produced significantly lower cane yield than the elite and average families.

Table 2.6: The intercepts (kg), slopes, family mean (kg) and standard deviations (kg) (STDEV), clone mean cane yield of seedlings selected with ≥10 kg of the elite, average, below average and discard group of families

Family Group	Intercept (kg)	Slope	Family Mean (kg)	Selected Clone mean (kg)	STDEV (kg) (clones)
Elite	21.84	2.48	38.86	56.36	18.39
Average	23.17	1.67	33.22	48.62	15.69
Below average	19.49	1.51	31.23	38.45	13.25
Discard	37.79	-0.01	37.75	34.69	13.99

2.3.8 Distribution Patterns Within the Four Classified Family Groups

The seedling cane yield (*x*-axis) and clonal cane yield (*y*-axis) of the elite, average, below average and discarded families were plotted separately to evaluate their distribution patterns (Figure 2.4). Most of the scatter points of the elite families were located above the 1:1 and population lines indicating higher yield potential (Figure 2.4 (a)). These points also showed an ascending banding pattern, indicating higher repeatability. The trend of the elite families produced a larger slope than the population (Table 2.6) also indicating higher repeatability. The average family group was located around the population trend line (Figure 2.4(b)). Most of the

scatter plots of the below average group were located largely below the population trend line indicating lower yield potential than the population (Figure 2.4(c)). The discarded families produced a random distribution indicating no association between seedling and clonal cane yield (Figure 2.4(d)). From Figure 2.4, it can be deduced that if one selected seedlings that produced ≥10 kg cane yield more clones with cane yield greater than 45 kg would be recovered from the elite families and fewer clones from the discarded families in the clonal stage. Selecting seedlings from the discard families would be equivalent to a random selection for cane yield.

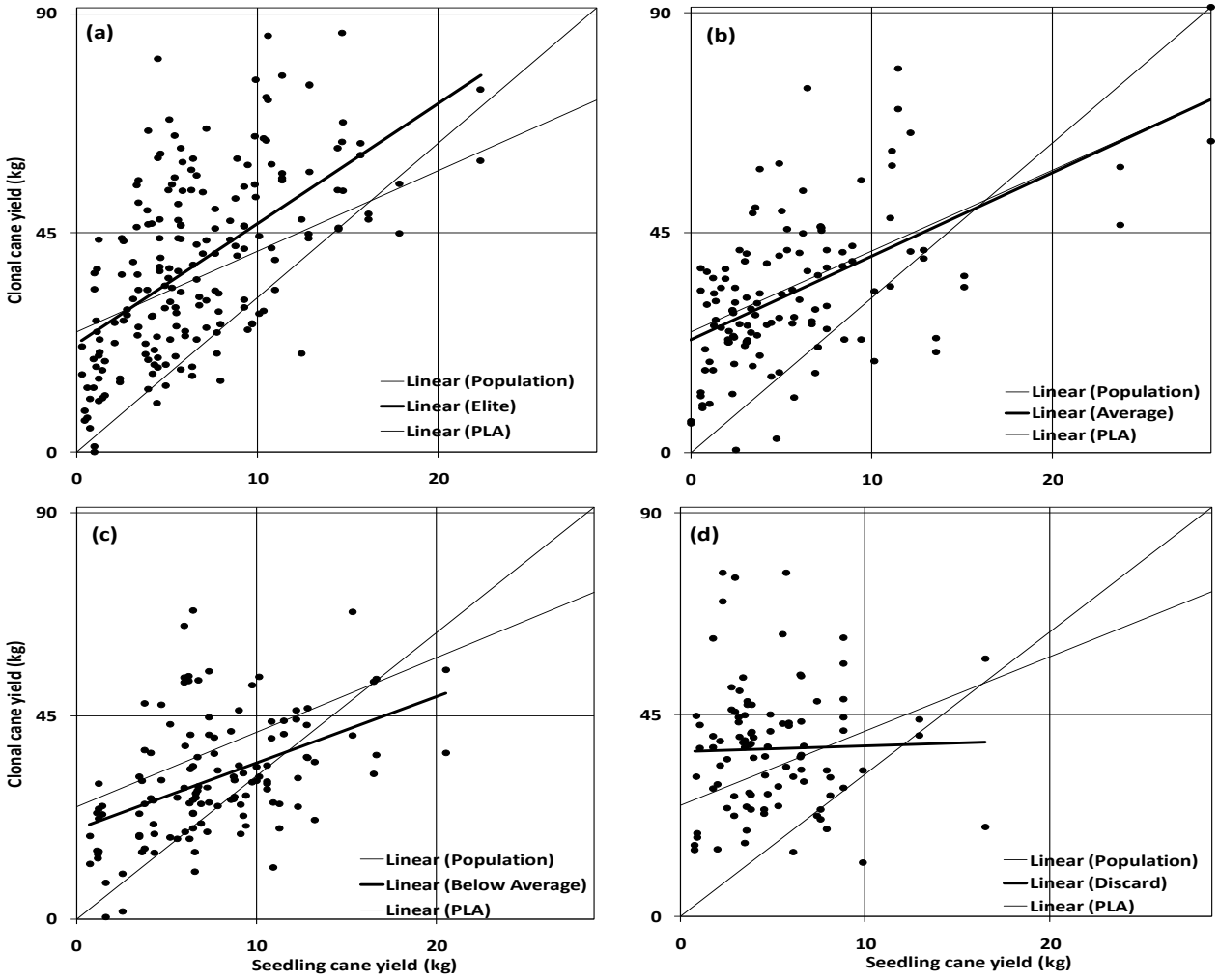

Figure 2.4: The scatter and trend lines of the clonal cane yield (*y*-axis) plotted against seedling cane yield (*x*-axis) of the elite (a), average (b), below average (c) and discard (d) families compared to the perfect linear association (PLA) and the population trends.

2.3.9 Comparison of Families Selected Using RCM Analysis, Family Means and ANCOVA

Family mean cane yield of the seedlings was ranked to identify the top 6 (high cane yield) and bottom 6 (low cane yield) families that would be selected using ANOVA, mimicking the method currently used for family evaluation. The ANOVA (Equations 2.9 and 2.10) showed no significant differences among families ($P < 0.05$) for seedling and clonal yield (data not shown). The 17 families were also ranked using their correlation coefficient (Table 2.2) to identify the top 6 and bottom 6 families that would have been selected using ANCOVA. The RCM analysis output (Table 2.4, Figure 2.3) was used to select the top 6 and bottom 6 families using intercept effects (yield potential) and slopes (repeatability). The means for the top 6 and bottom 6 families were calculated for the seedling and clonal stages as selected using ANOVA, ANCOVA and RCM analysis (Table 2.7). The means of the top 6 seedling families were 47 % (ANOVA), 16 % (ANCOVA) and 11 % (RCM) greater than that of the bottom 6 families. The ANOVA and ANCOVA clonal mean yields were similar for the top 6 and bottom 6 families indicating that these family evaluation methods failed to predict clonal cane yield of the families. The top 6 families selected using RCM analysis produced 14 % greater clonal yield than the bottom 6 families. The P-value ($P = 0.11$) of the difference between RCM top 6 and bottom 6 clonal cane yield was much smaller than that of the family means (0.86) and ANCOVA (0.92), indicating that the RCM analysis was more discriminating between high cane yield and low cane yield families than ANOVA and ANCOVA. The families identified by RCM analysis as high cane yield produced high seedling cane yield and high clonal cane yield indicating the ability of RCM analysis to identify families that produced high cane yield as seedlings and clones.

Table 2.7: The seedlings and clonal mean cane yields (kg) for the top 6 and bottom 6 families, and probability of a larger difference between top 6 and bottom 6 derived from family means, R^2 from ANCOVA and RCM family analysis methods.

Family Category	Family means		R^2 from ANCOVA		RCM analysis	
	Seedling	Clones	Seedlings	Clones	Seedlings	Clones
Top 6	7.89	36.53	6.90	37.19	6.85	38.86
Bottom 6	5.37	36.05	5.96	36.93	6.17	33.97
P-value	0.06	0.86	0.46	0.92	0.45	0.11

2.4 Discussions

The confounding of seed type between seedlings and clones that is ignored during family evaluation while using the ANOVA can be resolved by adopting RCM analysis. Family evaluation at the seedling stage as is the current practice ignores the existence of this confounding. The RCM analysis solved the influence of confounding by evaluating families using trends between cane yield of seedlings and clones. These trends provided for the evaluation for both the yield potential and repeatability between the seedlings and clones. The fact that the family means were not associated with slope, a measure of repeatability, indicated that family evaluation using means was not addressing the confounding between the seedling and clonal stages. The confounding effect of seed type results in smaller plots planted for seedlings and larger plots for clones.

Whereas ANOVA and ANCOVA produced large differences between the top 6 and bottom families for seedling cane yield these differences were not reflected among the clones, indicating the influence of confounding. The mean cane yields of the top 6 and bottom 6 families in the seedling and clonal stages also confirmed the superiority of RCM analysis and the deficiency of using means for family evaluation. Using RCM analysis, the seedling cane yield

differences between the top 6 and bottom 6 were reflected in the clones, indicating that RCM analysis identified families with higher repeatability.

The seedlings selected with high cane yield from the elite families that were identified by RCM analysis produced high cane yield clones. The seedlings selected from the elite families with ≥10 kg cane yield produced at least 47 % more clonal cane yield than seedlings selected with ≥10 kg from the rejected families (below average and discard). The elite families selected by RCM analysis also produced the largest proportion of high cane yielding clones. Therefore, seedlings selected as high in cane yield from the elite families that were identified by RCM analysis provides a greater chance of producing clones with high cane yield, for example, greater than 45 kg. The elite families identified by family means indicated that seedling selection from these families would be equivalent to random seedling selection compared to the elite families identified using RCM analysis.

The families identified as elite and average by RCM analysis produced larger standard deviations, indicating greater within family variability. Selection generally takes advantage of the within family variability. It is easier to select from a population where there is large variability among seedlings (Allard, 1960). Fewer seedlings with high cane yield from the below average and the discarded families produced higher yielding clones because of the low within family variability in these groups compared to the elite and average families. Selection from families with low within family variability appeared similar to random selection, and would result in limited or no gains in cane yield. This study showed that these families can be discarded because the effort expended in selecting from these families would not match the expected gains.

Family evaluation using RCM analysis produced more discriminating parameters than the family means. Sugarcane breeders are generally interested in discarding families that have low

number of seedling producing high cane yield. The slope, representing repeatability, was the most discriminating parameter. In family selection, the goal is to differentiate between families with a high proportion of high cane yielding seedlings from those with low population of high cane yielding seedlings. Larger slopes were associated with families whose high cane yield seedlings produced high cane yield clones. Larger slopes were also significantly correlated with larger standard deviations, indicating greater within family variability. Therefore, using RCM analysis, families that have a lower chance of producing high yielding clones can be discarded with greater precision than with family means, as is currently practiced.

Family mean cane yields produced from ANOVA failed to separate the elite and discarded families. Previous studies by Hogarth *et al*. (1990) and Kimbeng *et al*. (2000) reported significant deviations in the number of expected high cane yielding clones that were selected from the elite families evaluated using family means. In this study, family means derived from ANOVA produced no significant ($P < 0.05$) differences in cane yield for both the seedlings and clones, indicating that there was no statistical justification for family selection using family means. These families were statistically similar for cane yields, according to ANOVA. Statistically, family evaluation and selection would be valid if there were significant differences for cane yield. When there are significant differences, means separation using, for example, the least significant difference, could be used to identify families that are significantly higher yielding as elite families and those significantly lower yielding can be discarded.

The graphical presentation of the output data from the RCM analysis provided for easy interpretation of the results. The advantage of graphical presentation is in their ability to provide for the visualization of the trends in the data (Yan and Kang, 2002). The population trend in Figure 2.1 clearly showed the variability among the families. Figures 2.3 provided easy

visualization of the family groups and together with the output in Table 2.4, helped classify the families as well as identify the elite families. The graphs in Figure 2.4 provided easy visualization of the attributes of each of the family group and displayed the distribution patterns of the genotypes within the family groups. In Figure 2.4, the number of seedlings identified by a mock selection from each group of families can be evaluated visually for their potential clonal cane yield.

The data requirements will remain a challenge for the adoption of RCM analysis for family evaluation. Most breeding programs do not measure yield of seedlings and the first clonal plots. For those breeding programs that collect this data, retrospective and parallel evaluations are suggested. Retrospective evaluation will use yield data from seedlings and clones after stage II harvest. Parallel evaluation requires a breeding program to establish a family selection stage before the seedling selection stage. Fewer seedlings per family, say 10 to 20, can be planted from each of the several families and the data generated would be used to select the elite families. Only seedlings from the elite families will be grown for individual seedling selection. The fewer seedlings planted for individual seedling selection and the expected higher yield gains could more than compensate for the extra cost. In programs with active family evaluation, this approach entails an extra year to grow and select the best families using RCM analysis.

2.5 Conclusions

Our study showed that the ability to account for the influence of confounding for seed type between seedlings and clones was important for family evaluation for cane yield. The elite families selected by the RCM analysis produced high clonal cane yield from seedlings selected with high cane yield. The elite families identified using RCM analysis produced the highest proportion of high cane yield clones selected from seedlings that were identified to have

produced high cane yield. The slope (repeatability) was the most discriminating parameter among families indicating the importance of evaluating family distribution trends for cane yield. Families 3101, 3174, 3255, 3256, 3276, and 3322 were selected as the elite families using RCM analysis. These families produced high family for cane yield in the seedling and clonal stages and high within family variability (larger standard deviations). From our study, family means were inadequate for family evaluation because they failed to account for repeatability and therefore confounding effect on cane yield between seedlings and clones. In our study, the means were statistically similar among the families, a situation that could also weaken family selection based on family means. We suggest that the RCM analysis can be implemented using the retrospective or the parallel approach. Retrospective evaluation can use yield data available from Stages I and II of sugarcane breeding programs. Parallel evaluation would involve establishing a parallel family evaluation stage before the individual seedling selection stage.

2.6 References

Allard, R.W. (1960). Principles of Plant Breeding. John Wiley and Sons. New York. U.S.A.

Balzarini, M.G. (2000). Biometrical models for predicting future performance in plant breeding. PhD Dissertation, Louisiana State University.

Bressiani, J.A., Vencovsky, R. and da Silva, J.A.G. (2003). Repeatability within and between selection stages in a sugarcane breeding program. *Journal of the American Society of Sugarcane Technologists* 23: 40 – 47.

Bryk, A.S. and Raudenbush, S.W. (1992). Hierarchical Linear Models. Newbury Park, CA, Sage Publications Inc.

Cesnik, R. and Venkovsky, R. (1974). Expected response to selection, heritability, genetic correlations and response to selection of some characters in sugarcane. *Proceedings of the International Society of Sugar Cane Technologists 15*: 96 – 101.

Chang, Y.S. and Milligan, S.B. (1992a). Estimating the potential of sugarcane families to produce elite genotypes using bivariate methods. *Theoretical and Applied Genetics* 84: 633 – 639.

Chang, Y.S. and Milligan, S.B. (1992b). Estimating the potential of sugarcane families to produce elite genotypes using univariate cross prediction methods. *Theoretical and Applied Genetics* 84: 662 – 671.

Cox, M.C. and Stringer, J.K. (1998). Efficacy of early generation selection in a sugarcane improvement program. *Proceedings of ASSCT* 20: 148 – 153.

De Resende, M.D.V. and Barbosa, M.H.P. (2006). Selection via simulated individual BLUP based on family genotypic effects in sugarcane. Pesq. Agropec. Bras., Brasilia, Volume 41 Number 3: 421 – 429.

De Sousa-Vieira, O. and Milligan, S.B. (1999). Intra-row spacing and family x environment effects on sugarcane family evaluation. *Crop Science* 39: 358 – 364.

De Sousa-Vieira, O. and Milligan, S. B. (2005). Interrelationships of cane yield components and their utility in sugarcane family selection: Path coefficient analysis. *Interciencia* Volume 30 number 2: 93 – 96.

Falconer, D.S. and Mackay, T.F.C. (1996). Introduction to quantitative genetics. Fourth Edition. Longman Group Ltd, UK.

Fieldsend, S., Longford, N.T. and McLeary, S. (1987). Industry effects and the proportionality assumption in ratio analysis: A variance component analysis. *Journal of Business Finance and Accounting*, 14: 557 – 572.

Goldstein, H. (1987). Multilevel Models in Educational and Social Research. New York: Oxford University Press.

Gravois, K.A., Milligan, S.B. and Martin, F.A. (1991). Indirect selection for increased sucrose yield in early sugarcane testing stages. *Field Crops Research* 26: 67 – 73.

Harrison, D. and Rubinfeld, D.A. (1978). Hedonic prices and the demand for clean air. *Journal of Environmental Economics and Management* 5: 81 – 102.

Heinz, D.J. and Tew, T.L. (1987). Hybridization procedures. In: Heinz, D.J. (editor). Sugarcane Improvement through Breeding. Elsevier. New York.

Hogarth, D.M. (1971). Quantitative inheritance studies in sugarcane. II. Correlations and predicted responses to selection. *Australian Journal of Agricultural Research* 22: 103 – 109.

Hogarth, D.M., Braithwaite, M.J. and Skinner, T.C. (1990). Selection of sugarcane families in the Burdekin district. *Proceedings of the Australian Society of Sugarcane Technologists* 12: 99 – 104.

Jackson, P.A., McRae, T.A. and Hogarth, D.M. (1995a). Selection of sugarcane families across variable environments. II. Patterns of response and association with environmental factors. *Field Crops Research* 42: 109 – 118.

Jackson, P.A., Bull, J.K. and McRae, T.A. (1995b). The role of family selection in sugarcane breeding programs and the effect of genotype x environment interactions. *Proceedings of the International Society of Sugar Cane Technologists* 22: 261 – 269.

James, N.I. and Miller, J.D. (1975). Selection in six crops of sugarcane. II. Efficiency and optimum selection intensities. *Crop Science* 15: 37 – 40.

Kang, M.S., Miller, J.D. and Tail, P.Y.P. (1983). Genetic and phenotypic path analysis and heritability in sugarcane. *Crop Science* 23: 643 – 647.

Kimbeng, C.A., McRae, T.A. and Stringer, J.K. (2000). Grains from family and visual selection in sugarcane, particularly for heavily lodged crops in the Burdekin region. *Proceedings of the Australian Society of Sugarcane Technologists* 22: 163 – 169.

Kimbeng, C.A., McRae, T.A. and Cox, M.C. (2001). Optimizing early generation selection in sugarcane breeding. *Proceedings of the International Society of Sugarcane Technologists* 24: 488 – 493.

Ladd, S.L., Heinz, D.J., Meyer, H.K. and Nishimoto, B.K. (1974). Selection studies in sugarcane (Saccharum spp. Hybrids). I. Repeatability between selection stages. *Proceedings of the International Society of Sugar Cane Technologists* 14: 102 – 105.

Littell, R.C., Milliken, G.A., Stroup, W.W. and Wolfinger, R.D. (2005). SAS System for Mixed Models. 7th edition. SAS Institute Inc., Cary, NC, USA.

Longford, N.T. (1993). Random Coefficient Models. Oxford Statistical Science Series. Oxford University Press.

Lundbye-Christensen, S. (1991). A multivariate growth curve model for pregnancy. *Biometrics* 47: 637 – 657.

Mariotti, J.A. (1974). The effect of environments on the effectiveness of clonal selection in sugarcane. *Proceedings of the International Society of Sugar Cane Technologists* 14: 89 – 95.

Mariotti, J.A. (1977). Sugarcane clonal selection research in Argentina: A review of experimental results. *Proceedings of the International Society of Sugar Cane Technologists* 14: 121 – 136.

Miller, J.D. and James, N.I. (1974). The influence of stalk density on cane yield. *Proceedings of the International Society of Sugarcane Technologists* 15: 177 – 184.

Miller, J.D. and James N.I. (1975). Selection in six sugarcane crops. I. Repeatability of three characters. *Crop Science* 15: 23 – 25.

Milligan, S.B. and Legendre, B.L. (1990). Development of a practical method for sugarcane cross appraisal. *Journal of the American Society of Sugar Cane Technologists* 11: 59 – 68.

Raudenbush, S.W. (1988). Educational applications of hierarchical linear models: A review. *Journal of Educational Statistics* 13: 85 – 116.

SAS Institute (2007). SAS/STAT user's guide, version 9.1.3. SAS Institute, Cary, North Carolina, USA.

Shanthi, R.M., Bhagyalakshmi, K.V., Hemaprabha, G., Alarmelu, S. and Nagarajan, R. (2008). Relative performance of the sugarcane families in early selection stages. Sugar Tech 100(2): 114 – 118.

Skinner, JC, Hogarth, DM and Wu, KK (1987). Selection methods, criteria and indices. 409 – 453. In DJ Heinz (editor). Sugarcane Improvement Through Breeding. Elsevier.

Stringer, J.K., McRae, T.A. and Cox, M.C. (1996). Best linear unbiased prediction as a method of estimating breeding value in sugarcane. In: Wilson, J.R., Hogarth, D.M., Campbell, J.A. and Garside, A.L. (editors). Sugarcane: Research Towards Efficient and Sustainable Production: 39 – 41. CSIRO Division of Tropical Crops and Pastures, Brisbane.

Walker, D.I.T. (1963). Family performance at early selection stages as a guide to the breeding program. *Proceedings of the International Society of Sugarcane Technologists* 11: 469 – 483.

Yan, W. and Kang, M.S. (2002). GGE Biplot Analysis: A Graphical Tool for Breeders, Geneticists, and Agronomists. CRC Press, Florida, USA.

Zhou, M.M., Singels, A. and Savage, M.J. (2003). Physiological parameters for modelling differences in canopy development between sugarcane cultivars. *Proceedings of the South African Sugarcane Technologists Association* 77: 610 – 621.

CHAPTER 3: ARTIFICIAL NEURAL NETWORK MODELS: A DECISION SUPPORT TOOL FOR ENHANCING SEEDLING SELECTION IN SUGARCANE BREEDING

3.1 Introduction

Sugarcane is grown commercially as a clone, yet the first episode of selection must occur among individual seedlings raised from true seed. The identification of seedlings with high cane yield potential remains a challenge facing sugarcane breeders. Visual appraisal for cane yield is used during individual seedling selection. Visual selection is subjective (Cox and Stringer, 1998) and likely to be inefficient (Hogarth and Berding, 2006). Visual selection is confounded by the effects of genotype by environment interaction and competition among seedlings. The effects of genotype by environment interaction are exacerbated because seedlings cannot be replicated at this stage of the program due to space limitation owing to the large numbers involved.

Seedlings are often closely spaced because of the need to plant large numbers on limited land. Closely spaced seedlings result in altered phenotypic expression for cane yield components (stalk number, stalk height, and stalk diameter) (Breaux and Miller, 1987; de Sousa-Vieira and Milligan, 1999). De Sousa-Vieira and Milligan (1999) reported reduced genetic expression for yield component traits in closely spaced seedlings. However, most breeding programs continue to use narrow spacing because of the limited land resources (Breaux and Miller, 1987). Planting smaller seedling populations which are better managed has been suggested as one strategy to increase selection efficiency (Hogarth and Berding, 2006; Kimbeng and Cox, 2003). Hogarth and Berding (2006) also suggested the exploration of more innovative statistical techniques to improve selection efficiency. In this study, we explore a novel statistical technique known as artificial neural network (ANN) for use in sugarcane seedling selection.

An ANN model, often called neural network model, is a mathematical or computational model based on biological neural networks (Nelson and Illingworth, 1991). The ANN is a supervised learning method and uses pattern learning from training data to produce models that generate predictions of response variables (Masters, 1993; Nelson and Illingworth, 1991). The ANN consists of a layered, free forward and completely connected network restricted to a single direction of flow (Nelson and Illingworth, 1991). The ANN has an input layer, a hidden layer, and an output layer (Figure 3.1). The ANN models complex relationships between input variables and outputs (Gurney, 1997; Fausett, 1994). The model must be 'trained' by processing data with input and output patterns similar to the data to be predicted. The model detects similarities in the new input data, and uses these similarities to generate output predictions (Smith, 1993). The logistic function calculates probabilities used to make predictions (Allison, 2003; Agresti, 2007). Multiple linear regression equations form the linear predictors (Hertz *et al*., 1990; Agresti, 2007).

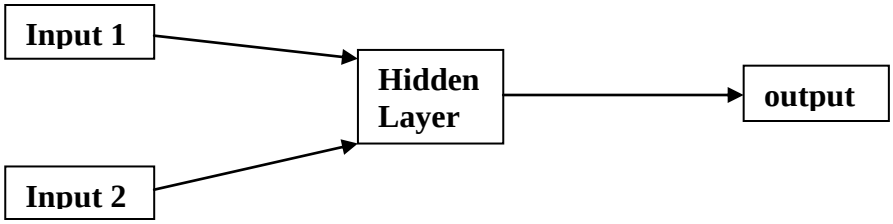

Figure 3.1: The input layers, hidden layer and output layer of the artificial neural network model.

The ANN models have been used in financial risk management (Huang *et al*., 2004; Sethuraman, 2006), process control in manufacturing (Lee and Paik, 2006), predicting credit scores and interest rates (Perkins and Brabazon, 2006), and predicting fish abundance (Iglesias *et al*., 2006). In predicting credit card scores and interest rates using ANN models, for example, the

payment history and other variables from other individuals are used as the training data set to calculate probabilities. The probabilities calculated for all individuals including the applicant using the full data set determines the relative risk of the applicant and are used to determine the interest rates on credit cards or loans. Few uses of the ANN techniques have been reported in plants. Recently the ANN model was used to assign tea accessions into taxonomic groups using leaf morphological measurements as input variables (Pandolfi *et al*., 2009). The same group of researchers used the ANN model to classify *Camellia japonica* using phyllometric and fractal parameters (Mugnai *et al*., 2008).

The same concept used above can be applied to sugarcane seedling selection. At early selection stages, Brix (% soluble solids in juice) measured by a hand refractometer is used to screen for sucrose content. Stalk diameter, stalk length and stalk number evaluate cane yield (Chang and Milligan, 1992). These yield components (Milligan *et al*., 1990) can be used as input variables in the ANN models to predict the probability of either selecting or rejecting a seedling. During selection, the decision to select or reject a seedling depends on the combination and magnitude of the cane yield components as assessed visually. The outcome or response variable would be to either select (1) or reject (0) a seedling which is binary in nature. In this case a training data set consisting of previously defined response variables (select or reject) and the input or independent variables (cane yield components) are used by the ANN model to determine the logistic regression function. Then a new data set consisting of input variables is fed into the logistic regression function which produces probabilities of either selecting or rejecting a seedling as the output (Figure 3.1).

The objective of this study was to evaluate the potential of using the SAS enterprise miner ANN models for identifying seedlings with high cane yield potential at the seedling stage

of a sugarcane breeding program. The yield of seedlings selected using the ANN models were compared to those selected using the visual method.

3.2 Materials and Methods

3.2.1 Experimental Materials and Data Collection

Data were collected from seedlings raised from true seed at the United States Department of Agriculture, Agricultural Research Service (USDA), Ardoyne Research Farm, at Schreiver, LA., and Louisiana State University Agricultural Center (LSU AgCenter) Sugar Research Station at St. Gabriel, LA. The seedlings from 17 crosses (USDA) and 5 crosses (LSU AgCenter) (Table 3.1) were first germinated and established in the greenhouse and then transplanted into the field as single stools in the summer of 2002. At the USDA, the seedlings of each cross were divided and transplanted into two replications. The crosses were replicated but not the seedlings. In each plot, two rows were planted, each to 16 seedlings. In 2003, eight seedlings (four from each row per plot) were randomly chosen from each plot and used for data collection.

At the LSU AgCenter, five crosses (Table 3.1), each with more than 500 seedlings, were selected from the seedling program. Thirty seedlings were randomly chosen from each cross in 2003. The chosen seedlings from the two populations were evaluated subjectively to determine if they would have been selected (1) or rejected (0). The decision to select (1) or reject (0) a seedling was based on a consensus by two experienced sugarcane breeders. From the chosen seedlings, stalk number was counted, stalk height was measured from the base of the stool to the top most visible dewlap, and stalk diameter was measured at the center of the stalk on three random stalks using a caliper and without reference to the bud.

Table 3.1: Cross showing female and male parents of sugarcane seedlings planted at the United States Department of Agriculture (USDA) and Louisiana State University Agricultural Center (LSU AgCenter) sugarcane research farms.

Cross ID	Female Parent	Male Parent	Cross	Female Parent	Male Parent
Crosses evaluated at the USDA research farm					
306	Ho94-856	HoCP96-540	3255	HoCP00-945	Ho94-856
3055	HoCP00-945	HoCP99-866	3256	HoCP00-950	Ho94-856
3074	HoCP00-950	HoCP96-540	3257	L98-207	Ho94-856
3093	HoCP00-945	HoCP96-540	3276	TUCCP77-42	HoCP99-866
3101	HoCP99-866	HoCP96-540	3322	TUCCP77-42	L98-207
3107	HoCP00-950	LCP85-384	3328	HoCP91-555	LCP85-384
3111	HoCP99-866	LCP85-384	3345	HoCP91-555	L98-207
3174	HoCP00-945	LCP85-384	3417	HoCP91-555	TUCCP77-42
3249	N27	LCP85-384			
Crosses evaluated at the LSU AgCenter research farm					
XL01-001	HoCP92-624	HoCP91-552			
XL01-050	LCP86-454	LC85-384			
XL01-059	HoCP95-951	HoCP96-540			
XL01-215	TucCP77-42	LCP85-384			
XL01-460	Ho95-988	L99-238			

3.2.2 Estimation of Seedling Cane Yield From Yield Components

The seedling cane yield was calculated based on the formula used by De Sousa-Vieira and Milligan (1999) (Equation 3.1). Their calculation assumed the sugarcane stalk was a perfect cylinder with specific gravity of one (1.0) as determined from previous studies (Miller and James, 1974; Gravois *et al*., 1991; Chang and Milligan, 1992).

$$\text{Seedling cane yield (g)} = ndπr^2L \qquad \text{Equation 3.1}$$

where n = seedling stalk number, d = density at 1.0 gcm^{-3}, r = stalk radius (cm), and L = stalk length (cm).

3.2.3 Data Analysis Using Artificial Neural Network Models

The training data consisted of 20 % (30 seedlings, LSU AgCenter) and 10 % (28 seedlings, USDA) of the original data. The input variables were stalk number, stalk height, and stalk diameter and the response was either to select (1) or reject (0) a seedling as determined by the two experienced sugarcane breeders. The training data was run in SAS enterprise miner (SAS Institute, 2007) to produce the coefficients of the multiple linear regressions. The data collected from 150 (LSU AgCenter) and 272 (USDA) seedlings constituted the prediction data. In the prediction data, the response values, select (1) or reject (0) a seedling were coded as missing values and were estimated by the model. The model selection criteria used was the 'average error' and the network architecture was the 'generalized linear model'. The training technique used was the 'Levenberg-Marquadt' set at 50 preliminary runs. The ANN flow chart for the analysis is shown in Figure 3.2.

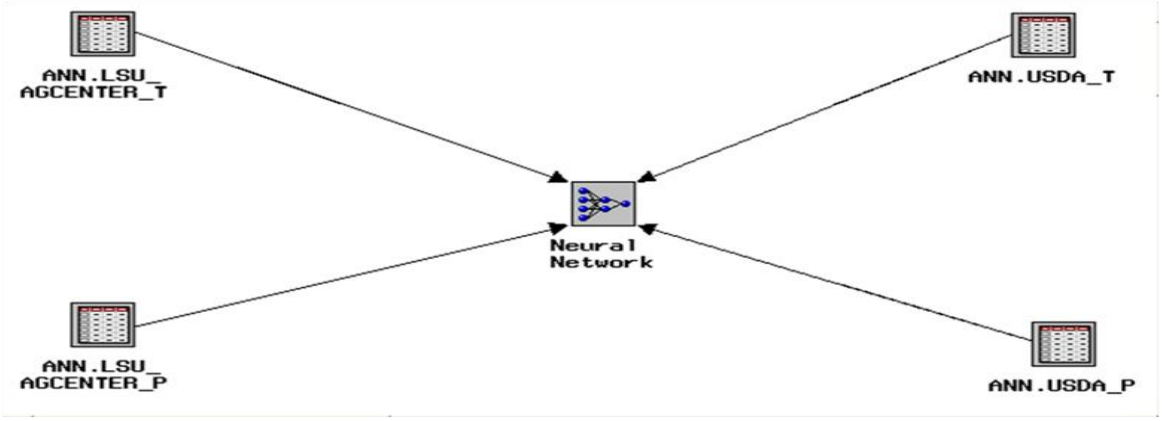

Figure 3.2: The artificial neural network flow chart used in analyzing the LSU AgCenter training (ANN.LSU_AGCENTER_T), LSU AgCenter prediction (ANN.LSU_AGCENTER_P), USDA training (ANN.USDA_T), and USDA prediction (ANN.USDA_P) data sets. ANN references the name of the SAS data library from where the data files were stored.

3.3 Results

3.3.1 Coefficients of the Prediction Models

The ANN models use the training data to produce coefficients that define the logistic regression functions. The coefficients represent the relative weighting of each input variable, similar to coefficients in multiple linear regressions (Table 3.2). The coefficients (Table 3.2) were used to build the prediction functions (Equations 3.2 and 3.3). Equations 3.2 and 3.3 were used to calculate the probability of either selecting or rejecting a seedling by plugging in the values of the stalk number, stalk height and stalk diameter of that seedling.

Table 3.2: Model coefficients for stalk diameter, stalk height, stalk number, and the intercept from artificial neural network analyses of data from the LSU AgCenter and USDA populations

Variable	LSU AgCenter	USDA
Diameter	11.20	5.71
Height	6.16	2.73
Stalk number	1.38	0.04
Intercept	-50.20	-18.11

$$P\,(Y = 1) = \frac{\exp\,(-50.2+1.38Stalks+6.16Height+11.2\ Diameter)}{1+\exp\,(-50.2+1.38Stalks+6.16Height+11.2\ Diameter)} \qquad \text{Equation 3.2}$$

$$P\,(Y = 1) = \frac{\exp\,(-18.11+0.04Stalks+2.73Height+5.71\ Diameter)}{1+\exp\,(-18.11+0.04Stalks+2.73Height+5.71\ Diameter)} \qquad \text{Equation 3.3}$$

3.3.2 Model Fit Statistics

The ANN analysis produces six fit statistics for evaluating the robustness of the model (Table 3.3). The average profit (prediction power) is estimated by the correlation between the response

variable (1 or 0) and probability (Agresti, 2007). A higher profit means the response variable was closely associated with the probability of selection. The misclassification rate is estimated as the proportion of the total observations that are classified by the model into different response categories from what was observed. Lower values indicate correct model classification and accurate training data. The average squared error (ASE) is calculated as,

$$ASE = \frac{SSE}{N} = \frac{(Observed\ Response - Prediction\ Probability)^2}{Number\ of\ observation\ in\ the\ Training\ data}. \qquad \text{Equation 3.4}$$

Smaller values indicate better model fit. The final prediction error (FPE) is estimated as,

$$FPE = \frac{SSE(N+P)}{N(N-P)}, \qquad \text{Equation 3.5}$$

where, *P* is the number of parameters including the intercept. FPE is an adjustment to ASE using (N+P)/(N–P). The adjustment penalizes for over-parameterization (model complexity) or the inclusion of too many input variables. Over-parameterization inflates FPE and increases prediction errors. It is generally desirable to achieve the best model fit by specifying the simplest or most parsimonious model. Just like with ASE, lower values indicate better model fit. The Akaike Information Criterion (AIC) (Akaike, 1974) and Schwarz Bayesian Criterion (SBC) (Schwarz, 1978) are used to compare the relative model fit for two or more models. Lower values indicate better model fit.

The fit statistics produced higher prediction power for the LSU AgCenter than the USDA data set (Table 3.3). Misclassification, ASE, FPE, AIC and SBC values were greater for the USDA population indicating poorer model fit of the data compared to that from the LSU AgCenter population.

Table 3.3: Model Fit Statistics from artificial neural network analysis of sugarcane seedling data from the USDA and LSU AgCenter populations

Fit Statistic	LSU AgCenter	USDA
Average profit	0.61	0.36
Misclassification rate	0.07	0.11
Average squared error	0.06	0.13
Final Prediction error	0.08	0.17
Akaike's Information Criterion (AIC)	19.79	31.72
Schwarz's Bayesian Criterion (SBC)	25.52	37.05

The distribution patterns of the two populations were evaluated graphically by plotting the estimated seedling cane yield (*x*-axis) against their corresponding probabilities of selection (*y*-axis). The LSU AgCenter data followed closely the theoretical logistic cumulative distribution function (Casella and Berger, 2003) compared to the USDA data (Figure 3.3). The distribution patterns depicted trends that were similar to the fit statistics (Table 3.3), confirming the larger variability found within the USDA than the LSU AgCenter data.

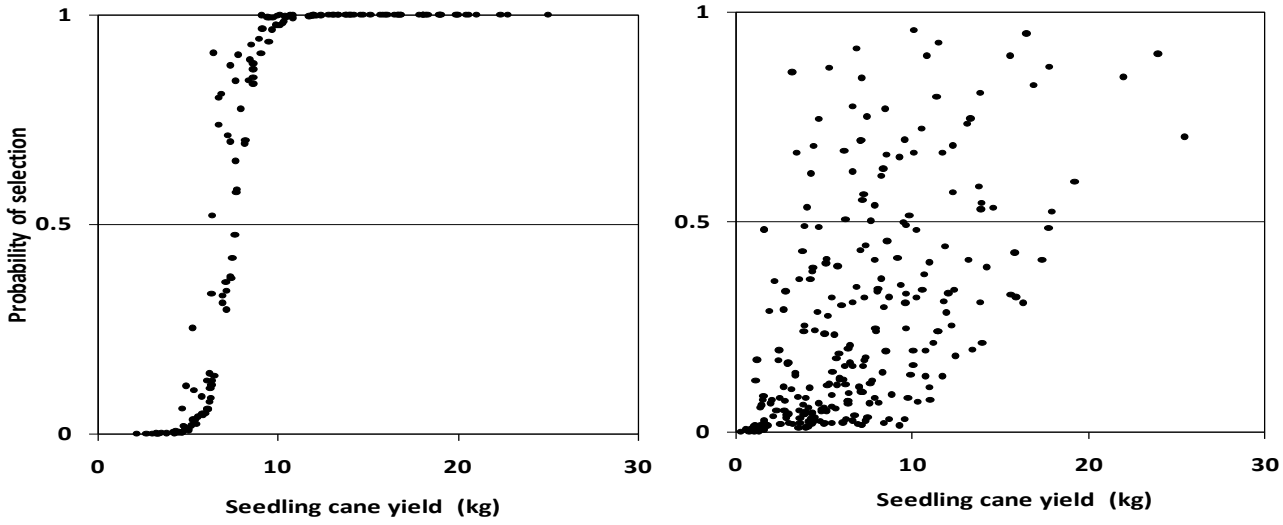

Figure 3.3: The logistic cumulative distribution functions for estimated seedling cane yield (kg) (*x*-axis) plotted against selection probabilities (*y*-axis) for the LSU AgCenter (a) and USDA (b) populations.

3.3.3 Probabilities and Seedling Selection

The probability is calculated by plugging in the input values of stalk number, stalk height, and stalk diameter of each seedling into the logistic regression function (Equations 3.2 and 3.3). The probability is larger for higher values of input variables and smaller for lower values. Only one probability (P) can be modeled, in this case the probability to select. The probability to reject is, therefore, 1–P. To predict the response, a threshold probability must be specified. If the select probability is modeled, the response would be to select when the probability is equal to or greater than the threshold and to reject when the probability is less than the threshold. In SAS ANN models, the default threshold is 0.5. Larger thresholds produces more stringent selection criteria and vice versa.

The probability to select was calculated using Equations 3.2 (LSU AgCenter) and 3.3 (USDA). In Tables 3.4 and 3.5, P(Y = 1) was the probability to select and P(Y = 0) was the probability to reject. The threshold probability was 0.5. When P(Y = 1) was equal to or greater than 0.5, the seedling was selected and categorized into select (1), otherwise it was categorized into reject (0). The column labeled ANN in Tables 3.4 and 3.5 represents the predicted categories. The input variables are included in the output, and can aid the breeder to decide border line seedlings. Other variables such as Brix, disease or insect damage can also be included to aid selection.

Generally, seedlings were selected at higher probability from the LSU AgCenter population (Table 3.4) than the USDA population (Table 3.5). Eighteen out of the 30 seedlings (LSU AgCenter population) were selected with probabilities ranging from 0.58 to 1.00 (mean = 0.88). Nine out of the 30 seedlings (USDA population) were selected with probabilities ranging from 0.53 to 0.91 (mean = 0.72). This indicated a greater precision of selection from the LSU

AgCenter population compared to the USDA population, reflecting the effect of relative variability depicted in Figure 3.3. In Figure 3.3(a) (LSU AgCenter population), a threshold of 0.5 selected seedlings with estimated seedling cane yield greater than 7.5 kg while, in Figure 3.3 (b) (USDA population), the same threshold selected seedlings with estimated seedling cane yield as low as 3.5 kg.

3.3.4 Discriminating Ability of the Artificial Neural Network Models Versus Visual Selection

The means of the seedling stalk number, stalk height, stalk diameter, and estimated cane yield were calculated for each group of selected and rejected seedlings. The difference between the means of the selected and rejected seedlings was calculated and expressed as a percentage of the rejected seedlings (Tables 3.6). This metric was used to describe and evaluate the discriminating ability of the ANN models and the visual method. A larger percentage of the difference between the means of the selected and rejected seedlings was used as an indicator of greater discriminating ability.

The ANN models produced greater discrimination between the selected and rejected seedlings than the visual method (Tables 3.6, Figure 3.4). The ANN models were twice (LSU AgCenter population) and 1.5 times (USDA population) more discriminating between the selected and rejected seedlings than the visual method. The seedlings selected by the ANN models produced more stalks than those selected by the visual method. These selected seedlings also produced stalks with greater diameter for both populations and taller stalks for the USDA population.

Table 3.4: Probability of selecting [P(Y = 1)] or rejecting [P(Y = 0)] a seedling, the predicted selection decision by the artificial neural network (ANN) model, the selection decision by the visual method (visual), stalk number, stalk height (height), stalk diameter (Diameter), and seedling cane yield (Cane) for the first 30 seedlings derived from the LSU AgCenter population.

Seedling	P(Y=1)	P(Y=0)	ANN	Visual	Number of stalks	Height (m)	Diameter (cm)	Cane (kg)
1	1	0	1	1	23	2.37	1.70	12.38
2	0.09	0.91	0	0	11	2.25	1.68	5.49
3	0.31	0.69	0	0	9	2.50	1.93	6.59
4	1	0	1	1	25	2.40	2.12	21.19
5	1	0	1	1	20	2.50	1.70	11.35
6	0.01	0.99	0	0	12	1.90	1.51	4.08
7	0	1	0	0	12	1.70	1.34	2.88
8	1	0	1	1	16	2.40	1.84	10.21
9	0	1	0	0	4	2.40	2.11	3.36
10	0.87	0.13	1	1	11	2.35	2.00	8.12
11	0.03	0.97	0	0	10	2.40	1.63	5.01
12	1	0	1	1	17	2.60	2.33	18.85
13	0.91	0.09	1	1	10	2.40	2.13	8.56
14	0.13	0.87	0	0	6	2.40	2.25	5.73
15	1	0	1	1	14	2.45	2.08	11.66
16	0.9	0.1	1	0	14	2.20	1.75	7.41
17	1	0	1	1	23	2.30	1.90	15.00
18	0.34	0.66	0	0	10	2.20	1.98	6.78
19	0.01	0.99	0	0	10	2.05	1.72	4.77
20	0	1	0	0	8	2.40	1.68	4.26
21	0.91	0.09	1	1	16	2.30	1.45	6.08
22	1	0	1	1	19	2.30	1.72	10.16
23	0.33	0.67	0	1	11	2.15	1.88	6.57
24	1	0	1	1	13	2.55	2.23	12.95
25	0.58	0.42	1	1	10	2.35	1.98	7.24
26	0	1	0	0	8	2.30	1.48	3.17
27	1	0	1	1	19	2.40	1.63	9.52
28	0.74	0.26	1	1	14	2.30	1.58	6.32
29	1	0	1	1	18	2.20	1.66	8.57
30	0.84	0.16	1	1	13	2.35	1.74	7.27

Table 3.5: Probability of selecting [P(Y = 1)] or rejecting [P(Y = 0)] a seedling, the predicted selection decision by the artificial neural network (ANN) model, the selection decision by the visual method (visual), stalk number, stalk height (height), stalk diameter (Diameter), and seedling cane yield (Cane) for the first 30 seedlings derived from the USDA population.

Seedling	P(Y=1)	P(Y=0)	ANN	Visual	Number of stalks	Height (m)	Diameter (cm)	Cane (kg)
1	0.53	0.47	1	0	19	2.31	1.95	13.11
2	0.05	0.95	0	0	10	1.88	1.67	4.12
3	0.08	0.92	0	0	22	1.68	1.78	9.20
4	0.53	0.47	1	1	20	2.06	2.07	13.87
5	0.33	0.67	0	0	19	2.01	1.95	11.41
6	0.05	0.95	0	0	6	1.88	1.72	2.62
7	0.12	0.88	0	0	15	2.06	1.73	7.27
8	0.11	0.89	0	0	5	2.08	1.77	2.56
9	0.75	0.25	1	1	15	2.24	2.18	12.55
10	0.07	0.93	0	0	4	1.78	1.83	1.87
11	0.48	0.52	0	0	14	1.93	2.13	9.63
12	0.91	0.09	1	1	6	2.39	2.40	6.49
13	0.03	0.97	0	0	11	1.63	1.70	4.07
14	0.85	0.15	1	1	24	2.08	2.30	20.75
15	0.29	0.71	0	0	7	2.29	1.87	4.40
16	0.02	0.98	0	0	11	1.63	1.67	3.93
17	0.35	0.65	0	1	10	2.13	1.97	6.49
18	0.02	0.98	0	0	15	1.91	1.45	4.73
19	0.28	0.72	0	0	20	2.26	1.78	11.25
20	0.23	0.77	0	0	9	2.16	1.86	5.28
21	0.13	0.87	0	0	11	2.03	1.78	5.56
22	0.87	0.13	1	0	5	2.16	2.43	5.01
23	0.75	0.25	1	0	5	2.49	2.13	4.44
24	0.14	0.86	0	0	10	2.24	1.72	5.21
25	0.12	0.88	0	0	2	1.85	1.93	1.08
26	0.29	0.71	0	0	4	2.24	1.92	2.60
27	0.11	0.89	0	0	24	2.11	1.62	10.44
28	0.48	0.52	0	0	2	2.24	2.08	1.52
29	0.70	0.3	1	1	8	2.11	2.25	6.71
30	0.61	0.39	1	0	10	2.29	2.08	7.78

Further evaluation of the discriminating ability was done for each of the five families from the LSU AgCenter population (Table 3.7). The ANN model produced greater discrimination between the selected and rejected seedlings than the visual method for all the families. The seedlings selected by the ANN model also produced more stalks that were thicker than those selected by the visual method. The magnitude of the discrimination of the ANN model was greater than that of the visual method in situations where the ANN model selected more seedlings than the visual method for example, families XL01-001, XL01-050, XL01-059, and XL01-460 (Table 3.7). When the number of seedlings selected by both methods was equal, for example, family XL01-215, the discriminating ability of the ANN model was similar to that of the visual method.

Table 3.6: The means for stalk number, stalk height, stalk diameter and cane yield for seedling selected (S) and rejected (R) by visual selection and artificial neural network models and the means expressed as a percent of rejected seedlings ((S-R)/R %) for the LSU AgCenter and USDA populations.

Population	Trait	Visual Selection			Artificial Neural Networks		
		Rejected	Selected	(S-R)/R%	Rejected	Selected	(S-R)/R %
LSU AgCenter	Stalks	9.74	15.58	60	7.83	14.28	82
	Height (m)	2.11	2.28	8	2.16	2.19	1
	Diameter (cm)	2.17	2.17	0	1.99	2.24	13
	Cane (kg)	7.62	12.62	66	5.08	12.01	136
USDA	Stalks	12.17	11.89	-2	11.89	13.08	10
	Height (m)	2.07	2.22	8	2.05	2.26	10
	Diameter (cm)	1.73	2.13	24	1.70	2.17	27
	Cane (kg)	6.02	9.37	56	5.65	10.44	85

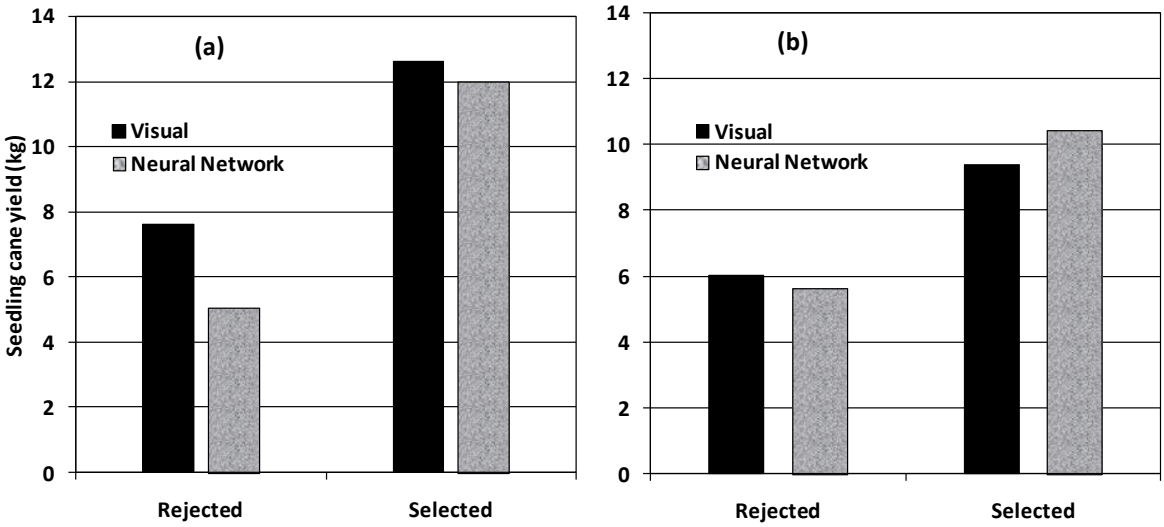

Figure 3.4: Comparison of mean cane yield (kg) for the seedlings selected and rejected using visual and artificial neural network models for the LSU AgCenter (a) and USDA (b) populations.

3.3.5 Selection Efficiency of the Artificial Neural Network Models Versus Visual Selection

Improving selection efficiency is a challenge shared by sugarcane breeders. Selection efficiency is the ability to discard a seedling that would eventually produce low cane yield and or select a seedling that would produce high cane yield. From the LSU AgCenter population, the visual method selected 57 seedlings while the ANN model selected 96 seedlings. Three out of the 57 seedlings selected by the visual method were rejected by the ANN model. The ANN model selected an additional 42 seedlings from those seedlings rejected by the visual method. From the USDA population, the visual method selected 46 seedlings while the ANN model selected 53 seedlings. Thirteen of the 46 seedlings selected by the visual method were rejected by the ANN model. An additional 20 seedlings were selected by the ANN model from those seedlings rejected by the visual method. The means of seedlings selected by the visual method and rejected by the ANN models, and the means of seedlings rejected by the visual method and selected by the ANN models were calculated (Table 3.8).

Table 3.7: The difference between the means of the selected and rejected seedlings expressed as a percent of the rejected seedlings for the seedlings selected using the visual method (Visual) and the artificial neural network model (ANN) for stalk number, stalk height, stalk diameter and cane yield and the number of seedlings selected (Number selected) for the individual crosses derived from the LSU AgCenter population.

Trait	XL01-001		XL01-050		XL01-059		XL01-215		XL01-460	
	Visual	ANN	Visual	ANN	Visual	ANN	Visual	ANN	Visual	ANN
Stalk number	89	104	72	76	50	88	71	77	45	60
Stalk height (m)	-1	-1	9	5	4	9	7	7	0	2
Stalk diameter (cm)	3	6	7	16	2	6	7	6	-14	14
Cane yield (kg)	104	126	115	166	59	144	119	126	73	100
Number selected	16	21	6	16	10	14	18	18	7	27

The seedlings selected by the ANN models and rejected by the visual method produced 75 % (LSU AgCenter population) and 51 % (USDA population) more cane yield than seedlings rejected by the ANN models and selected by the visual method (Table 3.8, Figure 3.5). The seedling selected by the ANN model and rejected by the visual method produced 22 % (LSU AgCenter population) and 30 % (USDA population) more stalks that were generally thicker and taller (USDA population) than seedlings rejected by the ANN model and selected by the visual method.

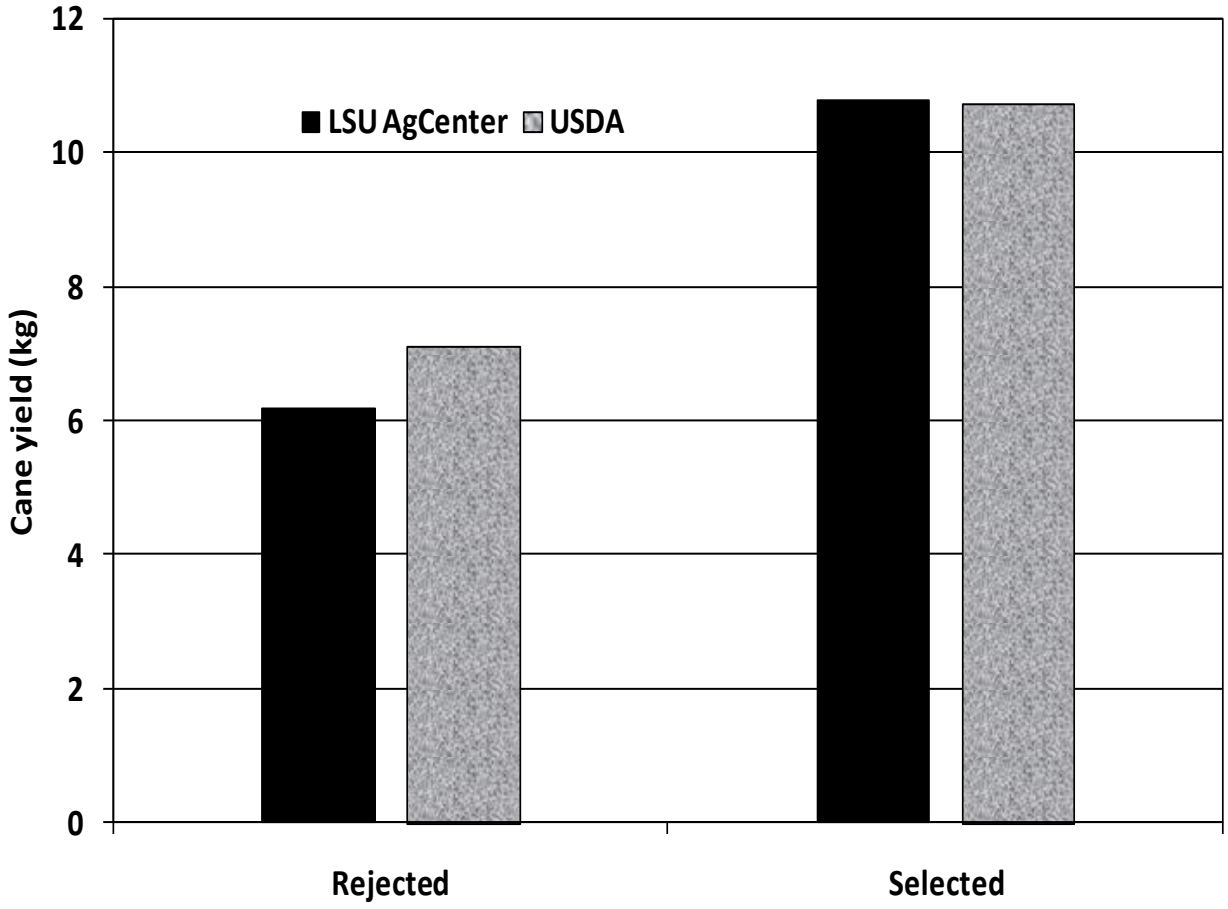

Figure 3.5: The mean cane yield (kg) for the seedlings rejected by the artificial neural network (ANN) model and selected by the visual method (Rejected) and seedlings selected by the ANN model and rejected by the visual method (Selected) for the LSU AgCenter and USDA populations.

Table 3.8: The means of the rejected and selected seedlings, and the difference of the means of selected and rejected seedlings expressed as a percent of rejected seedlings ((S-R)/R%) for stalk number, stalk height, stalk diameter and seedling cane yield for the LSU AgCenter and USDA populations.

Trait	LSU AgCenter			USDA		
	Rejected	Selected	(S-R)/R%	Rejected	Selected	(S-R)/R%
Stalk number	10.00	12.21	22	11.31	14.65	30
Stalks height (m)	2.15	2.05	-4	2.12	2.26	7
Stalk Diameter (cm)	1.91	2.39	25	1.97	2.12	8
Cane yield (kg)	6.17	10.78	75	7.09	10.74	51

†Rejected refers to seedlings selected by the visual method and rejected by the ANN model.
‡Selected refers to seedlings rejected by the visual method and selected by the ANN model.

The number of seedlings that were selected by each method and produced less cane yield than the population mean, and the number of seedlings that produced higher cane yield than the population mean and were rejected were counted for both populations. From the LSU AgCenter population, all the seedlings that produced higher cane yield than the population mean were selected by the ANN model. From the same population, the visual method rejected 21 seedlings that produced higher cane yield than the population mean. The ANN and the visual method erroneously included similar numbers of low yielding seedlings in the select category although most of the seedlings included by the ANN model had lower probability of selection and could have been rejected by raising the threshold probability. From the USDA population, the ANN model selected 12 seedlings that produced lower cane yield than the population mean while the visual method selected 14 seedlings that produced lower cane yield than the population mean. The ANN model rejected 79 seedlings that produced higher cane yield than the population mean while the visual method rejected 88 seedlings that produced higher cane yield than the population mean. The seedlings rejected by the ANN model could have been selected by lowering the threshold probability since they were rejected with marginally lower probability

than the threshold. Generally, the visual method rejected more higher yielding and included more lower yielding seedlings than the ANN model, indicating lower selection efficiency. The ANN model was also more efficient than the visual method when selecting from the more variable USDA population. Recognizing that the USDA data had a poor fit to the model (Table 3.3, Figure 3.3(b)), one could, after gaining experience, learn to adjust the threshold probability for the ANN model to further improve selection efficiency when dealing with this type of population.

3.3.6 Seedling Cane Yield Increased With Increasing Selection Probabilities

We investigated the relationship between the probability and estimated seedling cane yield. The ANN output data were ranked in ascending order of probability. The 150 seedlings (LSU AgCenter) and the 272 seedlings (USDA) were divided into 10 groups each. Group 1 had the lowest probability and group 10, the highest. The means of each group for each trait were calculated. The means (*y*-axis) were plotted against group probability rankings (*x*-axis). The trends for cane yield and stalk number from the LSU AgCenter population were similar and increased with probability rankings (Figure 3.6(a)). The trends for stalk height and diameter were less similar to that for cane yield and marginally increased with probability rankings. From the USDA population, the trends for stalk diameter and stalk height were very similar to that for cane yield and increased with probability rankings (Figure 3.6 (b)). The trend for stalk number fluctuated and showed no clear pattern across probability rankings.

3.3.7 Artificial Neural Network Models Versus Visual Method at Identical Selection Rates

Comparison of the ANN model and the visual method at different selection rates likely obscured their impact on selection. A balanced comparison should use identical selection rates. Therefore,

to produce a balanced comparison of the ANN models and the visual method during seedling selection, identical selection rates were used. From the LSU AgCenter population, 57 out of 150 seedlings (38 %) were selected by the visual method while from the USDA population, 46 out of 272 (17 %) were selected. The ANN model selected 96 out of the 150 seedlings (64 %) from the LSU AgCenter population and 53 out of the 272 seedlings (19 %) from the USDA population. To produce identical comparisons, the visual selection rates were used as standard for the ANN models. The number of seedlings selected by the ANN model was adjusted to equal that of the visual method by ranking the probability and adjusting the probability threshold. The means of the highest 38 % for the LSU AgCenter population and 17 % for the USDA population were used for the comparison (Table 3.9). The seedlings selected by the ANN model produced 16 % (LSU AgCenter population) and 8 % (USDA population) more cane yield than those selected by the visual method. The seedlings selected by the ANN model produced 8 % more stalks that were thicker than those selected by the visual method.

Table 3.9: The means for stalk number, stalk height, stalk diameter and estimated seedling cane yield of seedlings selected by the artificial neural network models (ANN) and the visual method (Visual), and of seedlings selected by the ANN method expressed as a percent of seedlings selected by the visual method (ANN % Visual) for the LSU AgCenter (38 % selection rate) and USDA (17 % selection rate) populations.

Trait	LSU AgCenter			USDA		
	Visual	ANN	ANN % Visual	Visual	ANN	ANN % Visual
Stalk number	15.58	16.77	108	11.89	12.87	108
Height (m)	2.28	2.25	98	2.22	2.28	102
Diameter (cm)	2.12	2.24	106	2.13	2.19	103
Cane yield (kg)	12.62	14.65	116	9.37	10.45	108

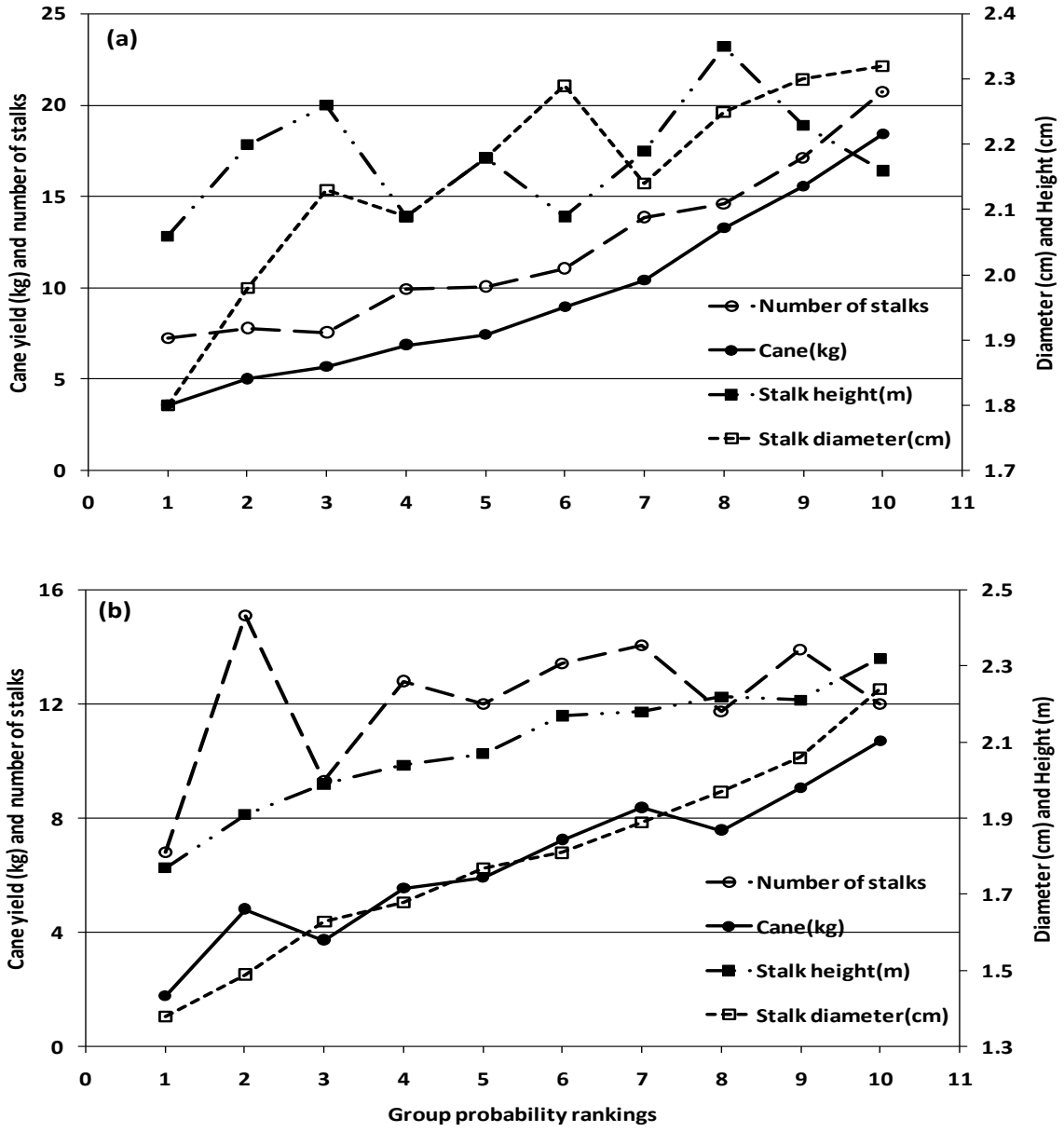

Figure 3.6: Trends for means of seedling stalk number, stalk height, stalk diameter, and cane yield (kg) (y-axis) plotted against the group probability rankings (x-axis) for the LSU AgCenter (a) and USDA (b) populations.

3.4 Discussions

The ANN model was superior to visual selection in identifying seedlings with high cane yield potential as evidenced from several comparisons between the two selection methods. For

example, the proportion of high yielding seedlings selected by the ANN model was greater than that selected by the visual method. This proportion increased when similar selection rates were used for both methods. Generally, seedlings selected by the ANN model produced more stalks that were thicker and taller than those selected by the visual method. The visual method rejected a greater proportion of seedlings that produced estimated cane yields higher than the population mean compared to the ANN model. A good number of these seedlings rejected by the visual method were selected by the ANN model. Conversely, the ANN model rejected low yielding seedlings that were selected by the visual method. Because only a limited number of seedlings can be advanced to the next stage, the low efficiency of the visual method would greatly reduce the overall efficiency of a breeding program. The ANN model uses fast and automated computations and was superior to the visual method even for a variable data set with poor model fit such as was the situation with the USDA population. A good aspect of the ANN model is that as the breeder gains in experience, they will be in a better position to recognize data with a poor model fit and adjust the probability threshold accordingly.

The effects of genotype by environment interaction are known to be large in sugarcane particularly for cane yield because of competition effects (Jackson and McRae, 2001) and the fact that cane yield is controlled by quantitative genes (Falconer and Mackay, 1996; Jackson and McRae, 1998). Genotype by environment interaction is expected to be even larger in non-replicated seedling plots. During seedling selection, it is also not possible to precisely pin point the seedlings that will eventually produce high clonal cane yield. Therefore, sugarcane breeders are inadvertently discarding low yielding seedlings rather than directly selecting for high yielding ones. Erroneously advancing seedlings that should have been rejected increases the costs of the clonal evaluation stages and reduces the efficiency as more clones are handled than

is necessary. Our data suggest that the ANN model was better at rejecting low cane yielding seedlings than the visual method.

The ANN model selected seedlings based on those traits that exhibited the largest variability within the population. Conversely, the traits with the low variability would be less associated with the estimated seedling cane yield and would have little influence in determining the probability of seedling selection. This aspect of the ANN models confirmed the long held view by breeders to base their selection on traits exhibiting the greatest genetic variability. Genetic variability creates the best opportunity for selection (Allard, 1960). Therefore, the ability of the ANN models to use the most genetically variable traits during seedling selection leads to higher selection efficiency than is the case with the visual method. In this study, the ANN models selected seedlings that produced more stalk numbers than visual selection. Research done on early selection stages in Zimbabwe showed that the stalk numbers was positively associated with cane yield (Zhou, 2004b). In Louisiana, seedlings producing high stalk numbers are routinely selected to enhance cane yield and ratooning ability (Milligan *et al*., 1990).

Since land is always a limiting resource in most breeding programs, the breeder has little choice but to design the best allocation of resources. The ANN model offers the breeder greater flexibility for adjusting the numbers of seedlings to advance during seedling selection. The breeder can increase or decrease the number of seedlings to advance by decreasing or increasing the threshold probability, respectively. These adjustments can be used to refine selection using the trait values that can be included in the output, for example, disease and insect resistance scores that were not used in developing the ANN prediction model. Other traits such as Brix can be added to the model. To reduce the number of seedlings to be advanced using the visual method, the breeder will have to go back to the field and review all the selected seedlings and

decide on the seedlings to discard. To increase the number of seedlings to be advanced, the breeder will have an equally daunting task of physically reviewing all the rejected seedlings in order to identify those seedlings that have to be included. With the ANN model, adjusting the numbers can be done easily using the probability of selection and the associated trait values of the seedlings.

A disadvantage of the ANN model for seedling selection is the required measurements of variables such as stalk number, stalk height, stalk diameter, Brix, disease and pest resistance. Most breeding programs cannot afford to measure these variables because of the high cost associated with the manual labor in some of these countries. However in some programs, for example the Zimbabwe sugarcane breeding program, these variables are routinely measured during visual selection and used for adjusting the numbers to advance (Zhou, 2004a). Measurement costs can be reduced by excluding seedlings that are too inferior and would probably never be selected. Visual scores for stalk numbers, stalk height, and stalk diameter can be used as input variables. The scores are easier and quicker to collect but will reduce precision. Scores may be more useful as a validation tool and their precision may improve with time as staff become more experienced.

The other disadvantage and limitation of the ANN model is their strong dependence on the amount, suitability and precision of measurements of the training data (Pandolfi *et al*., 2009). Pandolfi *et al*. (2009) noted that the ANN model training data should capture the variation in the population to attain the best results. However, Pandolfi *et al*. (2006) noted that even when the parameters of the training data were not statistically representative of the target population, the neural network models appeared capable of generalizations beyond the training data and produced correct results even in different populations. Pandolfi *et al*. (2009) applied the ANN

model for the classification of tea accessions. In their study, it was important to capture the variation in the population. This may not be entirely necessary during selection as the intent is to shift the trait values, in this case, cane yield in one direction.

In sugarcane selection, the training data can be collected from part of the seedling population or from special populations created from some of the elite families. This population of elite families would constitute a reservoir of the ideal trait combinations. With selection, the objective is to shift the population towards a desired direction of trait values, such as high cane yield. As previously indicated, therefore the ideal training data need not have similar variability to that of the target population. Rather, the training data should be a population with the desired combination of trait values that will be mimicked by the selection process. In this case, the ANN models provide the added advantage of allowing the breeder to directionally shift the population towards high cane yield more objectively than the visual method.

3.5 Conclusions

The greatest challenge facing sugarcane breeders is the identification of seedlings with high cane yield potential. The seedling stage is planted to a large number of single seedlings that are not replicated and visual selection is used as a proxy for cane yield. The ANN model is a statistical tool that can be used to increase selection efficiency at this stage. The ANN model requires the measurement of yield component traits such as stalk number, stalk height and stalk diameter and these are used as input variables to the logistic regression equations that compute probabilities that are used to decide whether to select or reject a seedling.

The ANN model was superior to the visual method in discriminating between the seedlings with high and low cane yield. The magnitude of the difference between the selected and rejected seedlings was greater for the ANN model than the visual method. The magnitude of

the difference increased when similar selection rates were applied for the visual method and the ANN model. The computations in the ANN model are automated by the SAS software and fast, and therefore large numbers of seedlings can be evaluated quickly. The output in the neural network models provides a decision to select or reject a seedling based on a threshold probability that is user-defined. Yield component traits with high variability have a greater influence in determining threshold probabilities thus mimicking what breeders try to achieve during selection. Although the ANN model was demonstrated on the seedling stage using cane yield components, traits such as Brix (that are less affected by competition in small plots (Jackson and McRae, 2001)) and insect and disease resistance can be added to improve the training data and the model. Additionally, the model can be applied across all the stages of a breeding program, and would be particularly useful in the non-replicated stages.

3.6 References

Agresti, A. (2007). An Introduction to Categorical Data Analysis. Second Edition. John Wiley and Sons, USA.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* AIC-19: 716 – 723.

Allard, R.W. (1960). Principles of Plant Breeding. John Wiley and Sons. New York. USA.

Allison, P.D. (2003). Logistic Regression Using the SAS System: Theory and Applications. SAS Institute Inc., Cary, NC, USA.

Breaux, R.D. and Miller, J.D. (1987). Seed handling, germination and seedling propagation. In DJ Heinz (editor). Sugarcane Improvement Through Breeding. Elsevier.: 385 – 407.

Casella, G. and Berger, R.L. (2003). Statistical Inference. 2nd Edition. Thomson publications.

Chang, Y.S. and Milligan, S.B. (1992). Estimating the potential of sugarcane families to produce elite genotypes using univariate cross prediction methods. *Theoretical and Applied Genetics* 84: 662 – 671.

Cox, M.C. and Stringer, J.K. (1998). Efficacy of early generation selection in a sugarcane improvement program. *Proceedings of the Australian Society of Sugarcane Technologist* 20: 148 – 153.

De Sousa-Vieira, O. and Milligan, S.B. (1999). Intra-row spacing and family x environment effects on sugarcane family evaluation. *Crop Science* 39: 358 – 364.

Fausett, L. (1994). Fundamentals of Neural Networks. Prentice-Hall. Englewood Cliffs. New Jersey. USA.

Falconer, D.S. and Mackay, T.F.C. (1996). Introduction to Quantitative Genetics. Fourth Edition. Longman Group Ltd, UK.

Gravois, K.A., Milligan, S.B. and Martin, F.A. (1991). Indirect selection for increased sucrose yield in early sugarcane testing stages. *Field Crops Research* 26: 67 – 73.

Gurney, K. (1997). An Introduction to Neural Networks. UCL Press.

Hertz, J., Palmer, R.G. and Krogh, A.S. (1990). Introduction to the theory of neural computation. Persus Books.

Hogarth, D.M. and Berding, N. (2006). Breeding for a better industry: Conventional breeding. *Sugarcane International* 24 *Number* 2: 26 – 31.

Huang, S., Tan, K.K. and Tang, K.Z. (2004). Neural Network: Theory and Applications. Research Studies Press Ltd. Baldock, Hertfordshire, England.

Iglesias, A., Arcay, B. and Cotos, J.M. (2006). Connectionist systems for fishing prediction. In: Rabunal, J.R. and Dorado, J. (editors). Artificial Neural Networks in Real-Life Applications. Idea Group Publishing, London: 265 – 296.

Jackson, P.A. and McRae, T.A. (1998). Gains from selection of broadly adapted and specifically adapted sugarcane families. *Field Crops Research* 59: 151 – 162.

Jackson, P.A. and McRae, T.A. (2001). Selection of sugarcane clones in small plots: Effects of plot size and selection criteria. *Crop Science* 41: 315 – 322.

Kimbeng, C.A. and Cox, M.C. (2003). Early generation selection of sugarcane families and clones in Australia: A Review. *Journal of the American Society of Sugarcane Technologists* 23: 20 – 39.

Lee, K.C. and Paik, T.Y. (2006). A neural approach to cost minimization in a production scheduling setting. In: Rabunal, J.R. and Dorado, J. (editors). Artificial Neural Networks in Real-Life Applications. Idea Group Publishing, London: 297 – 313.

Masters, T. (1993). Practical Neural Network Recipes in C++. Academic Press Inc., San Diego, USA.

Miller, J.D. and James, N.I. (1974). The influence of stalk density on cane yield. *Proceedings of the International Society of Sugarcane Technologists* 15: 177 – 184.

Milligan, S.B., Gravois, K.A., Bischoff, K.P. and Martin, F.A. (1990). Crop effects on broad sense heritabilities and genetic variances of sugarcane yield components. *Crop Science* 30: 344 – 349.

Mugnai, S., Pandolfi, C., Azzarello, E., Masi, E. and Mancuso, S. (2008). *Camellia japonica* L. genotypes identified by an artificial neural network based on phyllometric and fractal parameters. Plant Systematics and Evolution 270: 95 – 108.

Nelson, M.N. and Illingworth, W.T. (1991). A Practical Guide to Neural Nets. Addison-Wesley Publishing Company, Inc.

Pandolfi, C., Mugnai, S., Azzarello, E., Masi, E. and Mancuso, S. (2006). Fractal geometry and neural networks for the identification and characterization of ornamental plants. In: Teixeira da Silva (editor). Floriculture, Ornamental and Plant Biotechnology: Advances and Topical Issues. Volume IV. Global Science Books, Kyoto: 213 – 225.

Pandolfi, C., Mugnai, S., Bergamasco, S., Masi, E. and Mancuso, S. (2009). Artificial neural networks as a tool for plant identification: a case study on Vietnamese tea accessions. *Euphytica* 166: 411-21.

Perkins, R. and Brabazon, A. (2006). Predicting credit ratings with a GA-MLP hybrid. In: Rabunal, J.R. and Dorado, J. (editors). Artificial Neural Networks in Real-Life Applications. Idea Group Publishing, London: 220 – 237.

SAS Institute (2007). The SAS System for Windows Version 9.1.3. SAS Institute, Cary, North Carolina, USA.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* 6: 461 – 464.

Sethuraman, J. (2006). Soft computing approach for bond rating prediction. In: Rabunal, JR and Dorado, J (editors). Artificial Neural Networks in Real-Life Applications. Idea Group Publishing, London: 202 – 219.

Smith, M. (1993). Neural Networks for Statistical Modelling. Van Nostrand Reinhold. New York.

Zhou, M. (2004a). Strategies for variety selection in the breeding program at the Zimbabwe Sugar Association Experiment Station. *Proceedings of the South African Sugar Technologists Association* 78: 125 - 131.

Zhou, M. (2004b). Stalk population control of yield, quality and agronomic traits of sugarcane population in early selection stages. *Sugar Cane International*: *Volume* 22 *number* 5: 14 – 20.

CHAPTER 4: LOGISTIC REGRESSION MODELS: A DECISION SUPPORT STATISTICAL TOOL FOR ENHANCING SEEDLING SELECTION IN SUGARCANE BREEDING

4.1 Introduction

The correct identification of seedlings with the potential to produce high cane yield is a major challenge faced by sugarcane breeders during selection. Currently, visual appraisal of seedlings for cane yield using stalk number, stalk height and stalk diameter is used for individual seedling selection. Visual selection is subjective (Cox and Stringer, 1998) and therefore can be inefficient (Hogarth and Berding, 2006). The confounding effect of genotype by environment interaction and the competition among closely spaced seedlings reduces the efficiency of visual selection. The influence of these confounding effects to seedling selection cannot be resolved by replication. The large number of seedlings planted meant that there would be insufficient land to plant the seedlings in replicated plots. At planting, each seedling is represented by one plant and therefore replication is also practically impossible because of limited planting material. Closely spaced seedlings alter phenotypic expression for stalk number, stalk height, and stalk diameter (De Sousa-Vieira and Milligan, 1999) making the visual identification of high cane yielding seedlings less precise. Sugarcane breeders generally plant closely spaced seedlings because of limited land and the need to plant large number of seedlings to enable high selection intensity (Breaux and Miller, 1987). Large numbers of seedlings are planted in order to capture the transgressive segregants that combine the unique and desirable traits. While family selection identified high cane yield crosses (Cox and Hogarth, 1993; Hogarth and Mullins, 1989; Kimbeng *et al*., 2001b), seedling selection from the elite crosses has remained inefficient because of the dependence on visual appraisal for yield (Kimbeng *et al*., 2001a).

Path coefficient analysis studies (De Sousa-Vieira and Milligan, 2005; Kang *et al*., 1983, 1989; Milligan *et al*., 1990; Singh *et al*., 2005) has proven the contributions of the yield components (stalk number, stalk height and stalk diameter) to cane yield. However, despite this knowledge, the yield components are not directly used in selection. One of the reasons for this could be the expense involved in measuring these yield components and, the unavailability of appropriate statistical models that incorporate the yield components to create a decision support tool that can be used for individual seedling selection. To increase seedling selection efficiency, Hogarth and Berding (2006) suggested that sugarcane breeders should explore and adapt available statistical decision support tools. In this study, we propose and demonstrate the potential of using the logistic regression model as a statistical decision support tool for individual seedling selection. The logistic regression model would use the stalk number, stalk height and stalk diameter as predictor variables. The output from the model, a probability would be used as the decision support tool for deciding to either select or reject a seedling. This statistical decision support tool is expected to reduce the bias and subjectivity associated with the visual appraisal method during seedling selection for high cane yield.

The objective of this study was to introduce and demonstrate the utility of the logistic regression models as a statistical decision support tool for sugarcane seedling selection.

4.1.1 Statistical Considerations in Logistic Regression Models

Logistic regression models are part of the generalized linear models that are used to predict the probability of occurrence of binary events by fitting the data of predictor variables to a logistic curve (Agresti, 2007). Generalized linear models are made up of three components namely the random, the systematic, and the link function (Casella and Berger, 2003). The response variable

is the random component and in this study, it was the decision to either select or reject a seedling, which is binary. The systematic component is a linear function of predictor variables and in this study would be a function of the stalk numbers, stalk height and stalk diameter. The predictor function would follow a multiple linear regression. The link function is the logit or logistic transformation. The link function is used to linearize the relationship between the random component (the binary response variable) and the systematic component (Allison, 2003).

Logistic regression models can use both numerical and categorical predictor variables (Le, 1998). Logistic regression models have been used extensively in medical research to predict the onset of diseases (Hosmer and Lemeshow, 1989), in political sciences to determine opinions for candidates (Cohen, 2006), and in education to predict pass or fail of students (Bowie, 2006). The logistic regression cumulative distribution function is,

$$\boldsymbol{\pi(x_{i1}, x_{i2}, x_{i3}) = \frac{e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\beta_3 x_{i3}}}{1+e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\beta_3 x_{i3}}}.} \qquad \text{Equation 4.1}$$

By re-arranging the terms, Equation 4.1 can be expressed as,

$$\boldsymbol{log\left[\frac{\pi(x_{i1},x_{i2},x_{i3})}{1-\pi(x_{i1},x_{i2},x_{i3})}\right] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3},} \qquad \text{Equation 4.2}$$

where $\boldsymbol{\pi(x_{i1}, x_{i2}, x_{i3})}$ is the probability of selecting the *i*th seedling (*i* = 1, 2, …, *n*), $\boldsymbol{x_{i1}}$ is the *i*th seedling stalk number, $\boldsymbol{x_{i2}}$ is the *i*th seedling stalk height, and $\boldsymbol{x_{i3}}$ is the *i*th seedling stalk diameter, $\boldsymbol{\beta_0}$ is the intercept of the predictor function, $\boldsymbol{\beta_1}$ is the coefficient of the stalk number, $\boldsymbol{\beta_2}$ is the coefficient of the stalk height, $\boldsymbol{\beta_3}$ is the coefficient of the stalk diameter. The $\boldsymbol{log}$ transforms the odds to produce log of odds. It is the log of odds that are modelled by the multiple linear regression function in Equation 4.2 (Hosmer and Lemeshow, 1989). The log

transformation linearizes the relationship between the odds and the function of predictor variables.

By re-arranging the terms, Equation 4.2 can be expressed as,

$$\frac{\pi(x_{i1},x_{i2},x_{i3})}{1-\pi(x_{i1},x_{i2},x_{i3})} = e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\beta_3 x_{i3}}, \qquad \text{Equation 4.3}$$

where the odds of selecting a seedling are equal to the exponential of the function of the predictor variables. The probability of selecting a seedling is estimated using Equation 4.1. The confidence intervals of the probability of selecting a seedling are calculated indirectly from the confidence intervals of the log of odds of selecting a seedling. Equation 4.2 is used to calculate the log of odds of selecting a seedling. Let $\hat{L}$ be the log of odds of selecting a seedling where,

$$\hat{L} = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i3}. \qquad \text{Equation 4.4}$$

The coefficients $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$ and $\hat{\beta}_3$ are estimated from a training data set. The variance of $\hat{L}$ is,

$$Var(\hat{L}) = Var(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i3}), \qquad \text{Equation 4.5}$$

which is equal to,

$$Var(\hat{L}) = Var(\hat{\beta}_0) + x_{i1}^2 Var(\hat{\beta}_1) + x_{i2}^2 Var(\hat{\beta}_2) + x_{i3}^2 Var(\hat{\beta}_3) + 2x_{i1} Cov(\hat{\beta}_0, \hat{\beta}_1)$$
$$+2x_{i2} Cov(\hat{\beta}_0, \hat{\beta}_2) + 2x_{i3} Cov(\hat{\beta}_0, \hat{\beta}_3) + 2x_{i1} x_{i2} Cov(\hat{\beta}_1, \hat{\beta}_2) + 2x_{i1} x_{i3} Cov(\hat{\beta}_1, \hat{\beta}_3)$$
$$+2x_{i2} x_{i3} Cov(\hat{\beta}_2, \hat{\beta}_3). \qquad \text{Equation 4.6}$$

Equation 4.6 includes the variances and covariance of the coefficients of the predictor variables. The use of the covariance improves the estimates of the confidence intervals by accounting for the correlation between the predictor variables when computing the variances. The standard error

($\boldsymbol{S.E.}$) is equal to $\sqrt{\boldsymbol{Var}(\hat{\boldsymbol{L}})}$**.** The 95 % confidence limits for $\hat{\boldsymbol{L}}$ is equal to $\hat{\boldsymbol{L}} \pm \mathbf{1.96} * \boldsymbol{S.E.}$**.** The confidence limits for the probability of selecting a seedling is calculated by,

$$\boldsymbol{C.I.}\,\hat{\boldsymbol{\pi}}(\boldsymbol{x_{i1}}, \boldsymbol{x_{i2}}, \boldsymbol{x_{i3}}) = \left(\frac{e^{\hat{L}_{lower}}}{1+e^{\hat{L}_{lower}}}, \frac{e^{\hat{L}_{upper}}}{1+e^{\hat{L}_{upper}}}\right). \qquad \text{Equation 4.7}$$

4.2 Materials and Methods

4.2.1 Experimental Materials and Data Collection

The data were collected from seedlings germinated and grown from true seed. Seedlings from 17 crosses were grown at the United States Department of Agriculture, Agricultural Research Service (USDA-ARS), Sugarcane Research Station at Houma (Table 4.1) and 5 crosses grown at the Louisiana State University Agricultural Center (LSU AgCenter), Sugar Research Station, St. Gabriel, (Table 4.2), Louisiana, USA. The seedlings were transplanted in the summer of 2002, harvested in the fall of 2002 and left to over winter. At the USDA, one set of the individual seedlings was planted in the first replication and the other set in the second replication. Families but not seedlings were replicated. The number of seedlings that survived winter was counted in 2003 (Table 4.1). In each plot, there were two rows, each with 16 seedlings. Eight seedlings (four from each row per plot) were randomly chosen from each plot.

For the LSU AgCenter population, five crosses (Table 4.2), each with more than 500 seedlings, were selected from the breeding program. From the seedlings that survived the winter, thirty seedlings were randomly chosen from each cross. For both the USDA and LSU AgCenter populations, the chosen seedlings were evaluated visually to determine if they would have been selected (1) or rejected (0). The decision was based on a consensus by two experienced sugarcane breeders. The stalk number was counted for each of the chosen seedlings. The stalk

diameter of three stalks from each seedling was measured at the middle of the stalk using a caliper (without reference to the node) and used to estimate a mean. The stalk height of each seedling was measured from the base of the seedling to the top most visible dewlap.

Table 4.1 Series, cross number, female parent, male parent and number of seedlings in replication 1 and 2 of the USDA population.

Series	Family	Female Parent	Male Parent	Rep 1	Rep2
HB00	306	Ho94-856	HoCP96-540	18	17
HB01	3055	HoCP00-945	HoCP99-866	18	17
HB01	3074	HoCP00-950	HoCP96-540	16	19
HB01	3093	HoCP00-945	HoCP96-540	17	18
HB01	3101	HoCP99-866	HoCP96-540	17	16
HB01	3107	HoCP00-950	LCP85-384	18	16
HB01	3111	HoCP99-866	LCP85-384	18	18
HB01	3174	HoCP00-945	LCP85-384	18	18
HB01	3249	N27	LCP85-384	23	11
HB01	3255	HoCP00-945	Ho94-856	18	18
HB01	3256	HoCP00-950	Ho94-856	17	18
HB01	3257	L98-207	Ho94-856	18	18
HB01	3276	TUCCP77-42	HoCP99-866	17	17
HB01	3322	TUCCP77-42	L98-207	18	18
HB01	3328	HoCP91-555	LCP85-384	18	20
HB01	3345	HoCP91-555	L98-207	17	17
HB01	3417	HoCP91-555	TUCCP77-42	18	18

4.2.2 Estimation of Seedling Cane Yield From Yield Components

The seedling cane yield was estimated using the formula used by De Sousa-Vieira and Milligan (1999) (Equation 4.8). Their calculation assumed the sugarcane stalk was a perfect cylinder with

specific gravity of 1.0 g cm^3 (Miller and James, 1974; Gravois *et al*., 1991; Chang and Milligan, 1992).

$$Seedling\ cane\ yield\ (g) = nd\pi r^2 l, \qquad \text{Equation 4.8}$$

where n = seedling stalk number, d = density at 1.0 g cm^{-3}, r = mean stalk radius (cm), and l = seedling stalk length (cm).

Table 4.2: The female and male parent of seedlings from crosses derived from the LSU AgCenter population.

Cross	Female Parent	Male Parent
XL01-001	HoCP92-624	HoCP91-552
XL01-050	LCP86-454	LC85-384
XL01-059	HoCP95-951	HoCP96-540
XL01-215	TucCP77-42	LCP85-384
XL01-460	Ho95-988	L99-238

4.2.3 Data Analysis

The data were analyzed using the logistic procedure of SAS (SAS Institute, 2007). The data were divided into the training data set (30 %) and prediction data set (70 %). The prediction data had the values of the response variable coded as missing. The training data set was used to produce the parameters that were used to build the logistic regression models. The SAS code used for data analysis is shown in Appendix 2.

4.3 Results

4.3.1 Likelihood Ratio, Score and Wald Statistical Tests

The logistic regression analysis (Appendix 2) produced the Likelihood Ratio, the Score and the Wald statistics that were used to test the robustness of the model (Table 4.3). The Likelihood Ratio, the Score and the Wald statistics are generated from the training data set and follow a Chi-square distribution. A significant statistic means that at least one of the predictor variables (stalk numbers, stalk height, stalk diameter) was significantly associated with the response variable (the decision to either select or reject a seedling). The Likelihood Ratio statistic is the most powerful, while the Wald statistic is the least. If any one of the Likelihood Ratio, Score and Wald statistic is not significant, then the model may be unreliable (Agresti, 2007). The Likelihood Ratio, the Score and the Wald statistics were highly significant ($P < 0.01$) for both populations, indicating that at least one of the predictor variables was significantly associated with the response variable (Table 4.3). The Likelihood Ratio statistic (greatest power) produced the largest Chi-square value and the Wald produced the least.

4.3.2 Variable Selection and Logistic Regression Cumulative Distribution Functions

The parameter estimates of the coefficients of the predictor variables (Table 4.4) are generated from the training data set. The coefficients of the predictor variables are interpreted the same way as with the multiple linear regression models. As is the case with multiple linear regression models, the intercept of the model has no meaningful interpretation. A significant coefficient of the predictor variables means that the predictor variable significantly influences the decision to either select or reject a seedling. Higher levels of significance mean higher levels of influence by a predictor variable on the selection decision. As is done with multiple linear regressions,

variable selection is used to eliminate non significant predictor variables from the model. Multi-collinearity among variables is considered during variable selection and occurs when two or more predictor variables are highly correlated. Multi-collinearity leads to variance inflation and variance inflation causes poor predictions. The effects of multi-collinearity are corrected by including only one of the highly correlated predictor variables in the model.

Table 4.3: The Chi-square statistic and the probability of obtaining a larger statistic (P-value) for the Likelihood Ratio, Score and Wald tests for the USDA and LSU AgCenter populations

Statistic	USDA		LSU AgCenter	
	Chi-square	P-value	Chi-square	P-value
Likelihood Ratio	42.64	0.0001	32.67	0.0001
Score	33.77	0.0001	25.20	0.0001
Wald	20.59	0.0001	13.54	0.0011

The parameter estimates of the coefficients of the predictor variables for the USDA and LSU AgCenter populations are shown in Table 4.4. From Table 4.4, the estimate column presents the coefficients for intercept, stalk number, stalk height and stalk diameter. Each coefficient is divided by its standard error to produce a t-statistic. The t-statistic is squared to produce the Chi-square statistic with one degree of freedom. The P-value is the probability of producing a larger Chi-square statistic.

For the USDA population, the stalk number and stalk diameter produced significant coefficients ($P < 0.05$) while stalk height ($P = 0.06$) was not significant (Table 4.4). The stalk diameter had the highest P-value, indicating that it was the most influential predictor variable for determining the decision to either select or reject a seedling compared to stalk number and stalk height. The stalk diameter was not significant ($P = 0.79$) for the LSU AgCenter population,

indicating that it had no significant influence on the decision to either select or reject a seedling. The stalk diameter was removed from the model during the variable selection process. The stalk numbers was highly significant ($P < 0.01$) and was the most influential predictor variable of the decision to either select or reject a seedling. The coefficients of the predictor variables were different between the USDA and the LSU AgCenter populations. The differences reflected the variability of the trait values of seedlings present within each population.

Table 4.4: The estimates, standard errors, chi-square statistic and probability (P-value) of obtaining a larger statistic for the coefficients of the parameters for the intercept, stalk number, stalk height and stalk diameter from the USDA and LSU AgCenter populations.

Population	Parameter	Estimate	Standard Error	Chi-Square	P-value
USDA	Intercept	-23.06	5.31	18.87	0.0001
	Stalk number	0.12	0.05	5.35	0.0207
	Stalk height	3.37	1.79	3.52	0.0605
	Stalk diameter	6.71	1.67	16.24	0.0001
LSU AgCenter	Intercept	-17.34	6.40	7.33	0.0068
	Stalk number	0.36	0.10	12.22	0.0005
	Stalk height	5.44	2.01	7.37	0.0066
	Stalk diameter	0.29	1.10	0.07	0.7936
LSU AgCenter (without Stalk diameter)	Intercept	-16.25	4.71	11.89	0.0006
	Stalk number	0.36	0.10	12.25	0.0005
	Stalk height	5.20	1.75	8.83	0.0030

The parameter estimates of the coefficients of the predictor variables (Table 4.4) were used to build the cumulative logistic regression distribution functions (Equations 4.9 and 4.10) that were in turn used to calculate the probability of selecting a seedling. The cumulative logistic regression distribution functions were constructed by substituting the values of the coefficients in

Table 4.4 into Equation 4.1. The cumulative logistic regression distribution function for the USDA population was,

$$\widehat{\boldsymbol{\pi}}(\boldsymbol{x_{i1}}, \boldsymbol{x_{i2}}, \boldsymbol{x_{i3}}) = \frac{e^{-23.06+0.12Stalks+3.37Height+6.71Diameter}}{1+e^{-23.06+0.12Stalks+3.37Height+6.71Diameter}}. \quad \text{Equation 4.9}$$

The logistic cumulative distribution function for the LSU AgCenter population was,

$$\widehat{\boldsymbol{\pi}}(\boldsymbol{x_{i1}}, \boldsymbol{x_{i2}}, \boldsymbol{x_{i3}}) = \frac{e^{-16.25+0.36Stalks+5.20Height}}{1+e^{-16.25+0.36Stalks+5.20Height}}. \quad \text{Equation 4.10}$$

The probability of selecting a seedling is calculated by substituting the values of the stalk numbers, stalk height and stalk diameter for that seedling in Equations 4.9 and 4.10. The probability is larger for higher values and smaller for lower values of the predictor variables.

The probability of selecting a seedling (*y*-axis) was plotted against the seedling cane yield (*x*-axis) to show the cumulative distribution patterns (Figure 4.1). The probability increased with increasing seedling cane yield for both populations. The LSU AgCenter population (Figure 4.1(b)) produced a slightly more variable pattern than the USDA population (Figure 4.1(a)). The patterns of the distributions reflect the variability within the populations.

4.3.3 Covariance Matrix of the Logistic Regression Coefficients

The covariance matrix is automatically generated by the SAS code (Appendix 2) from the training data set. The variances and covariance (Table 4.5) are substituted in Equation 4.6 to compute the variance of the log of odds that is in turn used to compute the standard errors of the log of odds. The standard errors of the log of odds are used for calculating the confidence limits of the log of odds. The confidence limits of the probability of selecting a seedling are calculated from the confidence limits of the log of odds using Equation 4.7. The covariance of two variables

can also be used to calculate their correlation coefficient. Using the covariance to compute the variance helps account for the correlation of predictor variables in calculating the probability confidence limits. In Table 4.5, the intersection of a coefficient row and its column represent the variance of that coefficient. For example, the variance of the intercept is 28.19 and that for the coefficient for stalk numbers is 0.0024. The intersection of the intercept and the coefficient of stalk numbers is their covariance, -0.128. In general, the diagonal represents the variances and the off-diagonals represent the covariance. Positive covariance mean positive correlation and negative covariance mean negative correlation between the two coefficients of the predictor variables. For both populations, the coefficients of stalk numbers, stalk number, stalk height and stalk diameter were negatively correlated with the intercept. The stalk number was positively correlated to stalk height and negatively correlated to stalk diameter.

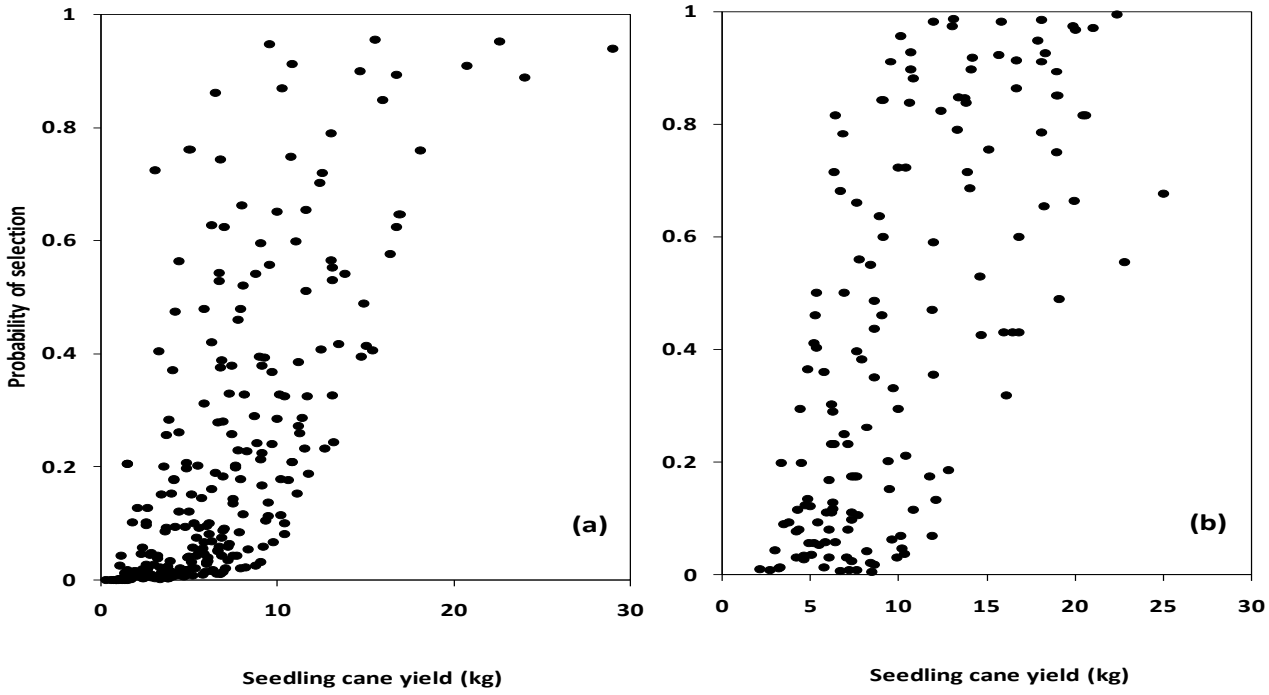

Figure 4.1: The cumulative logistic regression distribution patterns of the probability of selecting a seedling (*y*-axis) plotted against the seedling cane yield (*x*-axis) for the USDA (a) and LSU AgCenter (b) populations.

Table 4.5: The variances (diagonal) and covariances (off-diagonals) for the coefficients of the intercept, stalk number, stalk height and stalk diameter for the USDA and LSU AgCenter populations.

Population	Variable	Intercept	Stalk number	Stalk height	Stalk Diameter
USDA	Intercept	28.199	-0.128	-6.842	-5.801
	Stalk number	-0.128	0.003	0.009	0.036
	Stalk height	-6.842	0.009	3.221	-0.113
	Stalk diameter	-5.801	0.036	-0.113	2.776
LSU AgCenter	Intercept	22.204	-0.371	-7.974	†
	Stalk number	-0.371	0.011	0.106	
	Stalk height	-7.974	0.106	3.067	

†The stalk diameter for the LSU AgCenter population was not significant (P = 0.79, Table 4), and was excluded when calculating the covariance parameters after elimination during the variable selection process.

4.3.4 Output of Selection Probability and the Selection Probability Confidence Limits

A sample of the logistic regression analysis output is shown in Table 4.6 (USDA population) and Table 4.7 (LSU AgCenter population). The output includes variables used in building the logistic regression model (stalk number, stalk height and stalk diameter) and those variables not used to build the model such as cane yield. Other variables such as Brix, disease and insect damage ratings can be included in the output as aids to selection. The probability of selecting a seedling, P, is calculated using Equations 4.9 (USDA population) and Equation 4.10 (LSU AgCenter population) by plugging in the values of stalk numbers, stalk height and stalk diameter of a seedling. Only one probability, in this case, the probability of selecting a seedling, is computed. The probability of rejecting a seedling would be 1–P. The confidence limits of the probability of selecting a seedling are computed using Equations 4.4, 4.6, and 4.7. The seedling stalk number, stalk height and stalk diameter are plugged in the Equation 4.4 to compute the log of odds of

selecting a seedling. The seedling stalk number, stalk height and stalk diameter, and the variances and covariance from Table 4.5 are plugged in Equation 4.6 to compute the variance of the log of odds of selecting a seedling. The square root of the variance of the log of odds of selecting a seedling provides the standard error of the log of odds of selecting a seedling. The product of 1.96 (the 95 % confidence limit of the Z-distribution) by the standard error of the log of odds is added to (upper confidence limit) or subtracted from (lower confidence limit) the log of odds to provide the confidence limits of the log of odds of selecting a seedling. Equation 4.7 calculates the confidence limits of the probability of selecting a seedling by plugging in the confidents limits of the log of odds of selecting a seedling.

The probability and its confidence limits are used as selection aids. A selection threshold, that is user defined, determines the minimum probability of selecting a seedling. For example, using the threshold probability of 0.5, we selected seedlings 8, 9, 10, 11, 12, 13, 14, 15 from the USDA population (Table 4.6) and 1, 2, 3, 4, 5, 7, 10, 11, 14, 15 from the LSU AgCenter population (Table 4.7).

The probability confidence limits can be used to define a selection criterion where the seedlings are selected with a significantly larger probability than a given threshold. Using confidence limits, we can select seedlings with a probability that is significantly ($P < 0.05$) greater than the 0.5 threshold, that is the probability is greater than 0.5 and excludes 0.5 between the confidence limits of this probability. The seedlings 11, 14, 15 (USDA population) and 1, 2, 3, 4, 5, 7, 10, 14 (LSU AgCenter population) were selected at significantly ($P < 0.05$) larger threshold probability than 0.5. This criterion significantly reduced the number of seedlings selected compared to just selecting based on the probability threshold of 0.5.

Table 4.6: Sample output of the logistic regression analysis of the USDA population showing seedling number, stalk number, stalk height, stalk diameter, seedling cane yield, seedling selection probability and 95 % probability confidence limits.

Seedling	Stalk Numbers	Stalk Height (m)	Stalk Diameter (cm)	Cane Yield (kg)	Selection Probability	Lower Limit	Upper Limit
1	7	2.26	1.48	2.74	0.010	0.001	0.078
2	12	2.44	1.63	6.13	0.081	0.015	0.337
3	4	1.88	1.62	1.54	0.005	0.001	0.040
4	10	2.44	1.38	3.67	0.013	0.001	0.136
5	7	2.18	2.20	5.81	0.479	0.270	0.696
6	17	2.54	1.72	10.00	0.285	0.063	0.702
7	4	1.80	2.40	3.26	0.404	0.088	0.827
8	7	2.16	2.38	6.75	0.743	0.446	0.912
9	7	2.03	2.37	6.26	0.628	0.304	0.867
10	12	2.51	2.05	9.96	0.652	0.318	0.883
11	10	2.46	2.30	10.24	0.869	0.605	0.966
12	18	2.06	2.12	13.04	0.565	0.322	0.780
13	11	2.41	2.05	8.76	0.541	0.284	0.777
14	18	2.21	2.32	16.77	0.893	0.661	0.973
15	24	2.51	2.18	22.60	0.951	0.712	0.994

To further demonstrate the advantage of using confidence limits for seedling selection, the mean cane yield for the seedlings selected using the threshold probability of 0.5 and those selected using the probability significantly ($P < 0.05$) larger than the threshold of 0.5 was calculated. The mean seedling cane yield for the seedlings selected at the 0.5 threshold probability was 8.95 kg and that for significantly larger than 0.5 threshold probability was 16.54

kg for the USDA population. For the LSU AgCenter population, the mean was 7.60 kg (0.5 threshold probability) and 16.15 kg (significantly ($P < 0.05$) larger than the 0.5 threshold probability). Seedlings selected at significantly ($P < 0.05$) larger than 0.5 threshold probability produced 85 % (USDA population) and 112 % (LSU AgCenter population) more cane yield than those selected at the 0.5 threshold probability.

Table 4.7: Sample output of the logistic regression analysis of the LSU AgCenter population showing seedling number, stalk number, stalk height, stalk diameter, seedling cane yield, seedling selection probability and 95 % probability confidence limits.

Seedling	Stalk numbers	Stalk Height (m)	Stalk Diameter (cm)	Cane Yield (kg)	Selection Probability	Lower Limit	Upper Limit
1	16	2.50	2.35	18.36	0.926	0.688	0.986
2	17	2.40	2.22	16.71	0.914	0.687	0.981
3	15	2.50	1.85	10.70	0.897	0.637	0.977
4	12	2.70	2.65	19.01	0.893	0.551	0.983
5	22	2.30	1.76	13.06	0.974	0.814	0.997
6	5	2.35	2.75	7.40	0.098	0.021	0.359
7	15	2.30	2.30	15.16	0.754	0.522	0.896
8	8	2.40	2.50	10.03	0.294	0.109	0.585
9	5	2.70	2.19	5.38	0.402	0.094	0.813
10	14	2.60	2.44	18.11	0.911	0.629	0.984
11	12	2.40	1.93	8.97	0.637	0.379	0.835
12	9	2.50	1.70	5.40	0.501	0.217	0.784
13	8	2.30	1.72	4.54	0.198	0.070	0.450
14	25	2.20	1.99	18.09	0.985	0.846	0.999
15	11	2.40	1.96	8.42	0.550	0.301	0.777

4.3.5 Yield Trends of the Seedlings Identified Using Different Selection Strategies and Comparison to Visual Selection

Further investigation into the utilization of the probabilities and their confidence limits was done for the entire seedling populations. The output data from the logistic regression models were ranked using the probability of selecting a seedling. Within the populations, three groups were defined as rejected seedlings (seedlings to discard), average seedlings and elite seedlings (seedlings to advance). The rejected seedlings were defined as those seedlings with a probability of selection significantly ($P < 0.05$) lower than the 0.5 threshold, which is the probability lower than 0.5 and excluded 0.5 between their confidence limits. The average seedlings included 0.5 between their confidence limits. The probability of the elite seedling was significantly ($P < 0.05$) larger than the 0.5 threshold, which is larger than 0.5 and excluded 0.5 between their confidence limits. A data set was created with the group names as the class variables. This data set was subjected to analysis of variance and mean separation using Tukey's adjustment (Freund and Wilson, 2003). The elite seedlings produced significantly ($P < 0.05$) higher cane yield than the average and rejected seedlings (Table 4.8). The elite seedlings produced 39 % (USDA population) and 26 % (LSU AgCenter population) more yield than the average seedlings. The reject seedlings produced significantly ($P < 0.05$) less cane than the average seedlings. The elite seedlings produced significantly ($P < 0.05$) more stalks than the reject seedlings (USDA population), and the rejected and average seedlings (LSU AgCenter population). The elite seedlings were significantly ($P < 0.05$) taller for both populations and thicker (USDA population) than the rejected and average seedlings.

Table 4.8: The seedling means of selection probability and its confidence limits, stalk number, stalk height, stalk diameter, cane yield and cane yield expressed as percent of the elite for the reject, average and elite groups of the USDA and LSU AgCenter populations

Population	Group	Mean selection Probability (Confidence limits)	Number of stalks	Stalk height (m)	Stalk diameter (cm)	Cane yield (kg)	Cane yield % of elite
USDA	Reject	0.07 (0.03,0.18)	10.8a	2.01a	1.69a	4.84a	35
	Average	0.46 (0.20,0.73)	15.2b	2.28b	1.98b	10.00b	72
	Elite	0.88 (0.64,0.96)	14.5b	2.31b	2.33c	13.90c	100
LSU AgCenter	Reject	0.10 (0.03,0.29)	8.3a	2.04a	2.20a	6.73a	46
	Average	0.50 (0.27,0.73)	13.0b	2.23b	2.17a	11.48b	79
	Elite	0.89 (0.66,0.96)	17.2c	2.39c	2.05b	14.48c	100

The means for cane yield, stalk number, stalk height and stalk diameter of the selected and the rejected seedlings (visual appraisal method) were calculated (Table 4.9) and compared to means of seedlings selected using probability (Table 4.8). The elite seedlings selected by probability produced 48 % (USDA population) and 15 % (LSU AgCenter population) more cane yield than the seedlings selected by the visual appraisal method. For the USDA population, seedlings selected by visual appraisal produced 6 % less seedling cane yield than those seedlings classified in the average group by the logistic regression probability. The elite seedlings selected using probability produced more stalks that were taller and thicker than those seedlings selected using the visual method. Table 4.9 shows that gains can be made using visual selection but these gains could be increased significantly by using logistic regression probability as a decision support tool for seedling selection.

Table 4.9: The means of stalk number, stalk height, stalk diameter, cane yield and cane yield expressed as a percent of selected for seedlings selected and rejected by the visual method for the USDA and LSU AgCenter populations

Population	Selection Decision†	Stalks Numbers	Stalk Height (m)	Stalk Diameter (cm)	Cane yield (kg)	Cane yield %Selected
USDA	Rejected	12.17	2.07	1.73	6.02	64
	Selected	11.89	2.22	2.13	9.37	100
LSU AgCenter	Rejected	9.74	2.11	2.17	7.62	60
	Selected	15.58	2.28	2.12	12.62	100

†The decision to select a seedling was based on a consensus of the visual appraisal for cane yield using stalk number, stalk height and stalk diameter by two experienced sugarcane breeders.

4.3.6 Relationship of Seedling Stalk Numbers, Stalk Height and Stalk Diameter to Seedling Cane Yield Within the Populations

We investigated the trends in cane yield and the yield components across probabilities. The output data was ranked using the probability of seedling selection. Groups 1, 2, 3, and 4 were comprised of seedlings that produced a probability of selection significantly ($P < 0.05$) less than 0.1, 0.2, 0.3, and 0.5, respectively but without any overlapping of the probabilities of the groups for the USDA population. Group 5 comprised seedlings with probability similar to 0.5, that is, included 0.5 between their confidence limits, and group 6 comprised seedlings with probability significantly ($P < 0.05$) larger than 0.5. For the LSU AgCenter population, groups 1, 2, and 3 comprised seedlings with probability significantly ($P < 0.05$) less than 0.2, 0.3, and 0.5, respectively without any overlapping of groups. Group 4 comprised seedlings with probability similar to 0.5 and group 5 comprised seedlings with probability significantly ($P < 0.05$) larger than 0.5. The group mean probability rankings (*x*-axis) were plotted against the group trait means (*y*-axis) for stalk number, stalk height, stalk diameter and cane yield (Figures 4.2 and 4.3). The trends for the stalk number, stalk height, stalk diameter and cane yield increased with increasing mean probability rankings for the USDA population (Figure 4.2). The trend for cane yield was

more similar to that of stalk diameter. The trends for stalk number and stalk height were similar. Previously, we showed that stalk diameter was more associated with the decision to select seedlings with high cane yield (Table 4.4). For the LSU AgCenter population, the trends for the seedling cane yield, stalk number and stalk height increased with increasing mean probability rankings (Figure 4.3). The trend for the number of stalk was similar to that for cane yield. The stalk number was shown to be more associated with the decision to select seedlings with high cane yield than were stalk height and stalk diameter (Table 4.4). The stalk diameter produced a different trend to that for cane yield. Stalk diameter was not significantly associated with selection decision for seedling cane yield (Table 4.4). The trends in Figures 4.2 and 4.3 showed that the probability of seedling selection can also be used to study the relationship of traits to cane yield in breeding populations.

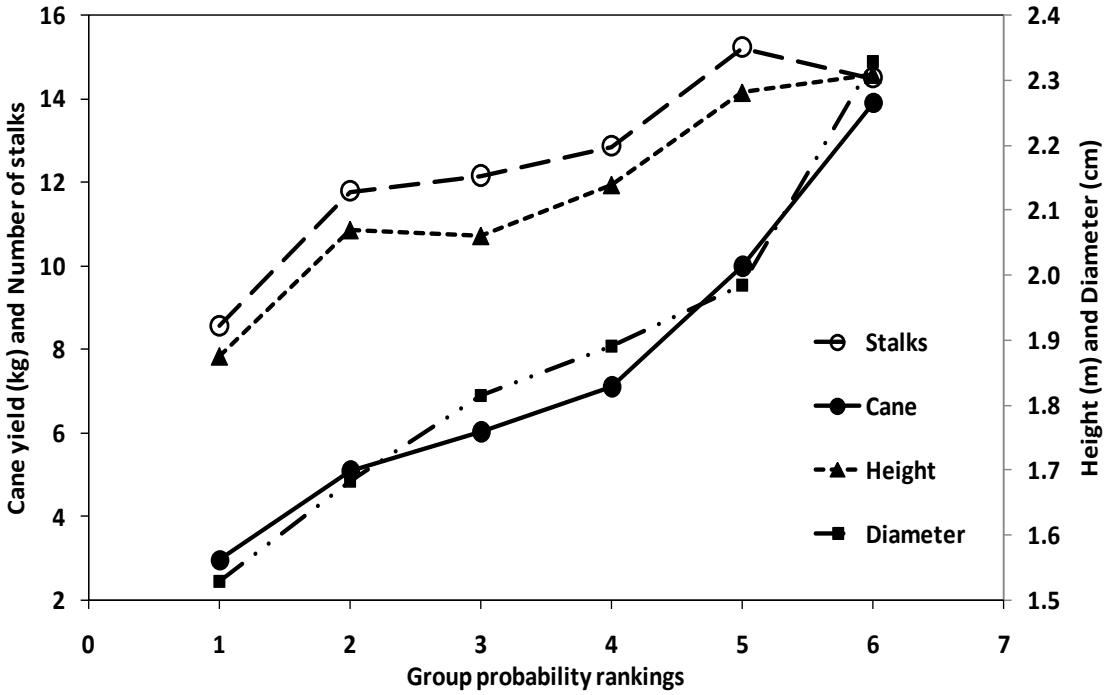

Figure 4.2: The trends of the means of seedling cane yield (kg), stalk number, stalk height (m) and stalk diameter (m) (*y*-axis) plotted against mean group probability rankings (*x*-axis) for the USDA population.

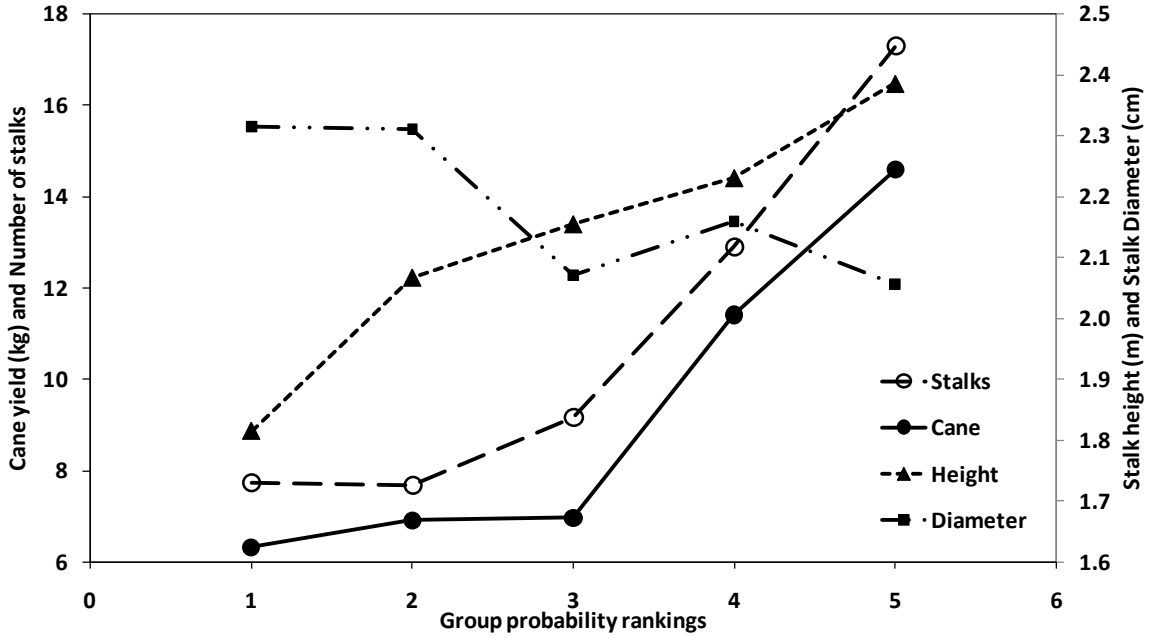

Figure 4.3: The trends of the means of seedling cane yield (kg), stalk number, stalk height (m) and stalk diameter (cm) (*y*-axis) plotted against probability rankings (*x*-axis) for the LSU AgCenter population.

4.4 Discussions

The logistic regression models selected higher cane yield seedlings than the visual appraisal method. The probability used in seedling selection by the logistic regression models is a function of the stalk number, stalk height and stalk diameter where higher seedling values produced higher seedling selection probability and were also associated with higher seedling cane yield than lower values. Seedlings selected using probability significantly ($P < 0.05$) greater than a selection threshold of 0.5 produced higher cane yield than those selected by visual selection. Logistic regression is a powerful statistical decision support tool for seedling selection that uses the yield components that are known to be strongly correlated with cane yield from path coefficient analysis studies (De Sousa-Vieira and Milligan, 2005; Kang *et al.*, 1983, 1989; Milligan *et al.*, 1990). It is easy to implement in SAS (Appendix 2) and other software.

The use of logistic regression models as a statistical decision support tool for sugarcane seedling selection can reduce the influence of genotype by environment interaction. At the seedling selection stage, the effects of genotype by environmental interaction are known to be very large (Kimbeng and Cox, 2003, Jackson and McRae, 2001). Genotype by environment interaction effects are particularly important for traits controlled by quantitative genes such as cane yield (Falconer and Mackay, 1996; Jackson and McRae, 1998). The logistic regression models would capture and account for the variability within a population when estimating the parameters used for building the models. The confidence intervals of the probability of selecting a seedling can also be used as indicators of the variability of the population. Larger confidence intervals indicate larger variability and the likely larger influence of genotype by environment interaction effects than lower confidence intervals. In such a situation, the selection can be adjusted according by relaxing the threshold probability of selection.

Using logistic regression models will provide sugarcane breeders with a decision support statistical tool for easily and quickly adjusting the number of seedlings to advance. Sugarcane breeders always have limited land to plant advanced seedlings necessitating the need to always adjust the number of seedlings selected to suit the available land. The breeder is interested in advancing the best yielding seedlings within the population after selection. With the logistic regression models, the probability and its confidence limits provide an easy and quick method to objectively adjust the numbers and also to help the breeder advance the best cane yielding seedlings. With the visual selection method, adjusting numbers would require the breeder to go back to the field and to reassess the seedlings to be added or removed, a daunting task. Because of the immense effort required to adjust the numbers and the limited time available to reassess the seedling to add or discard, the breeder would be tempted to just make a random selection for

seedling to be added or discarded. The output from the logistic regression models can also be programmed to include other variables such as disease and pest ratings that can be used as additional aids to selection.

The logistic regression models were also used to study the traits relationships within the populations providing insight into traits significantly influencing the decision to select. The logistic regression models can be used to identify the important traits within the breeding populations. This information on the important traits can also be used to identify parents with the appropriate traits that are required to improve the breeding populations, further improving the composition of desired traits in the breeding population. Over the years, breeding populations can shift because of the selection pressure imposed on the parents and their progenies. The logistic regression models can also be used to determine the magnitude of the shift in populations over the selection cycles. These studies can be done for individual crosses providing another mechanism for family evaluation. Those traits that do not respond to selection for cane yield such as diameter at the LSU AgCenter population can be assumed to be stable in a breeding population. Therefore emphasis during selection can be placed on responsive traits such as stalk number and stalk height. These evaluations can also be used to determine the progress achieved in improving traits in a breeding program.

The logistic regression model was capable of identifying the traits that were positively contributing to cane yield. For example, at the LSU AgCenter, the stalk number was the most significant trait contributing to the decision to select for cane yield. Previous studies in early selection stages by Zhou (1998, 2004b) also showed that stalk number were significantly and positively correlated with cane yield. The knowledge of the interrelationships of the traits being used as predictor variables for the selection decision is also important. In this study, the stalk

number was positively correlated to stalk height while the stalk height and stalk diameter were negatively correlated, a result also reported by Zhou (1998). This scenario requires that a balance be maintained between the stalk height and diameter during the planning of the development of the breeding populations and seedling selection.

The logistic regression models offer the breeder a method for directionally shifting the population to meet the breeding objectives. The objective of selection in plant breeding is to shift the population in the direction of interest to the breeder and the program objectives, for example high cane yield. With the logistic regression models, the shifting of the populations can be done easily and objectively using the training data. The training data is used to determine the parameters for building the logistic regression models that are in turn used to calculate the probability of selection. The training data can be defined and collected from a population that represents the desired outcome of the trait combination of the selected progenies. This training data will determine logistic regression parameters that when used to generate the probability to select will shift the selected population to resemble the training data. The populations for the training data can be derived from a fraction of the seedling population or from a set of families known to possess the desired trait combination expected in the selected progenies.

The potential constraint to the wide adoption of the logistic regression models in sugarcane seedling selection would be the required measurements of the stalk numbers, stalk height and stalk diameter. Because these measurements are labor intensive, most countries do not routinely measure the yield components at seedling selection. However, in countries where the labor costs are lower, such as in Zimbabwe, these measurements are routinely taken and used to adjust the numbers of seedlings to be advanced (Zhou, 2004a). We can also speculate that one of the reasons why these measurements are considered expensive to collect could be that there has

been no available statistical method for using them in seedling selection. From this study, the benefits associated with using logistic regression models for sugarcane seedling selection were significant compared to the visual selection method and could motivate sugarcane breeders to re-evaluate the cost versus the benefit to be gained by using this objective seedling selection decision support statistical tool.

While our study was primarily focused on the application of logistic regression models in seedling selection, there is potential for applications in other stages. The logistic regressions models could be particularly more useful in the first clonal stage that is generally planted to non-replicated and small plots. Application to replicated stages of the breeding programs could be of benefit where clones are planted to small plots (Jackson and McRae, 2001) and may aid in making selection decisions by reducing the influence of genotype by environment interaction effects.

4.5 Conclusions

The logistic regression models identified seedlings that produced higher cane yield than visual selection. The confounding effect of genotype by environment interaction is reduced by using logistic regression models for seedling selection. The logistic regression models provide for easy adjustment of the number of seedlings to advance by making use of the probability and their confidence intervals. Confidence intervals also help to account for the influence of genotype by environment interaction effects. The objective of selection in plant breeding is to shift the population towards desirable values of traits. Logistic regression models allow the plant breeders to achieve this shift by using the appropriate training data that would represents the desired outcome population after selection. Trends in traits and their influence on each other in breeding populations can also be investigated using the logistic regression models. These trends can also

be used to evaluate the progress made by the breeding program. While data for the predictor variables is labor intensive and costly to collect, the potential benefits and availability of logistic regression models that uses the data could motivate more breeding programs to re-evaluate the cost versus the benefit. The logistic regression models can be applied in other stages and would be particularly useful in the non-replicated stages where clones are planted to small plots.

4.6 References

Agresti, A. (2007). An Introduction to Categorical Data Analysis. Second Edition. John Wiley and Sons, USA.

Allison, P.D. (2003). Logistic Regression Using the SAS System: Theory and Applications. SAS Institute Inc., Cary, NC, USA.

Bowie, D.K. (2006). Using multivariate logistic regression analysis to predict black male persistence at a predominantly white institution: an approach investigating the relationship between student engagement and persistence. PhD Dissertation. Louisiana State University, Baton Rouge, Louisiana, USA.

Breaux, R.D. and Miller, J.D. (1987). Seed handling, germination and seedling propagation. In DJ Heinz (editor). Sugarcane Improvement Through Breeding. Elsevier.: 385 – 407.

Casella, G. and Berger, R.L. (2003). Statistical Inference. 2nd Edition. Thomson publications.

Chang, Y.S. and Milligan, S.B. (1992). Estimating the potential of sugarcane families to produce elite genotypes using univariate cross prediction methods. *Theoretical and Applied Genetics* 84: 662 – 671.

Cohen, J.E. (2006). Public Opinion in State Politics. Stanford University Press. California, USA.

Cox, M.C. and Hogarth, D.M. (1993). Progress and changes in the South Queensland Variety Development Program. *Proceedings of the International Society of Sugar Cane Technologist* 15: 251 – 255.

Cox, M.C. and Stringer, J.K. (1998). Efficacy of early generation selection in a sugarcane improvement program. *Proceedings of the Australian Society of Sugarcane Technologist* 20: 148 – 153.

De Sousa-Vieira, O. and Milligan, S.B. (1999). Intra-row spacing and family x environment effects on sugarcane family evaluation. *Crop Science* 39: 358 – 364.

De Sousa-Vieira, O. and Milligan, S.B. (2005). Interrelationships of cane yield components and their utility in sugarcane family selection. *Interciencia* 30(2): 93 – 96.

Falconer, D.S. and Mackay, T.F.C. (1996). Introduction to Quantitative Genetics. Fourth Edition. Longman Group Ltd, UK.

Freund, R.J. and Wilson, W.J. (2003). Statistical Methods. Academic Press. New York.

Gravois, K.A., Milligan, S.B. and Martin, F.A. (1991). Indirect selection for increased sucrose yield in early sugarcane testing stages. *Field Crops Research* 26: 67 – 73.

Hogarth, D.M. and Berding, N. (2006). Breeding for a better industry: Conventional breeding. *Sugarcane International* 24 *Number* 2: 26 – 31.

Hogarth, D.M. and Mullins, R.T. (1989). Changes in the BSES plant improvement program. *Proceedings of the International Society of Sugar Cane Technologist* 20: 956 – 961.

Hosmer, D.W.Jr. and Lemeshow, S. (1989). Applied Logistic Regression. John Wiley and Sons Inc. New York. USA.

Jackson, P.A. and McRae, T.A. (1998). Gains from selection of broadly adapted and specifically adapted sugarcane families. *Field Crops Research* 59: 151 – 162.

Jackson, P.A. and McRae, T.A. (2001). Selection of sugarcane clones in small plots: Effects of plot size and selection criteria. *Crop Science* 41: 315 – 322.

Kang, M.S., Miller, J.D. and Tai, P.Y.P. (1983). Genetic and phenotypic path analysis and heritability in sugarcane. *Crop Science* 23: 643 – 647.

Kang, M.S., Sosa, O. and Miller, J.D. (1989). Path analysis for percent fiber, and cane and sugar yield in sugarcane. *Crop Science* 29: 1481 – 1483.

Kimbeng, C.A., Froyland, D., Appo, D., Corcoran, A. and Hetherington, M. (2001a). An appraisal of early generation selection in the central Queensland sugarcane improvement program. *Proceedings of the Australian Society of Sugar Cane Technologists* 23: 129 – 135.

Kimbeng, C.A., McRae, T.A. and Cox, M.C. (2001b). Optimizing early generation selection in sugarcane breeding. *Proceedings of the International Society of Sugarcane Technologists* 24: 488 – 493.

Kimbeng, C.A. and Cox, M.C. (2003). Early generation selection of sugarcane families and clones in Australia: A Review. *Journal of the American Society of Sugarcane Technologists* 23: 20 – 39.

Le, C.T. (1998). Applied Categorical Data Analysis. John Wiley and Sons Inc. New York. USA.

Miller, J.D. and James, N.I. (1974). The influence of stalk density on cane yield. *Proceedings of the International Society of Sugarcane Technologists* 15: 177 – 184.

Milligan, S.B., Gravois, K.A., Bischoff, K.P. and Martin, F.A. (1990). Crop effects on genetic relationships among sugarcane traits. Crop Science 30: 927 – 931.

SAS Institute (2007). The SAS System for Windows Version 9.1.3. SAS Institute, Cary, North Carolina, USA.

Singh, R.K., Singh, S.P. and Singh, S.B. (2005). Correlation and path analysis in sugarcane ratoon. Sugar Tech 7(4): 176 – 178.

Zhou, M. (1998). Trends in yield, sucrose, stalk population and smut tolerance over 20 years of sugarcane selection in Zimbabwe. *Proceedings of the South African Sugarcane Technologists Association* 72: 47-50.

Zhou, M. (2004a). Strategies for variety selection in the breeding program at the Zimbabwe Sugar Association Experiment Station. *Proceedings of the South African Sugar Technologists Association* 78: 125 - 131.

Zhou, M. (2004b). Stalk population control of yield, quality and agronomic traits of sugarcane population in early selection stages. *Sugar Cane International*: *Volume* 22 *number* 5: 14 – 20.

CHAPTER 5: MULTIVARIATE REPEATED MEASURES ANALYSIS OF DATA FROM ADVANCED VARIETY TRIALS USING THE MIXED PROCEDURES OF SAS

5.1 Introduction

In sugarcane advanced variety trials, the data used to evaluate the differences between genotypes or entries are collected for several variables (yield, quality and agronomic traits) from individual plots every year for several years. The data collected in each crop every year for the several years are also used to determine the ratooning ability of the experimental genotypes (Berding *et al*., 2004).

Ratooning refers to the harvesting of several crops for several years from the same planting and it is important in sugarcane production economics. The sugarcane crops are harvested sequentially from the plant, first, and second ratoon, in successive years resulting in crop and year confounding to form crop-years (Kang *et al*., 1987). Planting varieties with high cane yield and high ratooning ability increases the profitability in sugarcane production (Berding *et al*., 2004; Clowes and Breakwell, 1998; Ellis and Merry, 2004; Salassi and Giesler, 1995). In sugarcane production, it is cheaper to maintain the ratoon crops than to plant a new crop every year. Planting sugarcane crops requires large quantities of bulky vegetative planting material that is expensive to transport from the source field to the field to be planted. In irrigated sugarcane production systems, the expensive land preparation, irrigation system rehabilitations, and planting operations add to the cost of establishing a crop. Growing more crops from each planting allows growers to recover these costs. Therefore, it is logical that ratooning ability is an important trait in sugarcane breeding and therefore varieties are evaluated for ratooning ability in advanced variety trials.

Data from multiple variables measured in each plot resemble a multivariate structure (Johnson and Wichern, 2002). Values of the multiple variables measured from each plot may not be independent because they are influenced by the same factors existing in that plot. For example, a plot that produced high cane yield is also likely to produce taller and thicker stalks. The result is that the multiple variables measured from the same plot could be correlated. The data of each variable measured from each plot over several crop-years resemble repeated measures (Littell *et al*., 2002, 2005). The measurements from one crop-year are likely not to be independent from measurements from other crop-years because the measurements come from sequential crop-years. The crop-years cannot be randomized to the plots (as would be done in an ideal split plot design). Additionally, a plot that produced high cane yield in crop-year 1 is also likely to produce high cane yield in crop-year 2 and subsequent crop-years. The result is that, for example cane yield from crop-year 1 could be correlated to cane yield measured in crop-year 2. Therefore the analysis of plant breeding data may need to account for the within plot correlation of the multiple variables (multivariate structure) and the correlation of the value of variables measured across crop-years (repeated measures).

Currently, the univariate analysis method that assumes a split-plot in time experimental design is used to analyze data from the advanced variety trials. The univariate method assumes independence between variables measured from the same plot and also assumes independence between values of a variable measured in successive crop-years (Freund and Wilson, 2003). The assumption of independence between data from multiple variables measured from the same plot and between data measured from the same plot across crop-years may not always be valid. The values of the multiple variables are influenced by the same factors that exist in the plot from which they are measured. These multiple variables are likely to be correlated. If these multiple

variables are significantly correlated, then there could be a violation of the assumption of independence. One of the consequences of the violation of the assumption of independence would be the underestimation or overestimation of the experimental errors. The underestimation or overestimation of experimental errors could increase Type I or Type II errors, respectively, leading to inaccurate statistical tests and incorrect interpretations. The underestimation or overestimation of experimental errors is caused by the exclusion of the covariance between variables as well as the covariance between crop-years in the computation of the variances. The covariance helps account for the correlation between the multiple variables and the correlation between crop-years. The ideal analysis should combine multivariate and repeated measures, to create a multivariate repeated measures analysis. The multivariate repeated measures analysis would account for the correlation between the multiple variables as well as the correlations between the sequential crop-years in a single analysis. We hypothesize that combining the multivariate and the repeated measures in one analysis will increase precision in the analysis of sugarcane advanced variety trials breeding data and therefore produce accurate tests and correct interpretation of the data.

The objectives of this study were to introduce and demonstrate the use of the multivariate repeated measures analysis method for sugarcane breeding advanced variety trials data using the linear mixed models of the SAS procedures (SAS Institute, 2007). Specifically we determined multivariate effects, the appropriate covariance structure for crop-years, and compared the univariate and multivariate repeated measures analysis methods for yield (cane and stalk dry matter yield), quality (sucrose % cane and Fiber % cane) and agronomic (stalk height and stalk diameter) traits.

5.1.1 Multivariate Repeated Measures Analysis Using the Mixed Procedures of SAS

The mixed model procedure of SAS can perform both the multivariate and the repeated measures analysis. The linear mixed model equation is,

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{u} + \boldsymbol{\varepsilon}, \qquad \text{Equation 5.1}$$

where $\boldsymbol{Y}$ is the column vector of the response variables, $\mathbf{X}$ is the fixed effects design matrix, $\boldsymbol{\beta}$ is the column vector of the fixed effects parameters, $\mathbf{Z}$ is the random effects design matrix, $\boldsymbol{u}$ is the column vector of the random effects parameters and $\boldsymbol{\varepsilon}$ is the column vector of the residual errors (Littell *et al*., 2005). The linear mixed model (Equation 5.1) combines the analysis of fixed ($\mathbf{X}\boldsymbol{\beta}$) and random ($\mathbf{Z}\boldsymbol{u}$) effects as well as modelling covariance parameters ($\boldsymbol{\varepsilon}$). The ability of the mixed models to perform multivariate and repeated measures analysis, and to model covariance parameters was utilized in this study to perform the multivariate repeated measures analysis.

The discovery of the direct (Kronecker) product structures allows the implementation of the multivariate repeated measures analyses (Galecki, 1994). The unstructured (UN) structure (representing the multivariate component) and the repeated measures covariance structures are merged by the direct product. In SAS, the products are coded TYPE = UN@AR(1), modeling the first order auto-regressive, TYPE = UN@CS, modeling the compound symmetry, and TYPE = UN@UN modeling the unstructured structure in the repeated measures. The direct product of the two matrices has rows equal to the product of rows for, say, UN and AR(1) and columns equal to the product of the columns for UN and AR(1). The UN@UN models unequal covariance, UN@CS models equal covariance, and UN@AR(1) models covariance decay over time.

5.1.2 Profile Analysis

Profile analysis provides detailed comparisons of the treatments involving multivariate data by incorporating tests that use linear combinations of the response variables (Moser, 2005). While the multivariate repeated measures analysis identifies the effects that are significant, it is also important to find out how the treatments vary over time that is across crop-years. The tests are done for parallel, coincident, and level profiles (Morisson, 1976; Srivastava and Carter, 1983).

The parallel profile test asks if the difference between treatments is the same across the times of measurement. The hypothesis being tested is,

$$\mathbf{H_0}: \boldsymbol{\mu_{1j} - \mu_{1(j-1)} = \mu_{2j} - \mu_{2(j-1)} \quad for\ j = 2, 3, \ldots, p,} \qquad \text{Equation 5.2}$$

where $\boldsymbol{\mu_1}$ is the mean of treatment 1, $\boldsymbol{\mu_2}$ is the mean of treatment 2, and *j* indexes the time intervals between the measurements being compared. When there are more than 2 treatments, say 5 treatments, treatments 1, 2, 3 and 4 will be tested against treatment 5 for the overall test.

The coincident profile test determines if the profiles are on top of each other. The hypothesis being tested is,

$$\mathrm{H_0}: \sum_{j=1}^{p} \mu_{1j} = \sum_{j=1}^{p} \mu_{2j}\,. \qquad \text{Equation 5.3}$$

If the means of the treatments at each time are the same, then the profiles are coincident. If the sums of the treatment means are equal, then the profiles are also coincident.

Level profiles should have the same mean for each time measurement for each treatment. The hypothesis being tested is,

$$\mathbf{H_0}: \boldsymbol{\mu_{1j} - \mu_{1(j-1)} = \mu_{2j} - \mu_{2(j-1)} = 0 \quad for\ j = 2, 3, \ldots, p.} \qquad \text{Equation 5.4}$$

Therefore, when the profiles are level, the slope of the profiles will be zero.

5.1.3 Covariance Structure Selection

The objective of covariance structure selection is to determine the most parsimonious structure (Moser, 2005). Information criteria are used to select and measure the relative fit of two or more competing models. The Akaike Information Criterion (AIC) (Akaike, 1974) and the Bayesian Information Criterion (BIC) (Schwarz, 1978) are used to compare the competing models. The AIC is calculated using

$$\text{AIC} = -2 \log (L) + 2k, \qquad \text{Equation 5.5}$$

where L is the maximum likelihood function of the model and k is the number of effective covariance parameters, that is, those that enter the optimization process, are not held fixed by the user, and are not zero.

The BIC method was developed using the Bayesian approach and is not sensitive to prior distributions when the sample size is large. The BIC is calculated as

$$\text{BIC} = -2 \log (L) + k \log(n), \qquad \text{Equation 5.6}$$

where n is the sample size. Studies by Guerin and Stroup (2000) found that larger values of BIC were associated with larger Type II errors.

5.2 Materials and Methods

5.2.1 Locations, Experimental Design, and Crop Management

Data were collected from the plant breeding advanced variety trials grown at the Mkwasine and Triangle locations in the South East Lowveld of Zimbabwe. The plots were arranged as a

randomized block design, blocking across irrigation furrows. The Mkwasine location had four blocks while the Triangle location had five. Each block was divided into 16 plots and each plot was planted to one of the 16 genotypes. The plots were made up of 6 rows that were 12m long and spaced 1.5m apart. The trials were planted on April 25, 1995 (Mkwasine) and April 26, 1995 (Triangle) and harvested at 12 months crop age every year for eight crop-years. At both locations, water was applied using furrow irrigation. Planting, fertilizer application, irrigation, and weed, disease, and insect control were done according to standard recommendations for the commercial crop (Clowes and Breakwell, 1998).

5.2.2 Data Collection

At harvest, all the sugarcane in the plots was burnt to remove the dry leaves. All the millable stalks in each plot were hand cut, hand trashed to remove the green leaves and hand topped at the natural breaking point. The millable stalks were weighed using a digital scale mounted on a tractor operated hydraulic boom. The weights per plot were divided by the plot area to calculate cane yield (Mg/ha). Twenty-four millable stalks were randomly picked from each plot, and bundled. The length of the bundle from bottom to the top provided the stalk height of each plot. The stalk diameter of each of the 24 stalks was measured at the center of the stalk using a caliper without reference to the bud and the average stalk diameter of the 24 stalks provided the values for each plot used in this study. After measuring the stalk diameter, the 24 stalks were divided into three groups of eight stalks each. From the first group, the bottom one-third of the stalk was cut. From the second group, the middle one-third was cut and from the third group, the top one-third was cut. The bottom, middle and top portions of the stalks were bundled together to form one sub-sample per plot. Each sub-sample was shredded to simulate milling. Two sub-sub-samples were collected from each shredded sub-sample. One sub-sub-sample was analyzed for

sucrose content which was expressed as estimable recoverable crystal (ERC % cane) using an empirical equation determined from mill sugar recovery data derived from the previous season. The other sub-sub-sample was dried for 24 hours in an oven at a constant temperature of 100 ^{o}C and used to determine the Fiber % cane and moisture content. The moisture content (MC) was then used to estimate the stalk dry matter (SDM) from cane yield (Equation 5.7),

$$SDM = \text{Cane yield} * (100 - MC) \div 100. \qquad \text{Equation 5.7}$$

Three groups of data emerged, that is, yield, quality, and agronomic traits data. Yield traits (cane and SDM) were measured at the plot level, the agronomic traits (stalk height and stalk diameter) were measured from the 24 stalks sampled from each plot, and the quality traits (ERC % cane and Fiber % cane) were measured from the sub-sub-sample derived from the shredded sub-sample of a third of the 24 stalk sample. As a result, the correlation within yield, quality, and agronomic traits was likely to be larger than the correlation between the trait groups. Therefore each trait group was analyzed separately.

5.2.3 Data Arrangement and Analysis Using the Multivariate Mixed Model of SAS

The multivariate repeated measures analysis was done using the mixed procedure of SAS. A response variable (Y) was created with all the response variables stacked. A class variable, RV, was created identifying each variable by stacking the corresponding variable names. The data was arranged as shown in Table 5.1. In Table 5.1, using yield data as an example, RV = 1 referenced cane yield and RV = 2 referenced SDM yield. Location = 1 referenced the Triangle location and location = 2 referenced the Mkwasine location. The effects were nested in RV and together with the NOINT option of SAS produced the multivariate analysis and testing of the effects (Appendices 3, 4, 5). The NOINT option allows each variable in RV (for example, cane

yield and SDM yield) to be treated as unique variables. With the NOINT option, the levels of RV are not compared. The comparisons are done within the levels of RV and the effects within the RV are added up for both the variables in the multivariate structure to provide the multivariate tests of the effects. The SAS codes used to implement the multivariate repeated measures analysis using the UN@UN, UN@CS and UN@AR(1) covariance structures are shown in Appendices 3, 4 and 5, respectively. Also shown in Appendix 4 is the SAS option statement that was used for performing the tests for the difference between the experimental genotypes and the control genotype using Dunnett's test using the UN@CS covariance structure.

5.2.4 Multivariate Repeated Measures Linear Mixed Model

The multivariate repeated measures linear mixed model for yield traits with two response variables (for example cane yield and SDM yield), two locations (1, 2), *r* replications per location planted to *g* genotypes and harvested for *c* crop-years, is

$$Y_{ijkmn} = \pi_i + \alpha(\pi)_{j(i)} + \rho\big(\alpha(\pi)\big)_{k(j(i))} + \gamma(\pi)_{m(i)} + \omega(\pi)_{n(i)} + \alpha\gamma(\pi)_{jm(i)} + \alpha\omega(\pi)_{jn(i)} + \gamma\omega(\pi)_{mn(i)} + \alpha\gamma\omega(\pi)_{jmn(i)} + \varepsilon_{ijkmn}, \qquad \text{Equation 5.8}$$

where Y_{ijkmn} is the response for the *i*th variable (i = 1, 2), *j*th location (j = 1, 2), *k*th replication within *j*th location (k = 1, 2, …, r), *m*th genotype (m = 1, 2, …, g) by replication (plot), and *n*th crop-year (n = 1, 2, …, c). The model effects are as follows: π_i is the effect of the *i*th response variable (RV), $\alpha(\pi)_{j(i)}$ is the effect of the *j*th location nested within the *i*th variable, $\rho(\alpha(\pi))_{k(j(i))}$ is the random effect of the *k*th replication nested within the *j*th location that is in turn nested within the *i*th variable, $\gamma(\pi)_{m(i)}$ is the effect of the *m*th genotype nested within the *i*th variable,

Table 5.1: Data arrangement for the response class variable (RV), location, replication, genotype, crop-year, and measured values (Y) for the multivariate repeated measures analysis using the linear mixed model procedure of SAS

RV	Location	Replication	Genotype	Crop-year	Y
1	1	1	1	1	y_{11111}
1	1	1	1	2	y_{11112}
1	1	1	1	3	y_{11113}
.	.	.	.	.	.
.	.	.	.	.	.
1	1	1	1	*c*	y_{1111c}
1	1	1	2	1	y_{11121}
1	1	1	2	2	y_{11122}
1	1	1	2	3	y_{11123}
.	.	.	.	.	.
.	.	.	.	.	.
1	1	1	*g*	*c*	y_{111gc}
1	1	2	1	1	y_{11211}
1	1	2	1	2	y_{11212}
1	1	2	1	3	y_{11213}
.	.	.	.	.	.
.	.	.	.	.	.
1	1	*r*	*g*	*c*	y_{11rgc}
1	2	1	1	1	y_{12111}
1	2	1	1	2	y_{12112}
1	2	1	1	3	y_{12113}
.	.	.	.	.	.
.	.	.	.	.	.
1	2	*r*	*g*	*c*	y_{12rgc}
2	1	1	1	1	y_{21111}
2	1	1	1	2	y_{21112}
2	1	1	1	3	y_{21113}
.	.	.	.	.	.
.	.	.	.	.	.
2	1	*r*	*g*	*c*	y_{21rgc}
2	2	1	1	1	y_{22111}
2	2	1	1	2	y_{22112}
2	2	1	1	3	y_{22113}
.	.	.	.	.	.
.	.	.	.	.	.
2	2	*r*	*g*	*c*	y_{22rgc}

$\omega(\pi)_{n(i)}$ is the effect of the *n*th crop-year nested within the *i*th variable, $\alpha\gamma(\pi)_{jm(i)}$ is the interaction effect of the *j*th location and the *m*th genotype nested within the *i*th variable, $\alpha\omega(\pi)_{jn(i)}$ is the interaction effect of the *j*th location and the *n*th crop-year nested within the *i*th variable, $\gamma\omega(\pi)_{mn(i)}$ is the interaction effect of the *m*th genotype and the *n*th crop-year nested within the *i*th variable, $\alpha\gamma\omega(\pi)_{jmn(i)}$ is the interaction effect of the *j*th location by the *m*th genotype by the *n*th crop-year nested within the *i*th variable, and ε_{ijkmn} is the residual error. The above linear mixed model (Equation 5.8) was used for the quality and agronomic traits. All the effects in Equation 5.8 are nested within the response variable (RV) to create the multivariate analysis and multivariate testing of the effects.

5.2.5 Comparison of the Efficiency of the Univariate and the Multivariate Repeated Measures Analysis

The multivariate repeated measures and univariate analysis were compared for their ability to account for the variability in the data, which is their model fitness. The model fit of the multivariate repeated measures and univariate analysis were compared using the fit statistics and the likelihood ratio tests. The multivariate repeated measures analysis was compared to the univariate analysis to evaluate the efficiency in the discriminating ability of the statistical methods on the experimental genotypes mean for the yield, quality and agronomic traits. The multivariate repeated measures and univariate analysis methods were compared for their discriminating ability of the difference in trait values between the experimental genotypes and the control. The difference between the experimental genotypes and the control is routinely used by plant breeders to identify superior genotypes in variety trials. The experimental genotypes were compared to genotype 16, the control genotype, using Dunnett's test for both the multivariate repeated measures and univariate analysis methods (Appendix 4). The P-value,

which is the probability of obtaining a larger value of the difference between the experimental genotype and genotype 16, was used to compare the discriminating ability between the multivariate repeated measures and univariate analysis methods. The statistical method that produced greater differences among the genotypes for the difference between the experimental genotypes and the control was considered to be more discriminating.

5.3 Results

The objective of the sugarcane breeding advanced variety trials is to evaluate the performance of the experimental genotypes for yield potential, quality, agronomic traits and ratooning ability and to determine the potential of these genotypes for release as commercial cultivars as well as their potential as parents for use in future crosses. Genotypes with potential for commercial varieties must produce similar or greater sugar yield and greater ratooning ability than the current cultivars in addition to excelling in other important traits such as disease and insect pest resistance. The sugarcane breeder is interested in evaluating genotype yield across locations (genotype by environment interaction) and ratooning ability (genotype by crop-year interaction). The genotype within RV, location by genotype within RV, genotype by crop-year within RV and location by genotype by crop-year within RV effects are used to evaluate genotype yield potential, determine the influence of locations, crop-years, and location by crop-year interactions, respectively, on the genotype yield potential. The location effects test environmental adaptation to factors such as soil type, changes in temperature and rainfall across locations while the crop-year effects test the ratooning ability of the genotypes, which is the fluctuation in yield across crop-years.

5.3.1 Multivariate Repeated Measures Analysis of Yield, Quality and Agronomic Traits Data

The multivariate repeated measures mixed model analysis for yield traits (cane yield, SDM yield) produced highly significant ($P < 0.01$) P-values for all the effects for the UN@CS and UN@AR(1) covariance structures (Table 5.2). The UN@UN covariance structure failed to converge. The multivariate repeated measures analysis for the quality traits (ERC % cane and Fiber % cane) and the agronomic traits (stalk height and stalk diameter) produced highly significant ($P < 0.01$) P-values for all the effects and for all the covariance structures (Tables 5.3 and 5.4).

Table 5.2: The numerator and denominator degrees of freedom and the probability of obtaining a larger Multivariate F-value (Multivariate P-values) for the multivariate effects for the yield traits (Cane (t/ha) and SDM (t/ha)) derived from the UN@UN, UN@CS and UN@AR(1) covariance structures.

Effect	Degrees of Freedom		Multivariate P-values		
	Numerator	Denominator	UN@UN†	UN@CS	UN@AR(1)
RV	2	14	Did not converge	<0.0001	<0.0001
Location(RV)	2	14		<0.0001	0.0002
Genotype(RV)	30	1778		<0.0001	<0.0001
Crop-Year(RV)	14	1778		<0.0001	<0.0001
Genotype*Location(RV)	30	1778		<0.0001	<0.0001
Crop-Year*Location(RV)	14	1778		<0.0001	<0.0001
Genotype*Crop-Year(RV)	210	1778		<0.0001	<0.0001
Genotype*Location*Crop-Year(RV)	210	1778		<0.0001	<0.0001

†The model did not converge because it was unable to make hessian positive definite matrix

Table 5.3: The numerator and denominator degrees of freedom and the probability of obtaining a larger Multivariate F-value (Multivariate P-values) for the multivariate effects for quality traits (ERC % cane and Fiber % cane) derived from the UN@UN, UN@CS and UN@AR(1) covariance structures.

Effect	Degrees of Freedom		Multivariate P-values		
	Numerator	Denominator	UN@UN	UN@CS	UN@AR(1)
RV	2	14	<0.0001	<0.0001	<0.0001
Location(RV)	2	14	0.0004	0.0007	0.0006
Genotype(RV)	30	1778	<0.0001	<0.0001	<0.0001
Crop-Year(RV)	14	1778	<0.0001	<0.0001	<0.0001
Genotype*Location(RV)	30	1778	0.0223	0.0170	0.0072
Crop-Year*Location(RV)	14	1778	<0.0001	<0.0001	<0.0001
Genotype*Crop-Year(RV)	210	1778	<0.0001	<0.0001	<0.0001
Genotype*Location*Crop-Year(RV)	210	1778	<0.0001	0.0002	0.0002

Table 5.4: The numerator and denominator degrees of freedom and the probability of obtaining a larger Multivariate F-value (Multivariate P-values) for the multivariate effects for the agronomic traits (stalk height and stalk diameter) derived from the UN@UN, UN@CS and UN@AR(1) covariance structures.

Effect	Degrees of Freedom		Multivariate P-values		
	Numerator	Denominator	UN@UN	UN@CS	UN@AR(1)
RV	2	14	<0.0001	<0.0001	<0.0001
Location(RV)	2	14	0.0043	0.0014	0.0019
Genotype(RV)	30	1778	<0.0001	<0.0001	<0.0001
CropYear(RV)	14	1778	<0.0001	<0.0001	<0.0001
Genotype*Location(RV)	30	1778	<0.0001	<0.0001	<0.0001
CropYear*Location(RV)	14	1778	<0.0001	<0.0001	<0.0001
Genotype*CropYear(RV)	210	1778	<0.0001	<0.0001	<0.0001
Genotype*Location*Crop-Year(RV)	210	1778	<0.0001	<0.0001	0.0006

The interpretation of the multivariate effects must recognize that the effects are computed within each of the variables making up the multivariate component. The effects computed within

each variable are then added up to produce the values of the multivariate F-statistic that are tested. The multivariate P-values (Tables 5.2, 5.3, 5.4) refer to the probability of obtaining a larger value of the multivariate F-statistic for the multivariate effects. The multivariate F-statistic of each multivariate effect follows the F-distribution with the numerator and denominator degrees of freedom shown in Tables 5.2, 5.3, and 5.4. The significant multivariate F-statistic would mean that at least one of the variables making up the multivariate structure produced significant effects.

The significant genotype within RV effects for the yield traits, for example, meant that the genotype effects were significantly different for cane yield or SDM yield or both. This test is also equivalent to the multivariate coincident profiles test for all the 16 genotypes. The significant multivariate non-coincident profiles for the yield traits suggest that at least one pair of the 16 genotypes was significantly different for cane yield or SDM yield or both. The significant location by genotype within RV effects for yield traits suggests that the location by genotype interaction effects were significantly different for cane yield or SDM yield or both. The significant crop-year within RV effects for yield traits meant that the crop-year effects were significantly different for cane yield or SDM yield or both. The crop-year within RV effects is equivalent to the multivariate level profiles test for all the 8 crop-years. Significant multivariate non-level profiles for yield traits suggest that at least one pair of the 8 crop-years was significantly different for cane yield or SDM yield or both. Significant genotype by crop-year within RV effects for the yield traits meant that the genotype by crop-year interaction effects were significantly different for cane yield or SDM yield or both. This test is equivalent to the multivariate parallel profiles test for all the 16 genotypes across all the 8 crop-years. The significant multivariate non-parallel profiles for the yield traits suggests that at least one pair of

the 16 genotypes was significantly different in at least one pair of the 8 crop-years for cane yield or SDM yield or both. Significant location by genotype by crop-year within RV effects for the yield traits meant that the location by genotype by crop-year interaction effects were significantly different for cane or SDM yield or both. The interpretation for the quality traits (ERC % cane and Fiber % cane) (Table 5.3) and agronomic traits (stalk height and stalk diameter) (Table 5.4) followed the same pattern to that for yield traits.

5.3.2 Covariance Structure Selection

The covariance structure UN@CS was selected as the most appropriate because it used fewer parameters than the UN@UN covariance structure (simplicity) and produced lower AIC and BIC than the UN@AR(1) covariance structure (Table 5.5). The covariance structure UN@AR(1) produced higher BIC values indicating larger Type II errors particularly for the yield and agronomic traits than UN@CS. The UN@CS covariance structure was used in performing the Dunnett's tests comparing genotypes to the control. The probability values obtained from Dunnett's test were used to evaluate the efficiency of the univariate and multivariate repeated measures analysis in discriminating between the experimental genotypes.

Table 5.5 The number of fitted covariance parameters, the Akaike information criterion (AIC), and the Bayesian information criterion (BIC) derived from the multivariate repeated measures analysis for the yield, quality and agronomic traits using the UN@UN, UN@CS and UN@AR(1) covariance structures.

Covariance Structure	Number of Parameters	Yield traits		Quality traits		Agronomic traits	
		AIC	BIC	AIC	BIC	AIC	BIC
UN@UN	42	-	-	4967.7	4975.8	-363.5	-355.4
UN@CS	7	10403.0	10404.4	5047.4	5048.7	-321.6	-320.3
UN@AR(1)	7	10423.5	10424.9	5047.6	5048.9	-287.7	-286.4

5.3.3 Comparison of the Univariate and the Multivariate Repeated Measures Model Fit

Model fitness determines if a statistical model adequately explains the variation in the data and can be used to compare two or more competing models after statistical analysis (Littell *et al*, 2002, 2005). Three fit statistics, -2 Residual Log Likelihood (RLL), AIC (Akaike, 1974) and BIC (Schwarz, 1978) were used to compare the univariate and multivariate repeated measures analysis model fitness. Smaller values of the fit statistics indicated a better model fit.

The multivariate repeated measures analysis produced consistently lower values of the AIC, BIC and RLL for the yield traits (cane and SDM) and stalk height than the univariate analysis (Table 5.6). The quality traits (ERC % cane and Fiber % cane) and stalk diameter produced similar values of the fit statistics for the multivariate repeated measures and univariate analysis. The lower values of the AIC statistic of the yield traits and stalk height for multivariate repeated measures analysis indicated better model fit and lower Type I errors than the univariate analysis (Guerin and Stroup, 2000).

The differences in model fitness between the multivariate repeated measures and univariate analysis was further evaluated using likelihood ratio tests. The likelihood ratio test tests if the model that produced a lower RLL has a significantly better model fit to the data than the model that produced a larger RLL. The likelihood ratio statistic is calculated as the difference between the RLL of the univariate and the multivariate repeated measures analysis. The likelihood ratio statistic follows a Chi-square distribution with degrees of freedom equal to the difference between the numbers of covariance parameters modelled by the models being compared. The multivariate repeated measures fitted three covariance parameters and univariate analysis fitted two, producing one degree of freedom for the test. The multivariate repeated measures analysis produced significantly ($P < 0.001$) better model fit for yield traits and stalk

height than univariate analysis (Table 5.6). The likelihood ratio tests for the quality traits and diameter produced non-significant ($P > 0.05$) value of the likelihood ratio test statistic, indicating similar model fit between the univariate and the multivariate repeated measures analysis methods.

5.3.4 Efficiency of the Univariate and Multivariate Repeated Measures Analysis in Determining Differences Between Experimental Genotypes and the Control Cultivar

Sugarcane breeders are generally interested in comparing the experimental genotypes to the control genotype (usually the dominant or widely grown cultivar) using data from variety trials. Experimental genotypes that produce significantly greater yield than the control cultivar are recommended for release to growers particularly if other important traits such as disease and pest tolerance are acceptable. The 15 experimental genotypes in this study were compared to the control (genotype 16, the most widely grown cultivar in Zimbabwe (Zhou, 2004)) using Dunnett's test for both the univariate and multivariate repeated measures analysis using the SAS code in Appendix 4. At Triangle, the univariate analysis produced highly significant ($P < 0.001$) differences between the experimental genotypes and the control cultivar for cane yield for all genotypes (Table 5.7) while the multivariate repeated measures analysis showed that six of the experimental genotypes were similar to the control. At Mkwasine, two experimental genotypes that were significantly ($P < 0.01$) different from the control cultivar using the univariate analysis were found similar to the control by the multivariate repeated measures analysis. The SDM of all genotypes at Triangle was significantly ($P < 0.001$) different from the control using univariate analysis but seven genotypes were found similar to the control by the multivariate repeated measures analysis. The SDM yield of three genotypes at the Mkwasine location was similar to the control using the multivariate repeated measures analysis but showed significant ($P < 0.01$) differences using the univariate analysis.

Table 5.6: The Model Fit Statistics (-2Residual log likelihood (RLL), Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC)) and the probability of obtaining a larger value of the Likelihood ratio test statistic (P-value) for yield, quality and agronomic traits derived from the univariate and multivariate repeated measures (multivariate) analysis for the data from the Triangle and Mkwasine locations.

Fit Statistic	Location	Method	Yield traits		Quality traits		Agronomic traits	
			Cane yield (t/ha)	SDM yield (t/ha)	ERC % cane	Fiber % cane	Stalk height (m)	Stalk diameter (cm)
RLL	Triangle	Univariate	4413.1	3215.2	1269.7	1795.1	96.2	-237.3
		Multivariate	4219.3	3047.2	1269.4	1793.9	-15.1	-239.9
		P-value	<0.001	<0.001	0.584	0.273	<0.001	0.107
	Mkwasine	Univariate	3065.2	2150.2	959.7	1256.0	10.1	-102.0
		Multivariate	3020.1	2112.0	959.3	1254.3	-12.7	-104.9
		P-value	<0.001	<0.001	0.527	0.192	<0.001	0.089
AIC	Triangle	Univariate	4417.1	3219.2	1273.7	1799.1	100.2	-233.3
		Multivariate	4225.3	3053.2	1275.4	1799.9	-9.1	-233.9
	Mkwasine	Univariate	3069.2	2154.2	963.7	1260.0	14.1	-98.0
		Multivariate	3026.1	2118.0	965.3	1260.3	-6.7	-100.9
BIC	Triangle	Univariate	4416.3	3218.5	1272.9	1798.3	99.4	-234.1
		Multivariate	4224.1	3052.1	1274.2	1798.7	-10.3	-235.0
	Mkwasine	Univariate	3068.0	2153.0	962.4	1258.8	12.9	-99.3
		Multivariate	3024.3	2116.1	963.4	1258.4	-8.5	-102.1

Table 5.7: The significance levels of the difference between the experimental genotypes and control cultivar for the cane and SDM yield when the data from Triangle and Mkwasine locations was analyzed using the univariate (UNIV) and multivariate repeated measures (MRM) analysis.

Genotype	Cane yield (t/ha)				SDM (t/ha)			
	Triangle		Mkwasine		Triangle		Mkwasine	
	UNIV	MRM	UNIV	MRM	UNIV	MRM	UNIV	MRM
1	***	***	***	***	***	***	***	***
2	***	***	***	***	***	***	***	***
3	***	***	***	***	***	***	***	***
4	***	***	***	***	***	***	***	***
5	***	*	***	***	***	NS	***	***
6	***	NS	NS	NS	***	NS	**	NS
7	***	**	***	***	***	NS	***	***
8	***	**	***	NS	***	***	***	***
9	***	NS	***	**	***	**	***	***
10	***	NS	NS	NS	***	NS	NS	NS
11	***	NS	**	NS	***	*	***	**
12	***	*	***	**	***	NS	***	NS
13	***	***	***	***	***	***	***	***
14	***	NS	NS	NS	***	NS	NS	NS
15	***	NS	NS	NS	***	NS	NS	NS

* = significant at 0.05.
** = significant at 0.01.
*** = significant at 0.001.
NS = not significant at P = 0.05.

For the quality traits (ERC % cane and Fiber % cane), both the univariate and multivariate repeated measures analyses produced similar trends in p-values for the tests of the differences between experimental genotypes and the control (Table 5.8). The results in Table 5.8 followed the same trends shown by the fit statistics and the likelihood ratio tests in Table 5.6.

The correlation coefficients between the crop-years for quality traits produced by multivariate repeated measures analysis ranged from -0.01 to 0.04 (data not shown) and were not significant ($P > 0.05$), indicating independence among crop-years. When the crop-years are independent, then the univariate and multivariate repeated measures analysis are expected to produce similar results.

Table 5.8: The significance levels of the difference between the experimental genotypes and control cultivar for the ERC % cane and Fiber % cane when the data from Triangle and Mkwasine locations was analyzed using the univariate (UNIV) and multivariate repeated measures (MRM) analysis.

Genotype	ERC % cane				Fiber % cane			
	Triangle		Mkwasine		Triangle		Mkwasine	
	UNIV	MRM	UNIV	MRM	UNIV	MRM	UNIV	MRM
1	***	***	NS	NS	NS	NS	NS	NS
2	***	***	***	***	NS	NS	NS	NS
3	***	***	***	***	***	***	***	***
4	***	***	***	***	***	***	***	***
5	***	***	***	***	***	***	***	***
6	***	***	**	**	***	***	***	***
7	***	***	***	***	NS	NS	NS	NS
8	***	***	***	***	***	***	***	***
9	***	***	***	***	***	***	***	***
10	***	***	***	***	***	***	***	***
11	***	***	***	***	***	***	***	***
12	***	***	***	***	NS	NS	NS	NS
13	***	***	***	***	***	***	***	***
14	NS	NS	NS	NS	NS	NS	NS	NS
15	NS	NS	NS	NS	NS	NS	NS	NS

* = significant at 0.05.
** = significant at 0.01.
*** = significant at 0.001.
NS = not significant at P = 0.05.

The means of the stalk height of four experimental genotypes from the data collected at the Triangle location were similar to that of the control when the data was analyzed using the multivariate repeated measures analysis. The same experimental genotypes showed significantly ($P < 0.05$) different mean stalk height compared to the control cultivar when the data was analyzed using the univariate method (Table 5.9). When the data collected at the Mkwasine location were analyzed using the multivariate repeated measures method, the mean of the stalk height of two experimental genotypes was similar to the control cultivar using the multivariate repeated measures analysis but the same experimental genotypes showed significant differences ($P < 0.01$) in stalk height between the experimental and the control cultivar when the data was analyzed using the univariate method.

There were similar trends in P-values for the differences in stalk diameter between the experimental genotypes and the control cultivar when the data was analyzed using the multivariate repeated measures and univariate methods (Table 5.9). The trends in P-values for the differences between the experimental genotypes and the control cultivar for stalk diameter were similar to the trends shown by the fit statistics and the likelihood ratio tests, where the multivariate repeated measures and the univariate analysis methods produced similar model fit statistics (Table 5.6). The correlation coefficient between the crop-years was -0.04 and non significant ($P > 0.05$) for the data from the Triangle and Mkwasine locations indicating independence of stalk diameter values across crop-years. Therefore the multivariate repeated measures and the univariate analysis methods would be expected to produce similar results.

5.4 Discussions

The multivariate repeated measures analysis method produced greater discrimination of the differences in cane and SDM yield between the experimental genotypes and the control cultivar

Table 5.9: The significance levels of the difference between the experimental genotypes and control cultivar for the stalk height and stalk diameter when the data from Triangle and Mkwasine locations were analyzed using the univariate (UNIV) and multivariate repeated measures (MRM) analysis.

Genotype	Stalk height (meters)				Stalk diameter (centimeters)			
	Triangle		Mkwasine		Triangle		Mkwasine	
	UNIV	MRM	UNIV	MRM	UNIV	MRM	UNIV	MRM
1	***	***	***	***	NS	NS	NS	NS
2	***	***	***	***	***	***	**	*
3	NS	NS	***	*	***	***	***	***
4	***	***	***	***	***	***	***	***
5	NS	NS	NS	NS	NS	NS	NS	NS
6	NS	NS	NS	NS	***	***	***	***
7	***	***	***	***	NS	NS	NS	NS
8	***	***	NS	NS	***	***	***	***
9	***	NS	NS	NS	***	***	***	***
10	NS	NS	NS	NS	***	***	***	***
11	*	NS	**	NS	***	***	***	***
12	**	NS	NS	NS	NS	NS	NS	NS
13	***	**	NS	NS	***	***	***	***
14	***	NS	**	NS	***	***	***	**
15	***	*	NS	NS	NS	NS	NS	NS

* = significant at 0.05.
** = significant at 0.01.
*** = significant at 0.001.
NS = not significant at P = 0.05.

than the univariate method. The univariate analysis method declared that most experimental genotypes were significantly different from the control but these genotypes were found to be similar to the control when the data was analyzed using the multivariate repeated measures method. The implications of this result to sugarcane breeders is that genotypes similar to the control are currently being declared significantly superior or inferior to the control using the univariate method, because the univariate method is widely and exclusively used by plant breeders to analyze advanced variety trials data. The implication of erroneously declaring that a genotype was significantly higher yielding than the control when it was similar to the control is that some genotypes that were released as higher yielding would produce lower yield than expected in the commercial crops. Such genotypes would eventually show no yield benefits to the growers than the current cultivar that they are intended to replace. This scenario has occurred many times in the sugarcane industries where released varieties have produced no yield gains in commercial crops. Conversely, erroneously rejecting genotypes as inferior when they are similar to the control could also result in the loss of parental germplasm that would be similar to the control but excelling in other important traits such as disease and insect pest resistance. An example could be the case with the sugarcane borer (White *et al*., 1996) where very few sugarcane borer resistant genotypes are advanced because of low yield. Erroneously discarding potential parental genotypes could also narrow genetic diversity in the breeding populations.

The poor discrimination of the differences in yield between the experimental genotypes and the control cultivar when the data was analyzed using the univariate method could explain the phenomenon of yield plateau alluded to by some sugarcane breeders (Garside *et al*., 1997). The univariate methods are currently widely used for the analysis of advanced variety trials data. Some sugarcane breeding programs including Australia (Garside *et al*., 1997) and South Africa

have reported yield plateaus. This yield plateau could be attributed to some of the varieties that are released being erroneously described as significantly higher yielding when they are similar to the control because of the error due to the use of the univariate analysis method. The use of the multivariate repeated measures analysis is likely to help alleviate the screening of varieties and quantify if indeed there is a yield plateau in sugarcane variety improvement. Multivariate repeated measures would also offer a potential to assist in breaking the yield plateau by producing more accurate statistical comparisons of genotypes for yield during selection and advanced variety testing.

The multivariate repeated measures analysis produced significantly better model fit to the data than the univariate analysis method. The better model fitness suggests that the multivariate repeated measures analysis is explaining more of the variation within the data than the univariate analysis method. Better model fitness also indicates that the variances used for computing the tests of the effects are neither inflated nor deflated and therefore correct variances would be used for testing the genotype effects. The multivariate repeated measures achieved better model fitness by accounting for the correlation between the variables as well as the correlations between measurements of the variables measured between the crop-years. The correlation between variables and between crop-years is ignored when the data is analyzed by the univariate method because of the assumption of independence by this analysis. The covariances that measure the correlations between the variables and the correlations between the crop-years are also added to the variances of each variable during the computation of experimental errors that are used to perform test of effects with the multivariate repeated measures analysis. The test of the differences between the experimental genotypes and the control cultivar were inflated by the univariate method because of the exclusion of the covariance in the computation of the

experimental errors. The statistical power of the multivariate repeated measures comes from the inclusion of covariance in the computation of experimental errors.

The large significant differences declared by the univariate method for yield traits are likely to be due to the higher Type I errors than with the multivariate repeated measures. The Type I errors occur when significant differences are erroneously declared during statistical tests (Allchin, 2001). In Table 5.6, AIC values, -2 Residual Log Likelihood and likelihood ratio tests showed that the univariate method produced a significantly poorer model fit than the multivariate repeated measures analysis. One of the consequences of a poorer model fit is the underestimation of experimental errors. Some of the variation in the data remains unaccounted for when there is a poor model fit. Underestimating experimental error increases Type I errors. The underestimation of experimental errors by the univariate analysis is caused by the assumption of independence. The covariance between the variables and the covariance between crop-years is ignored by the univariate method as the covariance is assumed to be negligible. This covariance is used to account for the correlation between the variables and the correlation between the crop-years. The multivariate repeated measures include the covariance between the variables and between crop-years to estimate experimental errors, thereby reducing Type I errors and increasing the power of the tests.

The yield traits showed the greatest difference between the tests of the difference between the experimental genotypes and control cultivar by the univariate and multivariate repeated measures analysis. Yield traits are generally more difficult to improve through plant breeding and selection compared to other traits because they are controlled by quantitative genes and therefore more susceptible to the influence of genotype by environment interaction effects (Falconer and Mackay, 1996; Mirzawan *et al.*, 1993). Because of the large genotype by

environment interaction effects, more accurate statistical methods are required to separate the genotypes effects from the environmental effects and thereby identifying true genotype differences. The likely large Type I errors associated with the univariate analysis would also decrease the precision of the tests when the genotype by environmental interaction effects increase. The negative effects of the genotype by environmental interaction are increased by the poorer model fit associated with the univariate analysis compared to the multivariate repeated measures analysis. In a study of genotype by environment interaction and resource allocation by Kimbeng *et al*. (2009), differences in cane yield of less than 15 – 20 % was proven to be more difficult to detect in advanced variety trials. However, contrary to the findings reported by Kimbeng *et al*. (2009), in this study, the univariate analysis method showed significant differences for cane yield between the experimental and the control ranging from 5 to 10 %, a result that is likely to be caused by the high Type I errors. The multivariate repeated measures analysis method found such differences not significant, a result likely to be correct. Therefore the multivariate repeated measures analysis method offer a more statistically powerful method for identifying true differences between the experimental genotypes and also for reducing the effects of genotype by environmental interaction for yield traits.

Significant gains have been achieved and continue to be achieved for sucrose content using the univariate analysis method. This study also explains one of the reasons why gains in sucrose content have remained higher than those for cane yield. Univariate analysis was shown to be similar to the multivariate repeated analysis in this study, indicating that these gains can partly be attributed to correct statistical analysis for quality traits using the univariate method. The study by Kimbeng *et al*. (2009) also showed that the influence of genotype by environment interaction effects was lower for sucrose content than for cane yield. Studies in Australia have

shown that the effects for genotype by environment interaction (Bull *et al*., 1992) and of competition between genotypes (Jackson and McRae, 2001) were lower for sucrose content than for cane yield.

The multivariate repeated measures analysis method was statistically more powerful at identifying true differences between the experimental genotypes and control by using correct experimental errors for the tests. This is important because sugarcane breeders are generally more interested in discarding low yield genotypes as well as identifying genotypes that significantly out yield the control. The significantly superior yielding genotypes would be targeted for release as commercial cultivars. Genotypes that are similar to the control would be useful as parents for the future particularly if they excel in other important traits such disease and insect pest resistance. The genotypes that are similar in yield to the control and have higher disease and insect resistance would be valuable for resistance breeding. Statistical methods that reject varieties that are similar to the control are undesirable as some potential good parental genotypes would be lost and could result in narrowing the genetic diversity among breeding populations. Equally important, statistical methods that declare significantly larger yield when the genotypes are similar to the control are misleading the breeder. These erroneously superior yielding genotypes would be included in further yield testing when they should have been discontinued. Therefore many varieties are included in advanced testing, increasing the costs of the breeding program. The multivariate repeated measures would reduce these erroneous interpretations and could probably reduce the cost of advanced variety testing by advancing fewer and higher yielding genotypes.

The multivariate component of the multivariate repeated measures is also important in determining the validity of the significance of further tests including univariate tests. Because the

multivariate tests include the covariance between variables, they are more precise than the univariate tests. When the multivariate tests are not significant, significant univariate tests should not be interpreted because they are likely to be due to Type I errors (Johnson and Wichern, 2002). Significant univariate tests should only be interpreted when the multivariate tests are significant when analyzing data that comprise a multivariate structure. Therefore the multivariate repeated measures analysis provides a quality control for the statistical analysis that includes multiple response variables measured from experimental units such as is the case with sugarcane breeding data and other crops with similar data structure.

5.5 Conclusions

The multivariate repeated measures produced significantly better model fits than the univariate analysis for yield traits. Multivariate repeated measures analysis method was more discriminating for the differences in yield between the experimental genotypes and the control than univariate analysis. Greater discrimination would result in correct selection decision during variety testing for the yield traits. Multivariate repeated measures analysis produced correct computation of experimental errors by including the covariance between variables as well as the covariance between crop-years leading to correct tests particularly for the yield traits. Univariate analysis was likely to have larger Type I errors because of the violation of the assumption of independence that would result in the underestimation of experimental errors. Multivariate repeated measures would reduce the erroneous interpretations for the yield traits likely to be associated with univariate analysis. Multivariate repeated measures analysis was a potentially powerful statistical tool for controlling the influence of genotype by environment interaction effects generally associated with complex traits like cane yield that are controlled by quantitative

genes. Quality traits showed that univariate analysis was adequate in identifying the true differences between genotypes.

5.6 References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* AIC-19: 716 – 723.

Allchin, D. (2001). Error types. Perspective on Science. Volume 9 Number 1: 38 – 58.

Berding, N., Hogarth, M. and Cox, M. (2004). Plant improvement of sugarcane. In: G. James (editor): Sugarcane: Second Edition. Blackwell Publishing: 20 – 77.

Bull, J.K., Hogarth, D.M. and Basford, K.E. (1992). Impact of genotype by environment interaction on response to selection in sugarcane. *Australian Journal of Experimental Agriculture* 32: 731 – 737.

Clowes, M.St.J. and Breakwell, W.L. (1998). Zimbabwe Sugarcane Production Manual. Zimbabwe Sugar Association Experiment Station, Chiredzi, Zimbabwe.

Ellis, R.D. and Merry, R.E. (2004). Sugarcane agriculture. In: G. James (editor): Sugarcane: Second Edition. Blackwell Publishing: 101 – 142.

Falconer, D.S. and Mackay, T.F.C. (1996). Introduction to Quantitative Genetics. Fourth Edition. Longman Group Ltd, UK.

Freund, R.J. and Wilson, W.J. (2003). Statistical Methods. Third Edition. Academic Press, New York.

Galecki, A.T. (1994). General class of covariance structures for two or more repeated factors in longitudinal data analysis. *Communications in Statistics, 23,* 3105 – 3119.

Garside, A.L., Smith, M.A., Chapman, L.S., Hurney, A.P. and Magarey, R.C. (1997). The yield plateau in the Australian Sugar industry: 1970 – 1990. In Keating, B.A. and Wilson, J.R. (editors). Intensive Sugarcane Production: Meeting the Challenges Beyond 2000. CAB International, Wallingford, United Kingdom: 103 – 124.

Guerin, L. and Stroup, W.W. (2000). A simulation study to evaluate PROC MIXED analysis of repeated measures data. *Proceedings of the 12th Annual Conference on Applied Statistics in Agriculture*. Manhattan, KS: Kansas State University.

Jackson, P.A. and McRae, T.A. (2001). Selection of sugarcane clones in small plots: Effects of plot size and selection criteria. *Crop Science* 41: 315 – 322.

Johnson, R.A. and Wichern, D.W. (2002). Applied Multivariate Statistical Analysis. Prentice Hall, New Jersey, USA.

Kang, M.S., Miller, J.D., Tai, P.Y.P., Dean, J.L. and Glaz, B. (1987). Implications of confounding of genotype x year and genotype x crop effects in sugarcane. *Field Crops Research* 15: 349 – 355.

Kimbeng, C.A., Zhou, M.M. and da Silva, J.A. (2009). Genotype by environment interactions and resource allocation in sugarcane yield trials in the Rio Grande valley region of Texas. *Journal of the American Society of Sugarcane Technologists* (In press).

Littell, R.C., Stroup, W.W. and Freund, R.J. (2002). SAS for Linear Models. Fourth Edition. SAS Institute Inc., Cary, NC, USA.

Littell, R.C., Milliken, G.A., Stroup, W.W. and Wolfinger, R.D. (2005). SAS System for Mixed Models. 7th edition. SAS Institute Inc., Cary, NC, USA.

Mirzawan, P.D.N., Cooper, M. and Hogarth, D.M. (1993). The impact of genotype by environment interactions for sugar yield on the use of indirect selection in southern Queensland. *Australian Journal of Experimental Agriculture* 33: 629 – 638.

Morrison, D.F. (1976). Multivariate Statistical Methods. McGraw-Hill Book Company, Second Edition, New York, USA.

Moser. E.B. (2005). Multivariate Statistical Data Analysis. Course Notes. Department of Experimental Statistics. Louisiana State University.

Salassi, M.E. and Giesler, G.G. (1995). Projected costs and returns – sugarcane, Louisiana, 1995. Department of Agricultural Economics and Agribusiness, AEA Information Series Number 132, LSU AgCenter, Baton Rouge, Louisiana, USA.

SAS Institute (2007). SAS for windows, Version 9.1.3. Cary, North Carolina, USA.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* 6: 461 – 464.

Srivastava, M.S. and Carter, E.M. (1983). An Introduction to Applied Multivariate Statistics. North-Holland, Amsterdam.

White, W.H., Legendre, B.L., and Miller, J.D. 1996. Progress in breeding for sugarcane borer resistance. *Sugar Cane* 5: 3 – 7.

Zhou, M. (2004). Performance of varieties N14 and NCo376 in the South East Lowveld of Zimbabwe. *Proceedings of the South African Sugarcane Technologists Association* 78: 153 – 160.

CHAPTER 6: CROSS RESISTANCE BETWEEN THE MEXICAN RICE BORER AND THE SUGARCANE BORER (*LEPIDOPTERA:* CRAMBIDAE): A CASE STUDY USING SUGARCANE BREEDING POPULATIONS

6.1 Introduction

Moth (Lepidoptera) stem borers are major pests of sugarcane (*Saccharum* spp. hybrids) and other important gramineous crops worldwide. Stem borers attacking tropical gramineous crops chiefly belong to the families Pyralidae, Noctuidae and Castniidae (Smith *et al.*, 1993). These authors list 19 important genera of the family Pyralidae; however, later taxonomic revisions separate the tribe Crambinae from Pyralidae creating Crambidae as an additional family (Munroe and Solis, 1999). In North America, two important stem borers are the crambids, *Diatraea saccharalis* (F.) (= sugarcane borer) and *Eoreuma loftini* (Dyar) (= Mexican rice borer). The sugarcane borer has been the dominant stem borer of sugarcane in the U.S.; however, in 1980, the Mexican rice borer became established in the Lower Rio Grande Valley of Texas (Johnson and van Leerdam, 1981) and subsequently supplanted the sugarcane borer as the dominant insect pest of that industry (Johnson, 1984). Reay-Jones *et al.* (2007) predicted the arrival of the Mexican rice borer into Louisiana in 2008 and the first specimens were indeed found near Vinton, Louisiana on December of 2008.

While being taxonomically closely related species and sharing many of the same cultivated and wild hosts, the species are a contrast to one another in certain aspects of their behavior: in particular, oviposition behavior. The ovipositor of the Mexican rice borer is laterally compressed, which allows oviposition in crevices, whereas the ovipositor of the sugarcane borer is vertically depressed, which facilitates oviposition on flat surfaces (Smith *et al.* 1993). Mexican rice borer eggs are found deep within the canopy and near the soil surface on dry leaf material while the moths of the sugarcane borer oviposite on the young leaves in the sugarcane canopy.

However, once the first instar larvae eclose from the egg, the behavior of the two species becomes similar. Specifically, both species' larvae move to the green leaf sheaths and begin feeding (Ring *et al*., 1991 and White, 1993) and later the larvae then bore into the young, developing internodes. The speed with which these larvae enter the stalk depends upon the genotype: more quickly in susceptible genotypes than resistant ones (White *et al*., 1996).

We hypothesize that, due to the similarities in larval feeding behavior of the two species, selecting for resistance to one species will obtain resistance to the other, henceforth referred to as cross resistance. Being able to accept this hypothesis would be a great benefit to both the Louisiana and Texas sugarcane breeding programs as it would eliminate the need for maintaining dual breeding programs needed to develop resistance to both species.

The objective of this study was to determine if cross resistance exist among sugarcane genotypes between the Mexican rice borer and the sugarcane borer using breeding populations derived from Louisiana and Texas breeding programs.

6.2 Materials and Methods

Eighty sugarcane genotypes were planted at the SRS farm, in Santa Rosa, Texas on November 11, 2005. The field design was a randomized complete block design with four replications. Individual plots were 6 m in length. Thirty of the genotypes were from Louisiana, and they represented clones chosen at random from different sub-populations. Sixteen of the thirty genotypes were from the USDA, ARS Sugarcane Research Laboratory's recurrent selection program for sugarcane borer resistance (White *et al*., 1996). Ten were sampled from among commercial genotypes and were identified as either resistant or susceptible to the sugarcane borer based on previous field evaluations. The remaining four genotypes were selected for the self-stripping trait (e.g. leaves and associated sheath drop from the cane stalk) and their

resistance status was unknown. The 50 genotypes from the Texas A&M University program were from the 2002 breeding series and their resistance status to either stem borer species was unknown and therefore, represented a random population.

Standard cultural practices for cultivating sugarcane in Texas were followed. However, no insecticide applications were made to the experiment and damage was ascribed to native infestations of both stem borers.

On August 2 to 9, 2006 random 10-stalk samples were hand-cut from each plot in the plant-cane. These stalks were topped at the last fully-expanded internode and stripped of all leaves and leaf-sheaths. The samples were returned to the laboratory where stalks were evaluated for insect damage. Stalks were split longitudinally and the pieces examined externally and internally for presence of larvae or presence of larval entrance and moth exit holes. Although both species damage sugarcane by boring into the internodes, the sugarcane borer primarily makes longitudinal tunnels in the internodes, whereas the Mexican rice borer often bores around and across the internode causing transverse tunnels (Johnson, 1981). Additionally, sugarcane borer larvae regularly deposit their frass outside the entrance of the tunnel. In contrast, the Mexican rice borer larvae maintain closed tunnels by plugging the traversed area with frass and detritus, thus packing the tunnels (Meagher *et al.,* 1994). These two contrasts in tunneling behavior were used to distinguish between the two stem borer species when making the damage assessment. The assessment involved counting the number of bored internodes on each stalk made by each of the pests. The data, percent borer-damaged internodes, were computed as the ratio of the bored internodes per plot to the total number of internodes per plot expressed as a percentage.

The sugarcane in the experiment was harvested and allowed to ratoon and data were again collected in the first-ratoon crop. The first-ratoon crop was harvested on August 6 to 8, 2007 and the data were collected following the same procedure as in the plant-cane crop.

6.2.1 Data Analysis

The data collected from field plots were coded to identify genotypes as subsets from either Louisiana or Texas. The Louisiana population was further coded to identify three subgroups based on prior information. One subgroup comprised of 18 genotypes that were known to be resistant to the sugarcane borer. A second subgroup comprised 7 genotypes known to be susceptible to the sugarcane borer, and the third subgroup had genotypes with unknown resistance status. The performance of the resistant and susceptible genotypes were used as a benchmark to classify all the 80 genotypes in the trial into two subgroups, that is, resistant or susceptible to the sugarcane borer. To achieve this we calculated the mean and their associated 95% confidence limits for each of the resistant and susceptible groups. Genotypes with % sugarcane borer-damaged internodes values lower than the upper confidence limit of the resistant group were classified as resistant while the rest were classified as susceptible to the sugarcane borer.

The experimental design variables in this experiment were populations (Texas and Louisiana), resistance status to the sugarcane borer (resistant and susceptible) and crop-year (plant and first-ratoon). The data were analyzed using the analysis of variance (ANOVA), analysis of covariance (ANCOVA) and log linear models. The rationale for these analyses was to determine the significant experimental design variables for % borer-damaged internodes, the strength of the association for % borer-damaged internodes between the two pests, and finally to

determine the strength of dependency of % borer-damaged internodes on experimental design variables (population/resistance status, and crop).

The ANOVA was done using SAS Mixed procedures (SAS Institute, 2007). The linear mixed model equation used was:

$$Y_{ijkm} = \mu + R_i + S_j + V(S)_{k(j)} + RV(S)_{ik(j)} + C_m + CS_{jm} + CV(S)_{km(j)} + RCV(S)_{ikm(j)},$$

Equation 6.1

where Y_{ijkm} is the % borer-damaged internodes from the *i*th replication (*i* = 1, 2), *j*th population (Texas vs. Louisiana) or resistance status (resistant vs susceptible to the sugarcane borer) (*j* = 1, 2), *k*th genotype (*k* = 1, 2, …, 80) and *m*th crop year (*m* = 1, 2), μ is the overall mean, R_i is the random effect from the *i*th replication, S_j is the fixed effect of the *j*th population, $V(S)_{k(j)}$ is the fixed effect of the *k*th genotype nested within the *j*th population, $RV(S)_{ik(j)}$ is the random effect of the interaction of the *i*th replication by the *k*th genotype nested within the *j*th population and was the experimental error for population and genotype nested within population effects, C_m is the fixed effect of the *m*th crop year, CS_{jm} is the interaction fixed effect of the *j*th population by the *m*th crop year, $CV(S)_{km(j)}$ is the interaction fixed effect of the *m*th crop year by the *k*th genotype nested within the *j*th population, and $RCV(S)_{ikm(j)}$ is the random interaction effect of the *i*th replication by the *m*th crop year by the *k*th genotype nested within the *j*th population and was the residual error.

ANCOVA was used to investigate the association in % borer-damaged internodes between the two borer species and if the type of association differed between the Louisiana vs Texas population and among resistant vs. susceptible genotypes. ANCOVA was run in SAS Mixed procedures with the % Mexican rice borer-damaged internodes as the response variable

and sugarcane borer % borer-damaged internodes as the covariate. The % sugarcane borer-damaged internodes was used as the covariate because the sugarcane borer resistance status of the Louisiana genotypes was known and was used to establish the resistance status of other genotypes in the trial, as the status of Mexican rice borer resistance is not known. The analysis determined the strength of the association between % borer-damaged internodes of the two species. The linear mixed model was:

$$Y_{ijkm} = R_i + S_j + V(S)_{k(i)} + RV(S)_{ik(j)} + C_m + CS_{jm} + CV(S)_{km(j)} + \beta x_{ijkm} + RCV(S)_{ikm(j)}, \quad \text{Equation 6.2}$$

where Y_{ijkm} is the % Mexican rice borer-damaged internodes in the *i*th replication, *j*th population or resistance status, *k*th genotype and *m*th crop year, β is the slope of the regression equation representing the association in the % borer-damaged internodes between the Mexican rice borer and sugarcane borer, x_{ijkm} is the % sugarcane borer-damaged internodes in the *i*th replication, *j*th population or resistance status, *k*th genotype and *m*th crop year and was the covariate. The ability of the mixed procedure to account for the random variation associated with experimental design variables increases statistical power of the tests compared to simple linear regression (Abraham and Ledolter, 2006).

The log linear analysis was done using the SAS GENMOD procedure and the linear model used was:

$$Log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ}, \quad \text{Equation 6.3}$$

where μ_{ijk} is the cell count from the combination of the *i*th crop year (plant or ratoon), *j*th population or resistance status and *k*th bored status (bored or not bored), λ is the intercept, λ_i^X is the effects of the *i*th crop, λ_j^Y is the effect of the *j*th population or resistance status, λ_k^Z is the

effect of the *k*th bored status, λ_{ij}^{XY} is the interaction effect of the *i*th crop year by the *j*th population (or resistance status), λ_{ik}^{XZ} is the interaction effect of the *i*th crop year by the *k*th bored status, λ_{jk}^{YZ} is the interaction effect of the *j*th population (or resistance status) by the *k*th bored status, and λ_{ijk}^{XYZ} is the interaction effect of the *i*th crop year by the *j*th population (resistance status) by the *k*th bored status. The interpretation of log linear analysis is done using odds ratios. The odds ratios are calculated by exponentiating the coefficients of the effects. For example, the odds ratios for the number of bored internodes in population Y_1 versus population Y_2 would be estimated as the ratio of the odds of being bored for population Y_1 to the odds of being bored for population Y_2 and were estimated using the formula,

$$\hat{\theta}_{YZ} = \frac{\left.\frac{\pi_Z}{1-\pi_Z}\right|_{Y=1}}{\left.\frac{\pi_Z}{1-\pi_Z}\right|_{Y=2}} = e^{\beta_{YZ}}. \qquad \text{Equation 6.4}$$

The confidence intervals of the odds ratios are calculated by exponentiating the coefficients of their confidence intervals.

6.3 Results

6.3.1 Analysis of Variance

The Mexican rice borer was the dominant species encountered during the study. The % borer-damaged internodes ascribed to the Mexican rice borer was approximately four times that of the sugarcane borer (Table 6.1). For both pests, borer damage was slightly greater (12% by the Mexican rice borer and 18% by the sugarcane borer) in the plant crop compared to the first ratoon crop, possibly because the plant crop is slower to establish and therefore, more susceptible to insect damage, e.g. the internodes remain susceptible to larval establishment for a greater length of time. However, the degree to which the two populations (Texas and Louisiana) were

damaged by both pests was not consistent across crops as evidenced from the significant population by crop effect (Table 6.1). The Louisiana population suffered significantly ($P < 0.01$) more (14%) damage by the Mexican rice borer compared to the Texas population in the plant crop whereas in the first ratoon crop, the Texas population suffered significantly ($P < 0.05$) more (11%) damage than the Louisiana population. Likewise for the sugarcane borer, the Texas population suffered significantly ($P < 0.01$) more damage (45 %) than the Louisiana population in the plant crop whereas in the ratoon crop, the Louisiana population suffered slightly more (non-significant) damage than the Texas population. When the data were averaged across crops, there was no clear indication of which population was superior against the Mexican rice borer while the Louisiana population suffered comparatively less damage from the sugarcane borer (Table 6.1). Little insight could be gained from this analysis with respect to the existence of cross-resistance between the two borer species.

A second data subset comprising sugarcane borer resistant and susceptible genotypes from the Louisiana population was subjected to ANOVA. As previously reported, the plant crop suffered significantly ($P < 0.05$) more borer damage from both pests (19 % by the Mexican rice borer and 32% by the sugarcane borer) compared to the first ratoon crop. As expected, genotypes previously identified as resistant to the sugarcane borer suffered significantly ($P < 0.01$) less damage from the sugarcane borer in both the plant (45 %) and first ratoon (52 %) crops compared to their susceptible counterparts. But more importantly, genotypes previously identified as resistant to the sugarcane borer also suffered significantly ($P < 0.01$) less damage from the Mexican rice borer in both crops compared to their susceptible counterpart. When the entire dataset comprising all 80 genotypes were subsequently coded as either resistant or susceptible to the sugarcane borer and analyzed, the trends were similar to those reported above

(Table 6.1). While the % borer-damaged internodes for both pests were significantly ($P < 0.05$) influenced by crop and prior sugarcane borer resistance status the interaction between resistance status and crop was not significant (Data not shown). The above results indicated that the significant variables for % borer-damaged internodes were similar for the Mexican rice borer and the sugarcane borer, and was suggestive of cross resistance in sugarcane between the sugarcane borer and the Mexican rice borer. However, it must be considered that, while the ANOVA can be used to show differences between groups, no reliable inferences can be made about the association between the borer-damaged internodes between the species. To accomplish this we performed the ANCOVA.

6.3.2 Analysis of Covariance

In the ANCOVA, the % Mexican rice borer-damaged internodes was used as the response variable while that of the sugarcane borer was used as the covariate (Table 6.2). A significant positive association would mean that the plants within the populations responded similarly to attack by both pests. The analyses produced two regression equations for each data set, one each for the Louisiana and Texas populations, and one each for the resistant and susceptible genotypes. The equations were,

$$MRBP_{Louisiana} = 20.54 + 0.36SCBP_{Louisiana}, \quad \text{Equation 6.5}$$

for the Louisiana population,

$$MRBP_{Texas} = 22.19 - 0.03SCBP_{Texas}, \quad \text{Equation 6.6}$$

for the Texas population,

$$MRBP_{Resistant} = 16.78 + 0.48SCBP_{Resistant}, \quad \text{Equation 6.7}$$

Table 6.1: Mean percent bored internodes and standard errors (S.E) for the Mexican Rice Borer and Sugarcane borer in the Plant and Ratoon Crops sampled from the Louisiana and Texas breeding populations, the Louisiana resistant and susceptible sub-populations and all genotypes reclassified into resistant and susceptible populations.

Population	Mexican Rice Borer			Sugarcane borer		
	Plant ± SE	Ratoon ± SE	Mean ± SE	Plant ± SE	Ratoon ± SE	Mean ± SE
Louisiana and Texas populations						
Louisiana	24.97 ± 1.02	19.83 ± 1.03	22.40 ± 0.85	5.24 ± 0.86	5.54 ± 0.86	5.39 ± 0.78
Texas	21.90 ± 0.87	22.10 ± 0.87	22.00 ± 0.76	7.61 ± 0.79	5.32 ± 0.79	6.47 ± 0.75
Mean	23.43 ± 0.80	20.97 ± 0.81	22.19 ± 0.42	6.43 ± 0.77	5.43 ± 0.77	5.93 ± 0.24
Significance	**	*	NS	***	NS	*
Louisiana Resistant and Susceptible Sub-populations						
Resistant	19.80 ± 1.17	16.90 ± 1.19	18.35 ± 1.00	4.01 ± 0.52	2.76 ± 0.53	3.38 ± 0.39
Susceptible	29.47 ± 1.74	24.60 ± 1.74	27.03 ± 1.36	7.26 ± 1.45	5.74 ± 0.85	6.50 ± 0.62
Resistant % Susceptible	67	69	68	55	48	52
Mean	24.63 ± 1.17	20.75 ± 1.17	20.66 ± 0.71	5.63 ± 0.51	4.25 ± 0.52	4.22 ± 0.34
Significance	***	***	**	**	**	*
All Genotypes reclassified into Resistant and Susceptible populations						
Resistant	19.53±1.29	17.20±1.31	18.37±1.03	3.74±0.94	2.62±0.94	3.18±0.87
Susceptible	25.57±1.16	24.32±1.17	24.95±0.95	8.76±0.90	7.12±0.91	7.94±0.85
Resistant % Susceptible	76	71	74	43	37	40
Mean	22.55±0.96	20.76±0.97	22.19±0.42	6.25±0.74	4.87±0.74	5.93±0.24
Significance	***	***	**	***	***	*

NS = not significant. *, **, *** significant at 0.05, 0.01, 0.001 respectively

for the resistant population (Louisiana subgroups),

$$MRBP_{Susceptible} = 24.84 - 0.34SCBP_{Susceptible}, \quad \text{Equation 6.8}$$

for the susceptible population (Louisiana subgroups),

$$MRBP_{Resistant} = 17.22 - 0.37SCBP_{Resistant}, \quad \text{Equation 6.9}$$

for the resistant population (80 genotypes grouped), and,

$$MRBP_{Susceptible} = 25.15 - 0.02SCBP_{Susceptible}, \quad \text{Equation 6.10}$$

for the susceptible population (80 genotypes grouped).

In the ANCOVA, the intercept of each population represents the level of damage caused by the Mexican rice borer whereas the slopes measure the strength of the association (Figure 6.1). The intercepts are represented by the Population (Vartype) or Resistance Status (Resistance) effect while the slopes are represented by the SCBP*Population or SCBP*Resistance Status effect (Table 6.2; Figure 6.1). The intercepts show that, similar levels of Mexican rice borer-damaged internodes were experienced by the Louisiana and Texas populations as previously indicated by the ANOVA (Table 6.1). However, the coefficient of the slope or SCBP*Population effect for the Louisiana population was positive and significant ($P < 0.01$) indicating a positive association in the % borer-damaged internodes between the Mexican rice borer and the sugarcane borer in this population. No such significant ($P > 0.05$) association was found for the Texas population.

The second data set comprised Louisiana genotypes with known sugarcane borer resistance status. The sugarcane borer resistant genotypes had a smaller intercept indicating that this group incurred less Mexican rice borer damage than their susceptible counterpart (Table 6.2). A significant association ($P < 0.01$) was found between the Mexican rice borer and the

Table 6.2: The estimate, standard error, t-value and probability of obtaining a larger t-value (Pr > |t|) for the intercepts (Vartype) and slopes (SCBP*Vartype) derived from the analysis of covariance of the % Mexican rice borer-damaged internodes (response variance) and % sugarcane borer-damaged internodes (covariate) for the Louisiana and Texas populations, the Louisiana resistant and susceptible sub-populations and all the genotypes reclassified into resistant and susceptible populations

Effect	Population	Estimate	Standard Error	t Value	Pr > |t|
Louisiana and Texas Populations					
Vartype	Louisiana	20.54	1.43	14.33	0.0001
Vartype	Texas	22.19	1.27	17.48	0.0001
SCBP*Vartype	Louisiana	0.36	0.12	3.01	0.0001
SCBP*Vartype	Texas	-0.03	0.08	-0.36	0.7181
Louisiana Resistant and Susceptible Sub-populations					
Resistance	Resistant	16.78	1.75	9.58	0.0001
Resistance	Susceptible	24.84	2.55	9.76	0.0001
SCBP*Resistance	Resistant	0.48	0.18	2.67	0.0095
SCBP*Resistance	Susceptible	0.34	0.22	1.51	0.1346
All Genotypes Reclassified into Resistant and Susceptible populations					
Resistance	Resistant	17.22	0.84	20.57	0.0001
Resistance	Susceptible	25.15	0.80	31.46	0.0001
SCBP*Resistance	Resistant	0.37	0.18	2.04	0.0423
SCBP*Resistance	Susceptible	-0.02	0.08	-0.27	0.7867

sugarcane borer among the group of genotypes identified as resistant to the sugarcane borer, whereas, no such association ($P > 0.05$) was found among the susceptible genotypes (Table 6.2). Similar trends were found when the data subset comprising the sugarcane borer resistant versus susceptible groups drawn from all 80 genotypes was analyzed (Table 6.2). These results indicate that factors that influence the % borer-damaged internodes were similar for the sugarcane borer and the Mexican rice borer (Equations 6.5, 6.7, 6.9) and are also suggestive of cross resistance between the sugarcane and Mexican rice borers. However, whereas the ANCOVA showed that there was significant association in the response of sugarcane genotypes to infestation by the two borer species, the method in strict statistical terms is not robust enough to identify factors responsible for the association. To accomplish this we used Log linear models which can determine the variables responsible for the response.

6.3.3 Log Linear Model Analysis

Log linear models have not typically been used in these types of analyses, therefore, additional details about how it is being applied in this study is warranted. The input data used by log linear models are counts arranged in a contingency table with the cell values in the table treated as the response variable (Table 6.3). In this study, the experimental design variables namely, the crop, population, and borer damage status (bored or not bored) were treated as the independent variables. The analysis was used to determine the association or independence of one factor (e.g. borer damage status) on the other factors (crop, population), but the interpretation is based on the odds of insect damage (borer damage status) occurring in one crop or population relative to another. This is accomplished by first identifying and interpreting the significant interaction effects in the model. For example, to determine the extent to which borer damage status is dependent on the crop, the population, and their interaction (that is, of crop and population), one

would first identify if the estimates or coefficients associated with the borer damage status by crop, borer damage status by population, and borer damage status by crop by population interaction effects, respectively, are significant in the model.

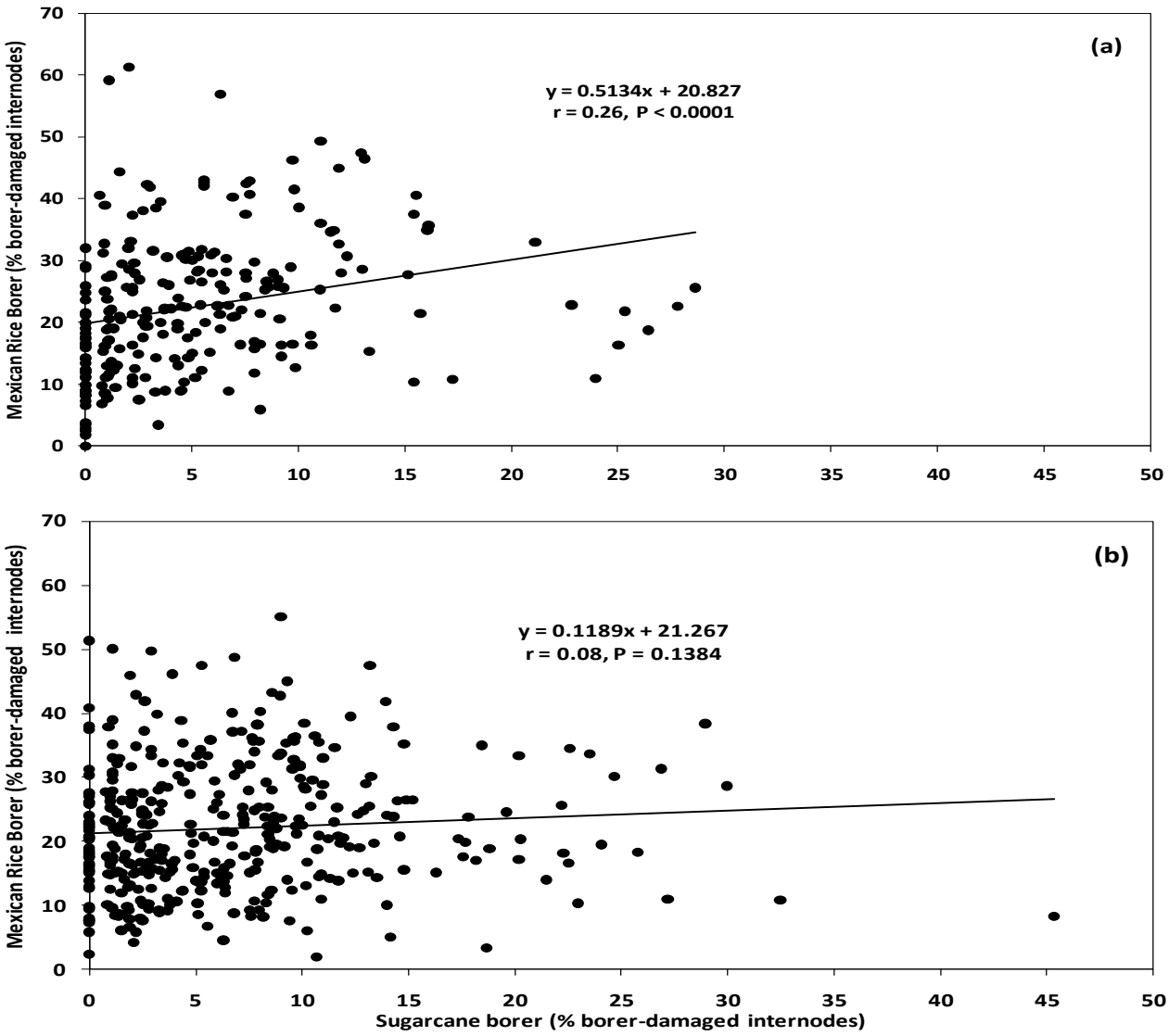

Figure 6.1: Mexican Rice Borer (% borer-damaged internodes) plotted against Sugarcane borer (% borer-damaged internodes) for genotypes selected from the Louisiana (a) and Texas (b) breeding populations. The trends in the graphs were fitted using simple linear regression and the coefficients are different from those in Table 2, that were derived from the analysis of covariance using the mixed procedure of SAS. The mixed procedure of SAS removes the variation associated with random variables such as replication.

Each of these coefficients is associated with a standard error, confidence intervals, chi-square and probability of the chi-square value which is used to test for its significance as shown in Table 6.4. A significant estimate for the effect of borer damage status by crop or borer damage status by population would indicate a departure from unity (1.0) meaning that the borer damage status was dependent on the crop or population, respectively. To derive the comparisons, these coefficients are exponentiated to produce odds ratios and it is the odds ratios that are interpreted. For example, the odds ratios for borer damage occurring in the Louisiana versus Texas population are calculated as the odds of being bored for the Louisiana population divided by the odds of being bored for the Texas population. Significant, positive odds ratios that are greater than 1.0 would mean that the odds of borer damage occurring in the Louisiana population were greater than that for the Texas population.

For example, when the dataset comprising sugarcane borer resistant and susceptible genotypes from the Louisiana population were subjected to log linear model analysis, the output would be as shown in Table 6.4. The highest significant effects were the two-way interactions of borer status by resistance and borer status by crop. From Table 6.4, the resistance by crop effect is not within the scope of this study and is therefore not interpreted. The resistance by Mexican rice borer and crop by Mexican rice borer effects are interpreted because they provide information on how the resistance status to the sugarcane borer and the crop affects the % Mexican rice borer-damaged internodes, respectively. The coefficients of the resistance by Mexican rice borer (-0.5055), from Table 4 represents the log of the odds ratio of being bored by the Mexican rice borer for the resistant genotypes versus the susceptible genotypes. The odds ratio (resistance versus susceptible) are calculated by exponentiating the log odds ratio (-0.5055) and is equal to 0.60 (Table 6.5). The confidence limits of the odds ratio 0.60, are calculated by

exponentiating the confidence limits of the log of the odds ratio (-0.5826, -0.4285) to provide the limits 0.56 and 0.65 in Table 6.5. The log of the odds ratio divided by the standard error (S.E.) provides a t-statistic which is squared to provide the Chi-square statistic (Chi-square) in Table 6.5. The probability (Pr > ChiSq) of obtaining a larger value of the Chi-Square statistic is given in Table 4 and this value can be read off the Chi-Square tables and indicates the strength of the dependency or association of the borer damage to the resistance status.

Log linear model analysis is interpreted the same way as factorial analysis. As with factorial analysis, only the highest order significant interaction effect is interpreted. In this study, the borer damage status by crop by population interaction effect for the Louisiana and Texas populations was significant ($P < 0.01$). Significant three-way interactions are interpreted by comparing two variables at each level of a third, just as is done with factorial analysis. Therefore, odds ratios for borer damage in the Louisiana versus Texas population were compared for each crop.

One area of similarity between ANOVA and log linear models is that both populations incurring borer damage were not consistent across crops (Table 6.5). In the plant crop Louisiana genotypes were 17 % (significant at $P < 0.05$) more likely to incur Mexican rice borer damage than Texas genotypes but the reverse was true in the first ratoon crop where Louisiana genotypes were 35 % (significant at $P < 0.05$) less likely to incur Mexican rice borer damage. For the sugarcane borer, Louisiana genotypes were 10 % (significant at $P < 0.05$) less likely to incur borer damage than Texas genotypes in the plant crop and in the first ratoon crop, genotypes from both populations were equally likely to incur borer damage as the odds ratio was close to 1.00 (Table 6.5) and the coefficient was not significant ($P > 0.05$).

Table 6.3: Number and proportion (in brackets) of internodes bored by the Mexican rice borer and Sugarcane borer in the plant and ratoon crops for the genotypes from the Louisiana and Texas populations, the Louisiana resistant and susceptible sub-populations, and all the genotypes reclassified into resistant and susceptible populations.

Crop	Population	Mexican Rice Borer		Sugarcane Borer	
		Bored	Not Bored	Bored	Not Bored
Louisiana and Texas populations					
Plant	Louisiana	3392 (0.25)	10334	678(0.05)	13048
	Texas	5016(0.22)	17833	1686(0.07)	21163
Ratoon	Louisiana	2066(0.20)	8245	507(0.05)	9804
	Texas	3685(0.22)	13280	846(0.05)	16119
Louisiana Resistant and Susceptible Sub-populations					
Plant	Resistant	1569(0.25)	6366	307(0.04)	7628
	Susceptible	773(0.42)	1851	185(0.08)	2439
Ratoon	Resistant	999(0.20)	4947	166(0.03)	5780
	Susceptible	518(0.32)	1596	123(0.06)	1991
All Genotypes Reclassified into Resistant and Susceptible populations					
Plant	Resistant	3042 (0.19)	12590	562 (0.04)	15070
	Susceptible	5366 (0.26)	15577	1802 (0.09)	19141
Ratoon	Resistant	2000 (0.17)	9658	304 (0.03)	11354
	Susceptible	3571 (0.24)	11867	1049 (0.07)	14569

Table 6.4: The output from log linear models showing the parameters, levels of resistance [Resistant (R) and Susceptible (S)], levels of crop [Plant (P) and Ratoon (R)], and levels of internode borer damage status (Damage (D) and Not Damaged (N)), Degrees of freedom (DF), parameter estimates (estimate), Standard error of the estimates (S.E.), the Wald 95 % confidence limits, Chi-square value (Chi-square) and the probability of obtaining a larger Chi-square value (Pr > ChiSq) for the Mexican rice borer (MRB) data.

Parameter	Levels		DF	Estimate	S.E.	Wald 95 % Confidence Limits		Chi-Square	Pr > ChiSq
Intercept			1	7.3701	0.0238	7.3234	7.4168	95691.2	<.0001
Resistance	R		1	1.1382	0.0269	1.0854	1.1909	1791.20	<.0001
Resistance	S		0	0.0000	0.0000	0.0000	0.0000	.	.
Crop	P		1	0.1579	0.0309	0.0974	0.2184	26.13	<.0001
Crop	R		0	0.0000	0.0000	0.0000	0.0000	.	.
MRB	D		1	-1.1042	0.0391	-1.1808	-1.0277	798.83	<.0001
MRB	N		0	0.0000	0.0000	0.0000	0.0000	.	.
Resistance*Crop	R	P	1	0.0913	0.0341	0.0245	0.1581	7.18	0.0074
Resistance*Crop	R	R	0	0.0000	0.0000	0.0000	0.0000	.	.
Resistance*Crop	S	P	0	0.0000	0.0000	0.0000	0.0000	.	.
Resistance*Crop	S	R	0	0.0000	0.0000	0.0000	0.0000	.	.
Resistance*MRB	R	D	1	-0.5055	0.0393	-0.5826	-0.4285	165.24	<.0001
Resistance*MRB	R	N	0	0.0000	0.0000	0.0000	0.0000	.	.
Resistance*MRB	S	D	0	0.0000	0.0000	0.0000	0.0000	.	.
Resistance*MRB	S	N	0	0.0000	0.0000	0.0000	0.0000	.	.
Crop*MRB	P	D	1	0.2158	0.0371	0.1432	0.2884	33.90	<.0001
Crop*MRB	P	N	0	0.0000	0.0000	0.0000	0.0000	.	.
Crop*MRB	R	D	0	0.0000	0.0000	0.0000	0.0000	.	.
Crop*MRB	R	N	0	0.0000	0.0000	0.0000	0.0000	.	.

Table 6.5: The odds ratios (and their confidence intervals) for the Louisiana versus Texas populations, Louisiana Resistant versus Susceptible sub-populations, and all the genotypes re-classified into resistant and susceptible populations and plant versus ratoon crop for the Mexican rice and sugarcane borers

Borer	$\hat{\theta}_{Louisiana\ vs\ Texas}$		$\hat{\theta}_{plant\ vs\ ratoon}$	
	Plant	Ratoon	Louisiana	Texas
Mexican rice borer	1.17 (1.11, 1.23)	0.65 (0.60, 0.71)	1.31 (1.23, 1.39)	1.01 (0.97, 1.06)
Sugarcane borer	0.90 (0.85, 0.96)	0.99 (0.88, 1.10)	1.00 (0.89, 1.13)	1.52 (1.39, 1.65)
Louisiana Resistant and Susceptible sub-populations $\hat{\theta}_{Resistant\ vs\ Susceptible}$				
Mexican rice borer	0.60 (0.56, 0.65)		1.24 (1.15, 1.33)	
Sugarcane borer	0.50 (0.44, 0.58)		1.33 (1.15, 1.54)	
All genotypes Re-classified into Resistant and Susceptible populations $\hat{\theta}_{Resistant\ vs\ Susceptible}$				
Mexican rice borer	0.68 (0.66, 0.71)		1.18 (1.08, 1.16)	
Sugarcane borer	0.39 (0.36, 0.42)		1.33 (1.24, 1.42)	

The sugarcane borer is believed to damage the plant crop more than the ratoon crops because the plant crop establishes more slowly than the ratoon crop and therefore remains in a vulnerable state longer. Also, the diversity and numbers of predators are lower in the plant-cane crop than in subsequent stubble crops. We therefore investigated the odds ratio for borer damage in the two populations between the plant and ratoon crops to determine if the same pattern of borer damage existed for the Mexican rice borer (Table 6.5). For the Louisiana population, the plant crop was significantly ($P < 0.05$; 31 %) more likely to be bored by the Mexican rice borer than the first ratoon crop. The Texas population was significantly ($P < 0.05$; 52 %) more likely to be bored by the sugarcane borer. Both the Louisiana and Texas populations were equally likely to be bored in the plant versus the ratoon crop by the sugarcane borer and Mexican rice borer,

respectively. From Table 6.5, it can be deduced that each population was more likely to be bored in the plant crop than the ratoon crop by the borer less prevalent in the area from which the population originated, that is, the Mexican rice borer for the Louisiana population and the sugarcane borer for the Texas population. As previously mentioned the similarity in damage by both borer species to the plant versus ratoon crops could indicate similar patterns of damage by both species.

Louisiana genotypes previously selected for resistance to the sugarcane borer were 40 % (significantly at $P < 0.01$) less likely to incur Mexican rice borer damage and 50 % less likely to incur sugarcane borer damage compared to susceptible genotypes (Table 6.5). When the analysis was repeated assuming prior knowledge of the sugarcane borer resistant and susceptible status of all 80 genotypes in the experiment, the results corroborated the above findings with resistant genotypes being 32 % less likely to incur Mexican rice borer damage and 61 % less likely to incur sugarcane borer damage than their susceptible counterparts (Table 6.5). The results corroborate previous evidence in suggesting that prior knowledge of the sugarcane borer resistance status of a plant could be useful as a predictive tool in determining how they would react when exposed to infestation by the Mexican rice borer. The dependency of Mexican rice borer-damaged internodes on the sugarcane borer resistance status of the plant provided evidence for cross resistance between the borer species. Therefore, mechanisms governing resistance to the sugarcane borer could also be active against the Mexican rice borer, although at marginally lower levels (Table 6.5). The plant crop was 24 % (Mexican rice borer) and 33 % (sugarcane borer) more likely to be bored than the ratoon crop; an indication that screening of both borers would be best carried out in the plant crop. That genotypes were more susceptible to damage by

both borers in the plant crop further highlights the notion that the patterns of damage for both borer species were similar, and perhaps one of the reasons for the cross resistance.

6.4 Discussion

The objective of this study was to determine whether cross resistance exist, among sugarcane genotypes, between the sugarcane borer and the Mexican rice borer. In particular, we wanted to know whether prior knowledge of the sugarcane borer resistance status of a genotype could be useful for predicting its reaction when exposed to the Mexican rice borer. The study was prompted as a measure of preparedness of the Louisiana sugarcane industry to the encroachment of the Mexican rice borer. The study sought to take advantage of resources developed through a long history of selection and breeding for sugarcane borer resistance in Louisiana (Hensley and Long, 1969; Kyle and Hensley, 1970; Pan and Hensley, 1973; White and Hensley, 1987; White, 1993; Milligan et al., 2003; Kimbeng et al., 2006). No formal program to select and breed for resistance to either pest existed in Texas. At the time the study was initiated, both the Mexican rice borer and the sugarcane borer were present in Texas whereas only the sugarcane borer was present in Louisiana. However, the Mexican rice borer was reported present in Louisiana by the time the study was concluded.

The study provided evidence of cross-resistance between the sugarcane borer and the Mexican rice borer. Evidence of a cross resistance between the sugarcane borer and the Mexican rice borer was more pronounced in the Louisiana population presumably because it had previously been selected for varying levels of resistance to the sugarcane borer (Figure 6.1; Table 6.2). This population was subdivided into two groups (resistant versus susceptible to the sugarcane borer) based on prior information. Using this information as a standard, it was also

possible to make a similar classification (resistant versus susceptible to the sugarcane borer) of all 80 entries in the trial. The analyses of these data showed that, % borer-damaged internodes for both the sugarcane borer and Mexican rice borer were substantially higher among the susceptible compared to the resistant group of genotypes (Tables 6.1 and 6.2). The resistant genotypes showed a strong positive association between the sugarcane borer and the Mexican rice borer (Figure 6.1 and Table 6.2) and were 32 % less likely to incur Mexican rice borer damage compared to their susceptible counterparts (Table 6.4).

Cross resistance would be a great benefit to the Louisiana sugarcane breeding program as it would eliminate the need for maintaining dual breeding programs needed to develop resistance to both species. The existence of cross resistance means the Louisiana sugarcane industry is not completely unprepared for the arrival of the Mexican rice borer. Genotypes identified as resistant to both pests in this study, especially Louisiana adapted germplasm, would form the base population of a program designed to elevate the level of resistance in sugarcane to both pests.

Most of the genotypes in the Louisiana population had previously been selected for their reaction to the sugarcane borer, whereas, none of the Texas genotypes had undergone any formal selection for either pests. Mean comparisons, from the ANOVA, of percent borer-damaged internodes for both pests, between the Louisiana and Texas population was probably not a reliable measure because the Louisiana samples deliberately included resistant and susceptible entries. However, the log linear model analysis showed that the Louisiana population was generally less likely to incur sugarcane borer damage compared to the Texas population. Log linear model analysis also showed that the Louisiana population responded more to fluctuations (within population comparison across crops) in Mexican rice borer pressure and less to fluctuations in sugarcane borer pressure while the Texas population succumbed more to

sugarcane borer pressure and less to Mexican rice borer pressure. The data seem to suggest that the Texas population has acquired at least a marginal to moderate level of resistance to the Mexican rice borer although no formal selection and breeding has been practiced for this trait. One can assume that native infestations of the Mexican rice borer in Texas are probably high enough to exert natural selection for this trait. It would, therefore, be possible to increase the level of Mexican rice borer resistance in sugarcane through active selection and breeding.

This study supports our contention that it is possible to simultaneously increase levels of stem borer resistance to both the Mexican rice borer and the sugarcane borer; unfortunately, those traits that possibly confer resistance to the stem borers are also inversely correlated with sucrose yields (White *et al*., 2006). We are currently reviewing advancement records to evaluate how valuable our sugarcane borer parental lines are to the commercial breeding programs in Louisiana. Our initial evaluations suggest that an additional cycle of backcrossing may be necessary to obtain sucrose yields required for a genotype to be accepted by growers. If borer resistance becomes diluted with subsequent backcrossing, then it may be necessary to identify other sources of resistance that are less correlated to low yields.

There could be negative side effects to increasing the cross resistance between the sugarcane borer and the Mexican rice borer. White *et al*. (2006) reported that the traits that confer resistance to the stem borers were negatively correlated with sucrose content. Increasing cross resistance could have the effect of significantly decreasing the sucrose content to uneconomic levels. Therefore information on the optimum levels of cross resistance that do not significantly reduce the values of other important traits like sucrose content could be used to reduce the negative side effects of high levels of cross resistance.

The choice of material to evaluate was an important consideration in confirming cross resistance in this study. A population comprising individuals with known levels of resistance to one of the borers is required in studying cross resistance. Little insight into the existence of cross resistance was gained when we analyzed the data from the Louisiana and Texas populations assuming all entries to be random with no defined resistance status for either one of the borer species. The use of a population that was defined for resistance to the sugarcane borer provided an experimental control for quantifying the existence and levels of cross resistance between the borer species.

Appropriate statistical methods are also important to investigate cross resistance. Borer damage by both species should depend on the same experimental design variables that are associated with the resistance. ANOVA can only identify significant experimental design variables but cannot reveal associations as well as test the dependency between borer damage and experimental design variables. Moreover, the ANOVA is not the most appropriate statistical method for analyzing categorical data that follows a Poisson distribution but it does help determine the important experimental design variables when the data follow a normal distribution as was the case in this study. While ANCOVA determines the strength of the association of the % borer-damaged internodes between species, and can suggest cross-resistance, it does not provide information of the design variables that are associated with the levels of borer damage. Log linear models test dependency of borer damage to experimental design variables. In this study, it was demonstrated that a combination of appropriate populations and statistical methods were required to determine the existence of cross resistance between the Mexican rice borer and sugarcane borer species.

This is only the first study to investigate cross resistance to pests in sugarcane. The study provides a good model for determining the ideal populations, experimental design and statistical methods for determining if cross resistance exists in other multi-pest agro-ecosystems. For example, in the Zimbabwe sugar industry, currently, the stem borer, *Eldana saccharina*, remains the major insect pest of sugarcane (Mazodze *et al*., 1999; Mutambara-Mabveni, 2007). Recently, another stem borer, *Chilo sacchariphagus* has been detected in Mozambique at the Mafambisse sugar estate (Conlong and Goebel, 2002) and is advancing from Mozambique to Zimbabwe, Malawi, Swaziland and South Africa sugarcane growing areas. Evidence of genotype resistance to *C. sacchariphagus* has been observed in Mozambique (Conlong *et al*., 2004). Currently, there is active selection for eldana resistance (Rutherford, 1998) but no active selection for *C. sacchariphagus*. Investigating cross resistance between the *E. saccharina* and *C. sacchariphagus* would help the Southern Africa sugarcane industries develop strategies to control both borer species.

6.5 Summary

There was a significant association between the Mexican rice borer damage and the sugarcane borer damage for the Louisiana population and the sugarcane borer resistant sub-population, suggesting the probable existence of cross resistance. Sugarcane borer resistant genotypes were significantly less bored by the Mexican rice borer than susceptible genotypes indicating that resistance to the sugarcane borer also imparted resistance to the Mexican rice borer. The analysis of covariance showed strong association in % bored internodes between the Mexican rice and sugarcane borers. Log linear models support the existence of cross resistance between the sugarcane borer and the Mexican rice borer. However, sugarcane borer resistance conferred marginally lower resistance to the Mexican rice borer. The plant crop was more likely to be

bored by both the borers because the plant crop was slower to establish and also indicating that screening for both borers could be best done in the plant crop. The approach followed in this study can be applied to breeding for pest resistance in other sugarcane industries and crop species with similar pest problems. These findings demonstrate that those traits conferring resistance to the sugarcane borer (i.e. fiber, rind-hardness, tight leaf sheaths) are likely conferring resistance to the Mexican rice borer. However, there was still a marginally lower resistance to the Mexican rice borer. If increasing levels of resistance to the Mexican rice borer are required to successful manage this new stem borer, the existing recurrent selection for sugarcane borer would provide germplasm for starting recurrent selection for the Mexican rice borer, but ultimately direct selection for the Mexican rice borer may become necessary. The marginal resistance to the Mexican rice borer among the genotypes from Texas indicates that natural selection was working in this population. Active screening for the Mexican rice is likely to significantly increase the resistance to the Mexican rice borer.

6.6 References

Abraham, B. and Ledolter, J. (2006). Introduction to Regression Modelling. Thomson Brooks/Cole, USA.

Agresti, A. (2007). An Introduction to Categorical Data Analysis. Second Edition Wiley-Inter-Science. New Jersey, USA.

Allison, P.D. (2003). Logistic Regression Using The SAS System : Theory and Applications. Fourth Edition. Cary, North Carolina: SAS Institute Inc.

Conlong, D.E. and Goebel, R. (2002). Biological control of *Chilo sacchariphagus* (Lepidoptera: Crambidae) in Mozambique: The firsts steps. *Proceedings of the South African Sugar Technologists' Association*, 76 : 310 – 320.

Conlong, D.E., Sweet, P. and Piwalo, J. (2004). Resistance of Southern African varieties of sugarcane to Chilo sacchhariphagus (Pyralidae: Crambidae) in Mozambique, and development of a non-destructive resistance rating system. *Proceedings of the South African Sugar Technologists' Association*, 78: 317 – 328.

Hensley, S.D. and Long, W.H. (1969). Differential response of commercial sugarcane varieties to sugarcane borer damage. *Journal of Economic Entomology* 62: 620 – 622.

Johnson, H.J.R. (1984). Identification of *Eoreuma loftini* (Dyar) (Lepidoptera: Pyralidae) in Texas, 1980: forerunner for other sugarcane boring pest immigrants from Mexico? *Bulletin of the American Society of America* 30: 47 – 52.

Johnson, K.J.R. and van Leerdam, M.B. (1981). Range extension of *Acigona loftini* into the Lower Rio Grande Valley of Texas. *Sugar y Azucar* 76: 34.

Kimbeng, C.A., White, W.H., Miller, J.D. and Legendre, B.L. (2006). Sugarcane resistance to the sugarcane borer: Response to infestation among progeny derived from resistant and susceptible parents. *Sugarcane International* 24(3): 14 – 21.

Kyle, M.L. and Hensley, S.D. (1970). Sugarcane borer host resistance studies. *Proceedings of the Louisiana Academy of Sciences* 33: 55 – 67.

Littell, R.C., Stroup, W.W. and Freund, R.J. (2002). SAS for Linear Models. Fourth Edition. SAS Institute Inc., Cary, North Carolina, USA.

Littell, R.C., Milliken, G.A., Stroup, W.W. and Wolfinger, R.D. (2005). SAS System for Mixed Models. Seventh edition. SAS Institute Inc., Cary, North Carolina, USA.

Mazodze, R., Nyanhete, C. and Chidoma, S. (1999). First outbreak of *Eldana saccharina* (Lepidoptera: Pyralidae) in sugarcane in the South-East Lowveld of Zimbabwe. *Proceedings of the South African Sugar Technologists' Association* 73: 107 – 111.

Meagher, R.L., Jr., Smith, J.W., Jr. and Johnson, K.J.R. (1994). Insecticidal management of *Eoreuma loftini* (Lepidoptera: Pyralidae) on Texas sugarcane: A critical review. *Journal of Economic Entomology*, Volume 87: 1332 – 1344.

Milligan, S.B., Balzarini, M. and White, W.H. (2003). Broad sense heritabilities, genetic correlations, and selection indices for sugarcane borer resistance and their relation to yield loss. *Crop Science* 43: 1729 – 1735.

Mutambara-Mabveni, A.R.S. (2007). *Eldana Saccharina* Walker (*Lepidoptera: Pyralidae*) in sugarcane: impact and implications for the Zimbabwe sugar industry. *Proceedings of the International Society of Sugar Cane Technologists* 26: 770 – 779.

Munroe, E. and M. A. Solis. (1999). The Pyraloidea, pp. 233 – 256. *In* N. P. Kristensen (editor), Handbuchder Zoologie, Band IV, Arthropoda: Insecta Teilband 35, Lepidoptera, moths and butterfies. Vol. 1: Evolution, Systematics, and Biogeography. Walter de Gruyter, Berlin, Germany.

Pan, Y.S. and Hensley, S.D. (1973). Evaluation of sugarcane seedlings for resistance to the sugarcane borer, *Diatraea saccharalis*. *Environmental Entomology* 2: 149 – 154.

Reay-Jones, F.P.F., Wilson, L.T., Reagan, T.E., Legendre, B.L. and Way, M.O. (2007). Predicting economic losses from the continued spread of the Mexican rice borer (*Lepidoptera: Crambidae*). *Journal of Economic Entomology*, Volume 101 number 2: 237 – 250.

Ring, D.R., Browning, H.W., Johnson, K.J.R., Smith, Jr., J.W., and Gates, C.E. (1991). Age-specific susceptibility of sugarcane internodes to attack by the Mexican rice borer (Lepidoptera: Pyralidae). *Journal of Economic Entomology*, Volume 84 number 3: 1001 – 1009.

Rutherford, R.S. (1998). Prediction of resistance in sugarcane to stalk borer *Eldana saccharina* by near infra-red spectroscopy on crude bud scale extracts: Involvement of chlorogenates and flavonoids. *Journal of Chemical Ecology* 24: 1147 – 1463.

SAS Institute (2007). The SAS System for Windows Version 9.1.3. SAS Institute, Cary, North Carolina, USA.

Smith, J.W., Jr., Wiedenmann, R.N. and Overholt, W.A. (1993). Parasites of lepidopteran stemborers of tropical gramineous plants. ICIPE Science Press, Nairobi, Kenya, 89 pages.

Stokes, M.E., Davis, C.S. and Koch, G.G. (1996). Categorical Data Analysis Using the SAS System. Second Edition. Cary, North Carolina: SAS Institute Inc.

White, W. H. 1993. Cluster analysis for assessing sugarcane borer resistance in sugarcane line trials. *Field Crops Res*. 33: 159 – 168.

White, W.H. and Hensley, S.D. (1987). Techniques to quantify the effect of *Diatraea saccharalis* (Lepidoptera: Pyralidae) on sugarcane quality. *Field Crops Research* 15: 341 – 348.

White, W.H., Legendre, B.L., and Miller, J.D. (1996). Progress in breeding for sugarcane borer resistance. *Sugar Cane* 5: 3 – 7.

White, W.H., Tew, T.L., and Richard, E.P., Jr. (2006). Association of sugarcane pith, rind hardness, and fiber with resistance to the sugarcane borer. *Journal American Society of Sugar Cane Technologist* 26: 87 – 101.

CHAPTER 7: GENERAL DISCUSSIONS, CONCLUSIONS AND PROSPECTS FOR FUTURE RESEARCH

7.1 Early Generation Selection Stages

To date, the challenges of the early generation selection have emphasized identifying families with greater proportions of high yielding seedlings as well as identifying the seedlings with potential to produce high cane yield from these families. Although current family evaluation methods using family means have increased gains in early generation selection, they remain inadequate. Current family evaluation methods are inadequate because they do not provide insight into the distribution patterns of the seedling cane yield within the families. The distribution patterns of seedling cane yield within families could be a key parameter that could help identify families with higher yield clones as well as identify those families where high cane yield seedlings can be easily identified during individual seedling selection. Visual seedling appraisal has produced gains in seedling selection and is also very easy to implement. However, the confounding influence of genotype by environment interaction effects and the competition among closely planted seedlings can significantly reduce the precision of visual selection. Statistical models with the capability of performing fast computations can provide decision support tools that are more objective for individual seedling selection. Such methods could also provide more flexible parameters for reviewing the selection process as well as evaluating the sugarcane breeding populations.

7.1.1 Family Evaluation and Selection

7.1.1.1 General Discussions

This study showed that accounting for the confounding influence on cane yield between the seedlings and the clones substantially improved the evaluation of families. The families

classified as elite using RCM analysis generally produced higher cane yield in both the seedlings and clones than those derived from family means. Seedling selection for high cane yield from these elite families (Stage I) produced clones with high cane yield (Stage II). The use of the intercept and slopes helped evaluate the families for yield potential and repeatability. Repeatability, as described by the slope was used to evaluate the distribution pattern of cane yield of the seedlings and clones. The elite families selected by RCM analysis produced larger within family variance among the seedlings and clones, making these families ideal populations for selection. The RCM analysis also produced the most discriminating parameter, the slope. The disadvantage associated with the implementation of the RCM analysis would be the required cane yield data from the seedlings, and first clonal stage. This data is not routinely collected in most breeding programs because of the high labor cost and availability. However, in breeding programs capable of collecting the data, parallel and retrospective family evaluation can be implemented. With the parallel family evaluation approach, fewer seedlings per family would be planted as seedlings and clones, and would be used to evaluate the families. Large numbers of seedlings for individual selection would only be planted from the elite families. The savings in planting fewer families for individual seedling selection could compensate for the cost associated with the data collection required for RCM family evaluation. The retrospective approach would use data collected in seedlings and first stage clones to evaluate families after seedling selection has already been done. However, the analysis would be used to determine the families to replant.

7.1.1.2 General Conclusions

In conclusion, the RCM analysis offers potential for family evaluation because of its ability to simultaneous screen families for yield potential and repeatability, thereby comparing the families

for their distribution patterns of cane yield of the seedlings and clones within the families. The elite families that were selected using RCM analysis comprised a greater proportion of seedlings that produced high cane yield. These elite families also showed better distribution patterns where the cane yield of the clones increased consistently with the increases in the cane yield of the seedlings. Because fewer high quality families will be planted for individual seedling selection, these seedlings are likely to be better managed and intensive selection can be practiced because of the fewer numbers involved than is the case with current methods. Parallel family evaluation appears more attractive as a strategy for implementing RCM analysis because the resources saved by planting fewer families for individual seedling selection are likely to compensate for the extra cost associated with data collection required to implement the RCM analysis.

7.1.2 Seedling Selection

7.1.2.1 General Discussions

Logistic regression offered great potential as an objective decision support statistical tool for seedling selection. The logistic regression model identified higher yielding seedlings and was more flexible in adjusting the number of seedlings selected than visual selection. The probability was the parameter used for selection. Statistical tests generated for the probability using confidence limits provided an extra aid to seedling selection. Trait relationships among families and between populations were evaluated from the trends of the probability plotted against the trait values, providing the breeder with a mechanism to quantify the effects of a selection strategy over time. The logistic models provided a statistical tool for shifting breeding populations towards, for example, high cane yield by using populations created with the desired trait combination that impart high cane yield as sources of the training data set. The cost of collecting the required seedling data would be a disadvantage to the wide adoption of the logistic

regression models as a decision support tool but this disadvantage could be overcome by calibrating models that could use scores derived from yield components.

The logistic regression models can be implemented using the SAS Enterprise Miner Artificial Neural Network (ANN) models. Both approaches produce similar results and interpretations. When using the ANN models, the user specifies the training and prediction data set and the selection threshold probability. The SAS code is not required because the SAS Enterprise Miner has the code behind the scenes. The output of the ANN models provides the probability of selection and the selection decision (elect or reject). When similar training data sets and selection threshold probability are used for both models, the outputted results would be identical for both methods and the selection decisions would also be identical. Because the ANN models do not require a SAS code, they therefore provide faster computations because the user does not need to spend time writing the code. Furthermore, the user need not have any knowledge of SAS programming to implement the analysis. The decision to select or reject that is automatically produced in the output of the ANN models reduces the need to make the decisions manually based on the threshold as is done with the logistic regression models, further saving the user more time.

7.1.2.2 General Conclusions

The logistic regression models provided a quick and objective decision support tool for objective individual seedling selection. The effect of genotype by environment interaction that would significantly confound visual selection is reduced by logistic regression models. The numbers of seedlings to advance to clonal stages can be easily and quickly adjusted using the probability of selection, saving the breeder a lot of time and also providing a more objective parameter for

refining the selection process. Gains achieved by using the RCM analysis for family evaluation could be further enhanced by incorporating logistic regression models for seedling selection. The interrelationships among the yield components in breeding populations can also be inferred from plots of logistic regression models output data. The logistic regression model analysis can be automated in SAS by using the Artificial Neural Network models, thereby increasing the speed of the analysis.

7.2 Multivariate Repeated Measures Analysis of Data From Advanced Variety Trials

7.2.1 General Discussions

The multivariate repeated measures analysis produced better model fit and greater discrimination between the differences in means of the yield traits of experimental genotypes and the control than the univariate analysis. The multivariate repeated measures analysis produced more appropriate experimental errors by including the covariance between variables and between crop-years in computing the variances used for the tests. The univariate analysis assumptions of a split plot in time as its experimental design and the independence between variables and between crop-years resulted in the univariate analysis significantly underestimating the standard errors used for testing the differences among genotypes for yield traits, causing incorrect interpretations. Yield traits are generally controlled by quantitative genes and therefore are more susceptible to genotype by environment interaction effects (Falconer and Mackay, 1996; Kimbeng *et al*., 2009; Mirzawan *et al*., 1993). The effects of genotype by environment interaction could be reduced by using the multivariate repeated measures approach. The yield plateau alluded to by Garside *et al*. (1997) could partially have been caused by some released varieties being erroneously defined as statistically higher yielding than the control because of the

Type I error associated with the univariate analysis. More appropriate analysis such as the multivariate repeated measures could result in correct interpretations of the results.

7.2.2 General Conclusions

The multivariate repeated measures analysis showed increased separation of the differences in means of the yield traits between the experimental genotypes and the control. However, the multivariate repeated measures analysis produced similar inferences to the univariate analysis for quality traits. Traits more influenced by genotype by environment interaction effects, for example, yield traits, would benefit immensely from the multivariate repeated measures analysis than those that are less influenced by genotype by environment interaction such as quality traits. The gains achieved from early generation selection using RCM analysis and logistic regression models could be further enhanced by adopting the multivariate repeated measures analysis in the advanced variety trials of sugarcane breeding programs.

7.3 Cross Resistance Between the Sugarcane Borer and the Mexican rice borer

7.3.1 General Discussions

This study proved that cross resistance existed between the sugarcane borer and the Mexican rice borer. The existence of cross-resistance implies that breeding for resistance to one of the borer species would achieve control for both, significantly lowering the costs associated with running parallel resistance breeding programs. The benefits of the recurrent selection for the sugarcane borer resistance can be used to enhance the genetic control of the Mexican rice borer. The type of populations and statistical methods were proved to be important in demonstrating the existence of cross resistance. Populations with known resistance status to one of the species were important in this study. The information of the resistance status to the sugarcane borer provided a

control variable that was used to determine the existence of cross resistance, thereby simplifying the study. While ANCOVA showed the associations, and ANOVA only identified significant experimental design variables, these statistical methods only provided indications of the possible existence of cross-resistance but were inadequate for proving concrete evidence of the existence of cross resistance. The log linear model analysis was able to prove cross resistance by identifying the variables that determined the levels of borer-damage between the populations. This study will provide a reference to other researchers facing similar dual pest problems.

7.3.2 General Summary

This study demonstrated that selecting for resistance to the sugarcane borer would produce resistance to the Mexican rice borer, that is, cross-resistance. This finding would benefit the Louisiana sugarcane industry by limiting the yield losses from the Mexican rice borer that was recently observed in Louisiana. The type of populations and the statistical methods used were important in demonstrating the existence of cross resistance between the borer species. The recurrent selection for sugarcane borer resistance in Louisiana should provide parent genotypes for initiating resistance breeding to the Mexican rice borer. This study would also provide a good reference for the experimental design and statistical procedures to other sugarcane industries that are facing similar multiple pest problems.

7.4 Prospects and Recommendations for Future Research

7.4.1 Early Generation Selection

Future research should focus on determining the optimum number of seedlings and or clones that should make up each family for RCM evaluation. Simulation studies can be used for optimizing the family population sizes and the optimum family selection rates. Previous studies have shown

the influence of genotype by environment interactions on family selection using family means (Jackson *et al*., 1995a, b). There is also a need to investigate the fluctuation in the family RCM parameters (intercept and slope) across environments. Such a study would provide more information on the stability of the RCM analysis parameters.

Using logistic regression models for seedling selection requires a training data set. The training data set determines the parameters that are used to build the models used to compute the probability of selection. Further studies, probably using simulation models, is required to determine the optimum population size for the training data set. Because the logistic regression models offer an opportunity for shifting the population using the training data set, there is an opportunity to investigate the potential and magnitudes of these shifts by changing the parameters of the training data set. Such studies can also be done using simulation models. Previous studies have reported significant influence of genotype by environment interaction effects on seedling selection (Bull *et al*., 1992; Jackson and McRae, 1998). There is a need to evaluate if the selection rates within families would change across different environments when the logistic regression model is used.

Further research should also evaluate the gains that would be achieved by combining RCM analysis and logistic regression models in the early generation selection stages. By evaluating the gains in yield across the stages of a selection program using RCM analysis and family means, the magnitude of the potential gains from RCM analysis over family analysis would be quantified for breeding programs. Alternatively, experiments designed specifically to evaluate these gains would be ideal. These studies would encourage the wide adoption of these models even in programs that may deem data collection expensive particularly if the gains would compensate for the cost of the data collection.

7.4.2 Multivariate Repeated Measures Analysis

The likely erroneous interpretations of results using univariate analysis need to be investigated using historical data. Data from advanced variety trials is plentiful among sugarcane breeding programs. Studies using this historical data should shed more light on the potential impact and benefits of using multivariate repeated measures in analyzing advanced variety trials data. In addition to providing information on the potential gains likely to be derived from using multivariate repeated measures analysis, cultivars already released may have to be redefined if studies on historical data conclude their yield advantages have been overstated particularly when the varieties were described erroneously as significantly higher yielding than the controls. Some rejected varieties discarded as inferior to the controls may be found similar to the controls, and if these varieties offer certain added advantages such as superior disease or pest resistance, these varieties could be useful as parents for future crossing.

Modelling Type I and Type II errors for both the univariate and the multivariate repeated measures analysis would shed more insight into the benefits of using the multivariate repeated measures analysis of yield traits in advanced variety trials. Such a study could also be used to determine experimental designs that would enhance the reduction of Type I errors and therefore enhance the effectiveness of the sugarcane breeding programs.

The gains achieved in early generation selection could be further enhanced by adopting the multivariate repeated measures analysis for the advanced variety testing stages. There is need for research to evaluate the gains from the use of a combination of RCM analysis, logistic regression models and multivariate repeated measures analysis in a breeding program. Following every series from Stage I to Stage V, for example in the Zimbabwe program, would provide data to evaluate these gains. Additionally, the gains can also be evaluated using simulation models for

the current system and the new approach that use RCM analysis, logistic regression and multivariate repeated measures analysis. Such studies would further enhance the value of these statistical methods.

While these methods have been demonstrated using SAS procedures, other software such as GENSTAT, R, SPSS and others that are widely used by plant breeders should be able to perform these analysis adequately. However, it is important to get the analysis evaluated with all the software that breeders are more familiar with and use frequently. The evaluation of the efficiency of the other statistical software may indicate the necessary improvements that may need to be incorporated in the deficient software. Additionally, such studies could also be used to identify the best software for performing these analyses, further providing the breeders with a wider choice.

While these studies used sugarcane as a case study crop, we can speculate that these statistical methods are likely to be applicable to most of the other perennial crops with similar growing patterns to sugarcane. Obvious examples will include perennial forage grasses, among others. Studies to evaluate their application in these situations would be valuable.

7.4.3 Cross Resistance

Further studies are needed to investigate, between the sugarcane borer and the Mexican rice borer, the species that would be easier to screen and then in future develop resistance for that species. Also, there is need to investigate which species, after selecting for resistance, would result in higher borer resistance levels for both species. Alternatively, the species more prevalent in an area would be selected for as this would also provide resistance for the less prevalent species.

Future studies could also quantify the strength of the cross resistance. It is necessary to investigate if the cross resistance is stable across environments and crop-years. To enhance the incorporation of the cross resistance into breeding populations, it will be important to evaluate those genotypes that show the greatest cross resistance but are deficient in important agronomic traits. More research is needed to quantify the agronomic traits of the progenies of crosses done between the high cross-resistant genotypes and the agronomically acceptable parental genotypes, and verify if recombinants with both high cross-resistance and acceptable agronomic qualities occur in large enough proportions. Such a study could also quantify the heritability of cross resistance.

Studies to determine the optimum selection rates that would allow the advancement of cross resistant genotypes for further use in the crossing program as parents is also needed. Such studies can be done using statistical simulation models. This approach would eventually have the effect of increasing agronomic value as well as capturing the borer cross resistance. This approach would act as a form of simultaneous recurrent selection for both cross resistance and agronomic traits.

The negative side effects of incorporating cross resistance between the Mexican rice borer and the sugarcane borer would need to be investigated further. White *et al*. (2006) reported that the traits that confer resistance to the stem borers were negatively correlated with sucrose content. Investigations that quantify the threshold levels of cross resistance that is unlikely to negatively impact sucrose content would provide guidance to plant breeders on the levels of backcrossing required to recover economically acceptable genotypes with acceptable borer cross resistance.

Southern Africa sugarcane growers encounter the same predicament presented here for the Louisiana and Texas industries. Two borers, *Eldana saccharina* and *Chilo sacchariphagus* have been reported to coexist in Mozambique. *Eldana saccharina* is currently known to be widely distributed throughout Southern Africa while *C. sacchariphagus* is reportedly spreading fast from Mozambique into the neighboring countries. Studies to determine the potential existence of cross resistance between the sugarcane borers would help prepare Southern Africa sugarcane growers to mitigate the impact brought by the two devastating sugarcane borers. Such studies would also determine if parallel resistance breeding programs are required.

7.5 References

Bull, J.K., Hogarth, D.M. and Basford, K.E. (1992). Impact of genotype by environment interaction on response to selection in sugarcane. *Australian Journal of Experimental Agriculture* 32: 731 – 737.

Falconer, D.S. and Mackay, T.F.C. (1996). Introduction to Quantitative Genetics. Fourth Edition. Longman Group Ltd, UK.

Garside, A.L., Smith, M.A., Chapman, L.S., Hurney, A.P. and Magarey, R.C. (1997). The yield plateau in the Australian Sugar industry: 1970 – 1990. In Keating, B.A. and Wilson, J.R. (editors). Intensive Sugarcane Production: Meeting the Challenges Beyond 2000. CAB International, Wallingford, United Kingdom: 103 – 124.

Jackson, P.A. and McRae, T.A. (1998). Gains from selection of broadly adapted and specifically adapted sugarcane families. *Field Crops Research* 59: 151 – 162.

Jackson, P.A., McRae, T. and Hogarth, M. (1995a). Selection of sugarcane families across variable environments I. Sources of variation and an optimal selection index. *Field Crops Research* 43: 109 – 118.

Jackson, P.A., McRae, T. and Hogarth, M. (1995b). Selection of sugarcane families across variable environments I. Patterns of response and association with environmental factors. *Field Crops Research* 43: 119 – 130.

Kimbeng, C.A., Zhou, M.M. and da Silva, J.A. (2009). Genotype by environment interactions and resource allocation in sugarcane yield trials in the Rio Grande valley region of Texas. *Journal of the American Society of Sugar Cane Technologists* (In press).

Mirzawan, P.D.N., Cooper, M. and Hogarth, D.M. (1993). The impact of genotype by environment interactions for sugar yield on the use of indirect selection in southern Queensland. *Australian Journal of Experimental Agriculture* 33: 629 – 638.

White, W.H., Tew, T.L., and Richard, E.P., Jr. 2006. Association of sugarcane pith, rind hardness, and fiber with resistance to the sugarcane borer. *Journal American Society of Sugar Cane Technologist* 26: 87 – 101.

APPENDIX 1 RANDOM COEFFICIENT MODELS ANALYSIS CODE

```
Proc MIXED data=one scoring=8 COVTEST Method=REML;
Class Family;
Model CY = SY/solution;
Random intercept SY/TYPE=UN subject=Family solution;
run;
```

The COVTEST Method=REML option provided the covariance test of the random effects using the Restricted Maximum Likelihood (REML). The TYPE=UN in the RANDOM statement provided the estimates of the variances and covariance of the slopes and intercepts and models an UNSTRUCTURED 2x2 covariance matrix for the random intercept and slope. The option SUBJECT=FAMILY in the RANDOM statement specified that the intercept and slope of each family was independently distributed from the intercepts and slopes of the other families, and the intercept and slope within each family were correlated and were random. The SOLUTION option on the MODEL statement provides a test for the population intercept and slope. The SOLUTION option on the RANDOM statement produces the tests for the effects of the intercept and slope (differences of the family intercept and slope from the population intercept and slope, respectively) of each family separately using the covariance matrix generated. The option SCORING=8 requested the PROC MIXED procedure to use Fisher's scoring method with eight iterations.

APPENDIX 2 LOGISTIC REGRESSION MODELS ANALYSIS CODE

```
Proc Logistic data=one Descending covout outest=Houma;
Model Response = Stalks Height Diameter;
Output out=predict p=Estimate lower=LCL upper=UCL;
run;
```

Where the DESCENDING option allows the modelling of the probability of selecting a seedling (1), COVOUT option produces the covariance matrix of the intercept and coefficients of the independent variables, the OUTEST = HOUMA option outputs the covariance matrix, OUT = PREDICT option outputs the predicted values, and P = ESTIMATE, LOWER = LCL, and UPPER = UCL option renames the predicted probability, the lower and upper confidence intervals for the predicted probability.

APPENDIX 3 MULTIVARIATE REPEATED MEASURES CODE FOR UN@UN

```
Proc mixed data=sugar;

Class Location Genotype CropYear Rep RV;

Model Y = RV Location(RV) Genotype(RV) CropYear(RV) Location*Genotype(RV)
Location*CropYear(RV) Genotype*CropYear(RV) Location*Genotype*CropYear(RV) / noint;

Random RV / Subject=Rep(Loc) type=UN;

Repeated RV CropYear / Subject=Genotype*Rep(Location) type=UN@UN;

Run;
```

The multivariate covariance structure was determined by the PROC MIXED random statement, RANDOM RV/SUBJECT=Rep(Loc) TYPE=UN. The PROC MIXED statement, REPEATED RV CropYear/SUBJECT=Genotype*Rep(Location) TYPE=UN@UN, determined the repeated measures covariance structure. The option TYPE = UN@UN or UN@AR(1) or UN@CS specified the covariance structure. The UN@UN modeled different covariance for the multivariate effects (UN) and different covariance for the repeated measures, while UN@AR(1) modeled a decay in covariance over time for the repeated measures and UN@CS modeled equal covariance for the repeated measures. For the yield, quality, and agronomic traits, each with two variables and eight crop-years, the UN@UN covariance had a 2x2 matrix for the multivariate effects (UN) and an 8x8 matrix for the repeated measures (UN) giving a 16x16 matrix for each of the covariance structures of UN@UN, UN@CS, and UN@AR(1).

APPENDIX 4 MULTIVARIATE REPEATED MEASURES CODE FOR UN@CS

```
Proc mixed data=sugar;

Class RV Location Genotype CropYear Rep;

Model Y = RV Location(RV) Genotype(RV) CropYear(RV) Location*Genotype(RV)
Location*CropYear(RV) Genotype*CropYear(RV) Location*Genotype*CropYear(RV) /noint;

Random RV / Subject=Rep(Location) type=UN;

Repeated RV CropYear / Subject=Genotype*Rep(Location) type=UN@CS;

Lsmeans Gen/pdiff adjust=Dunnett diff=control("16");

Run;
```

Where the statement LSMEANS GEN/PDIFF ADJUST=DUNNETT DIFF=CONTROL("16")

generates the comparison of each experimental genotype to the control, genotype 16.

APPENDIX 5 MULTIVARIATE REPEATED MEASURES CODE FOR UN@AR(1)

```
Proc mixed data=sugar;

Class RV Location Genotype CropYear Rep;

Model Y = RV Location(RV) Genotype(RV) CropYear(RV) Location*Genotype(RV)
Location*CropYear(RV) Genotype*CropYear(RV) Location*Genotype*CropYear(RV) /noint;

Random RV / Subject=Rep(Location) type=UN;

Repeated RV CropYear / Subject=Genotype*Rep(Location) type=UN@AR(1);

Run;
```

APPENDIX 6 LOG LINEAR MODELS ANALYSIS CODE

```
proc Genmod data=resistance;

class Resistance Crop MRB;

model Count = Resistance Crop MRB Resistance*Crop Resistance*MRB Crop*MRB /
Dist=POI Link=Log Obstats Residuals type3;

Title3 'Log Linear Mexican Rice Borer (Resistance/Susceptible Sub-Groups)';

run;
```

where DIST=POI option refers to the data following the Poisson distribution. LINK=LOG option refers to the log transformation that is used to linearise the data. OBSTATS RESIDUALS option outputs the observation statistics and the residuals. TYPE3 option produces the likelihood ration type 3 tests.

VITA

Marvellous Zhou was born in the Mberengwa district of the Midlands province of Zimbabwe. He attended Mpandashango Primary School, Chegato High School and studied for a Bachelor of Science Agriculture (Crop Science) degree at the University of Zimbabwe where he majored in plant breeding with a minor in biometry. He joined the Cotton Research Institute as a plant breeder for three and half years before accepting a similar post with the Zimbabwe Sugar Association Experiment Station (ZSAES). He was granted study leave by ZSAES to pursue a Master of Science in Agriculture at the University of Natal and later a Doctor of Philosophy at Louisiana State University. At LSU he also pursued and completed a Master of Applied Statistics in October 2008.

During high school, Marvellous was a recipient of the Lutheran Church Scholarship that financed his Advanced Level studies after passing the Cambridge Ordinary Level Examination as the best student among Lutheran schools. He participated in the Old Mutual Mathematics Olympiad (promoted excellence in Mathematics in high schools) and was the best student in the Midlands province. Marvellous was the best Cambridge Advanced Level student in the Midlands and as a result was awarded a Merit Scholarships by Ministry of Higher Education and Ministry of Agriculture, Zimbabwe. At the University of Natal, he was awarded two Certificates of Merit for being the best student in his course work. Marvellous is a member of the Crop Science Society of America, Sigma Xi honor Society, International Society of Sugarcane Technologists, South African Society of Sugarcane Technologists and Southern African Plant Breeders Association.