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

Design and analysis of sugarcane breeding experiments: a case study

Alessandra dos Santos

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

**Piracicaba
2017**

Alessandra dos Santos
Degree in Mathematics

Design and analysis of sugarcane breeding experiments: a case study

Advisor:

Profa. Dra. **CLARICE GARCIA BORGES DEMÉTRIO**

Thesis presented to obtain the degree of Doctor in Science.

Area: Statistics and Agricultural Experimentation

Piracicaba
2017

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

Santos, Alessandra dos

Design and analysis of sugarcane breeding experiments: a case study /
Alessandra dos Santos. – – Piracicaba, 2017 .
174 p.

Tese (Doutorado) – – USP / Escola Superior de Agricultura “Luiz de
Queiroz”.

1. Modelos mistos 2. Correlação autorregressiva 3. Correlação em banda
4. Tombamento 5. Delineamento não replicado 6. Delineamento ótimo 7.
Delineamento parcialmente replicado 8. Ganho genético 9. Assertividade na
seleção 10. Estudo de simulação I. Título.

ACKNOWLEDGMENTS

I thank God for all the moments, blessings and opportunities.

I thank my family and friends for all their love and care. I could not finish this work without their efforts to help me and keep my life happy, feeling close every time which it was not possible even when we were far apart. Thank you Leonardo, Aristides, Suely, Andressa, Aristides Neto, Bruna, Lourdes, Domingos, Renata, Roberta, Luís, Alyne, Priscila, Catherine, Pamela and Omar.

I thank my country, Brazil, for investing in education and developing good projects to improve more. I am proud to be Brazilian and have the opportunity to study in best universities in my country and work with the best researchers. I thank the Brazilian society and I hope to be able to contribute and return all the investments.

I am grateful to Dr. Clarice Garcia Borges Demétrio for accepting to be my supervisor and giving me incredible opportunities. I thank you very much for trusting me. Also, I am very grateful to work with Dr. Chris Brien. He is a genius and in one year in Australia I can say that my life changed a lot and I became better as a professional and a person. Chris, I thank you for all moments, all talking, all new ideas, all effort that you made to help me. You and your wife, Margaret, are incredible people and I thank you very much. Dr. Renata Alcarde Sermarini, I also would like to thank you very much, our talking and your points were very important to me.

Teachers and professors from primary school, high school, universities, thank you very much. I have only reached this stage because of your efforts. In particular I thank Dr. Vanderli Marino Melem, Londrina State University; because of her I decided to study statistics. Thanks for being an excellent professional and person.

For the financial support for this work, I thank CNPq, National Council for Scientific and Technological Development (38 months) and CAPES, Coordination for the Improvement of Higher Education Personnel (9 months). Lastly, I thank the University of São Paulo and the Centro de Tecnologia Canavieira for supporting the project through their partnership.

TABLE OF CONTENTS

Resumo	5
Abstract	6
List of Figures	7
List of Tables	12
1 Introduction	21
1.1 CTC company	22
1.2 Outline of thesis	23
2 Individual and joint analysis of sugarcane experiments to select test lines	25
2.1 Introduction	25
2.2 Material	27
2.3 Methodology	28
2.4 Data analysis results	30
2.5 Simulation studies	33
2.5.1 Single site experiments	33
2.5.2 Separate versus joint analyses	34
2.6 Simulation results	35
2.6.1 Single site experiments	35
2.6.2 Separate versus joint analyses	41
2.7 Conclusions	46
3 Negative residual correlation in sugarcane experiments	49
3.1 Introduction	49
3.2 Motivating example	50
3.3 Material and methods	50
3.4 Results	51
3.5 Conclusions	52
4 Comparison of a systematic design with some spatially optimized designs	53
4.1 Introduction	53
4.1.1 Optimality criteria for design searches	54
4.1.2 Comparing different design types using simulation	54
4.2 Motivating example	55
4.3 Material and methodology	56
4.4 Results	60
4.5 Conclusions	62
5 General conclusions and future works	65
References	67
Appendix	70

RESUMO

Delineamento e análise de experimentos de melhoramento com cana de açúcar: um estudo de caso

Um dos propósitos dos programas de melhoramento genético é a seleção de novos clones melhores (novos materiais). A acurácia de seleção pode ser melhorada usando delineamentos ótimos e modelos bem ajustados. Porém, descobrir isso não é fácil, especialmente, em experimentos grandes que possuem mais de cem clones sem a possibilidade de repetição devido à limitação de material, área e custos elevados, dadas as poucas repetições de parcelas com variedades comerciais (testemunhas) e o número de parâmetros de complexa variância estrutural que necessitam ser assumidos. Os principais objetivos desta tese foram modelar 21 experimentos de cana de açúcar fornecidos pelo Centro de Tecnologia Canavieira (CTC - empresa brasileira de cana de açúcar) e avaliar o delineamento empregado, o qual usa um número grande de clones não repetidos e testemunhas sistematicamente repetidas. O modelo linear misto foi usado, identificando três principais componentes de variação espacial nos erros de parcelas e efeitos de competição, em nível genético e residual. Os clones foram assumidos de efeitos aleatórios e as testemunhas de efeitos fixos, pois vieram de processos diferentes. As análises individuais e conjuntas foram desenvolvidas neste material pois os experimentos puderam ser agrupados em dois tipos: (i) um delineamento longitudinal (duas colheitas) e (ii) cinco grupos de experimentos (cada grupo uma região com três locais). Para os estudos de delineamentos, um tamanho fixo de experimento foi assumido para se avaliar a eficiência do delineamento não replicado (empregado nesses 21 experimentos) com os não replicados otimizado espacialmente, os parcialmente replicados com testemunhas e os parcialmente replicados otimizado espacialmente. Quatro estudos de simulação foram feitos para avaliar i) os modelos ajustados, sob condições de efeito de competição em nível genético, ii) a acurácia das estimativas vindas dos modelos de análise individual e conjunta; iii) a relação entre tombamento da cana e a correlação residual negativa, e iv) a eficiência dos delineamentos. Para concluir, as principais informações utilizadas nos estudos de simulação foram: o número de vezes que o algoritmo convergiu; a variância na estimativa dos parâmetros; a correlação entre os EBLUPs genético direto e os efeitos genéticos reais; a assertividade de seleção ou a semelhança média, sendo semelhança medida como a porcentagem dos 30 clones com os maiores EBLUPS genético e os 30 melhores verdadeiros clones; e a estimativa da herdabilidade ou do ganho genético.

Palavras-chave: Modelos mistos; Correlação autorregressiva; Correlação em banda; Competição; Tombamento; Delineamento não replicado; Delineamento ótimo; Delineamento parcialmente replicado; Ganho genético; Assertividade na seleção; Estudo de simulação

ABSTRACT

Design and analysis of sugarcane breeding experiments: a case study

One purpose of breeding programs is the selection of the better test lines. The accuracy of selection can be improved by using optimal design and using models that fit the data well. Finding this is not easy, especially in large experiments which assess more than one hundred lines without the possibility of replication due to the limited material, area and high costs. Thus, the large number of parameters in the complex variance structure that needs to be fitted relies on the limited number of replicated check varieties. The main objectives of this thesis were to model 21 trials of sugarcane provided by “Centro de Tecnologia Canavieira” (CTC - Brazilian company of sugarcane) and to evaluate the design employed, which uses a large number of unreplicated test lines (new varieties) and systematic replicated check (commercial) lines. The mixed linear model was used to identify the three major components of spatial variation in the plot errors and the competition effects at the genetic and residual levels. The test lines were assumed as random effects and check lines as fixed, because they came from different processes. The single and joint analyses were developed because the trials could be grouped into two types: (i) one longitudinal data set (two cuts) and (ii) five regional groups of experiment (each group was a region which had three sites). In a study of alternative designs, a fixed size trial was assumed to evaluate the efficiency of the type of unreplicated design employed in these 21 trials comparing to spatially optimized unreplicated and p -rep designs with checks and a spatially optimized p -rep design. To investigate models and design there were four simulation studies to assess mainly the i) fitted model, under conditions of competition effects at the genetic level, ii) accuracy of estimation in the separate versus joint analysis; iii) relation between the sugarcane lodging and the negative residual correlation, and iv) design efficiency. To conclude, the main information obtained from the simulation studies was: the convergence number of the algorithm model analyzed; the variance parameter estimates; the correlations between the direct genetic EBLUPs and the true direct genetic effects; the assertiveness of selection or the average similarity, where similarity was measured as the percentage of the 30 test lines with the highest direct genetic EBLUPs that are in the true 30 best test lines (generated); and the heritability estimates or the genetic gain.

Keywords: Mixed models; Autoregressive correlation; Banded correlation; Competition; Lodging; Unreplicated design; Optimal design; p -rep design; Genetic gain; Assertiveness of selection; Simulation study

LIST OF FIGURES

1.1	Map of Brazil with the estimated percentage of sugarcane production in April/2016. . . .	21
1.2	Map with the six regions where CTC carries out the experiments. On the left is the complete map of Brazil and right the specific CTC areas. Source: CTC (2013)	22
2.1	Boxplots of the correlations between direct genetic EBLUPs for pairs of models from each scenario (a, b, c and d). In (b) Model 2 converged only for 90 out of 1000 simulated data sets and hence the density plot for genetic components effects are not displayed (see Table 2.6).	36
2.2	Boxplots of the correlation between the true direct genetic effects and the respective EBLUP's from each model for each scenario (a, b, c and d). In (b) Model 2 converged only for 90 out of 1000 simulated data sets and hence the density plot for genetic components effects are not displayed (see Table 2.6).	37
2.3	Densities of the selection gain from each model for each scenario (a, b, c and d). In (b) Model 2 converged only for 90 out of 1000 simulated data sets and hence the density plot for genetic components effects are not displayed (see Table 2.6).	37
2.4	Percentage of average similarity of the 30 test lines selected as best compared to the true best 30 lines in each model and scenario (a, b, c and d). In scenario b, Model 2 converged only for 90 out of 1000 simulated data sets and hence the density plot for genetic components effects are not displayed (see Table 2.6).	38
2.5	Densities of the direct genetic (first column of plots), neighbour genetic (third) components along with their covariance (second column) for each scenario and model. Each row represents a different scenario and the vertical lines represent the assumed values of the parameters. Remember that the convergence was different in each one (see Table 2.6) and only Models 2 and 4 assumed neighbour genetic effect. In scenario b Model 2 converged only for 90 out of 1000 simulated data sets and hence the density plot for genetic components effects are not displayed.	39
2.6	Densities of the residual component(first column of plots) and the column (second column) and row (third column) correlations for each scenario and model. Each row represents a different scenario and the vertical lines represent the assumed values of the parameters. Remember that the convergence was different in each one (see Table 2.6) and only Model 3 and 4 assumed residual correlation. In scenario b Model 2 converged only for 90 out of 1000 simulated data sets and hence the density plot for genetic components effects are not displayed.	40
2.7	Densities of the global components, column (in the left) and row (in the right), for each scenario and model. Each row represents a different scenario and the vertical lines represent the assumed values of the parameters. Remember that the convergence was different in each one (see Table 2.6). In scenario b Model 2 converged only for 90 out of 1000 simulated data sets and hence the density plot for genetic components effects are not displayed.	41
2.8	Average EBLUP correlations in the individual (light symbols) and joint (dark symbols) analysis for autoregressive models in each scenario. The panels refer to comparison between (a) pairs of Locals; (b) individual and joint analysis at the same Local.	42
2.9	Average similarity between the true and estimated 30 best test lines (around top 7%). The dark symbols represent the result from joint analysis and the light are the individuals for autoregressive models in each scenario.	43

2.10	Densities of the heritability for each site from the joint analysis for autoregressive models in each scenario. The panels refer to the scenarios.	43
2.11	Densities of the direct genetic components from autoregressive and banded models for each Local from scenarios (a, b, c and d). The vertical lines represent the assumed values of the parameters.	44
2.12	Densities of the residual variance components from autoregressive and banded models for each Local from scenario a. The vertical lines represent the assumed values of the parameters.	45
2.13	Densities of the residual variance components from autoregressive and banded models for each Local from scenario b. The vertical lines represent the assumed values of the parameters.	45
2.14	Densities of the residual variance components from autoregressive and banded models for each Local from scenario c. The vertical lines represent the assumed values of the parameters.	46
2.15	Densities of the residual variance components from autoregressive and banded models for each Local from scenario d. The vertical lines represent the assumed values of the parameters.	46
3.1	Densities of the residual component and the column and row correlation for each scenario. The curves within a panel are for the different percentage of plots lodged. The plot of the residual components have been truncated to exclude a few estimated values that were in excess of 4.	52
4.1	Layout of the Design 0, where there are four commercial varieties, named “A”, “B”, “C” and “T” checks, and 414 test lines in the unnamed plots.	56
4.2	Layouts from the optimal designs 1 (panels (a) and (b)), 2 (panels (c) and (d)), 3 (panels (e) and (f)). Panels (a), (c) and (e) used $\phi_r = 0.4$ and the other panels used $\phi_r = -0.25$. “A”, “B”, “C” and “T” are the checks and the numbers are the duplicated test lines as well the unnamed plots.	58
4.3	Layouts from the optimal designs 4 (panels (a) and (b)) and 5 (panels (c) and (d)). Panels (a) and (c) used $\phi_r = 0.4$ and the other panels used $\phi_r = -0.25$. “A”, “B”, “C” and “T” are the checks and the numbers are the duplicated test lines as well the unnamed plots.	59
4.4	Densities of the correlation between the true and estimated genetic effects (first column of plots), relative genetic gain (second column) and prediction error variance - PEV (third) for each scenario and design.	61
4.5	Similarity percentage between true and estimated top 7% of test lines. The horizontal gray line indicates 50%. The panels represent scenarios the different scenarios.	62
B.1	Boxplot of TCH for each group of lines for Paraná cuts. Panel (a) refers to first cut and (b) second.	75
B.2	Heat map for TCH in the first cut of the Paraná.	76
B.3	Heat map for TCH in the second cut of the Paraná.	76
B.4	Plots of the row and column faces of the empirical variograms for the residuals referring to Models 4 (panels (a) and (d)); 6 (panels (b) and (e)) and 18 (panels (c) and (f)). Panels (a), (b) and (c) refer to the column direction and panels (d), (e) and (f) refer to the row direction.	79
B.5	Dispersion plot of the genetic EBLUPs values of the Model 18 and the traditional non-spatial analysis model for this design (without spatial dependence) for 1 st cut of Paraná. The cut-offs for the 30 best test lines (7 % upper) in each Model are indicated by the dotted line.	81

B.6	Panels (a) and (b) refer respectively to the plots of the column and row faces of the empirical variograms for the residuals referring to Model 3. Panel (c) is the semi-variogram of the Model 14.	83
B.7	Dispersion plot of the genetic EBLUPs values of the Model 14 and the traditional non-spatial analysis model for this design (without spatial dependence) for 2 nd cut of Paraná. The cut-offs for the 30 best test lines (7 % upper) in each Model are indicated by the dotted line.	85
B.8	Dispersion plot of the predicted values from selected model in each cut of Paraná. The cut-offs for the 30 best test lines (7% upper) in each Model are indicated by the dotted line.	85
B.9	Predicted values from individual (indiv) and joint analysis of Paraná cuts. The left panel is for the first cut and the second is to the right.	88
C.1	Boxplot of TCH for each group of lines for Locals of Paraná.	90
C.2	Heat map relating to TCH for Local 651 of Paraná.	90
C.3	Heat map relating to TCH for Local 851 of Paraná.	91
C.4	Heat map relating to TCH for Local 852 of Paraná.	91
C.5	Plots of the row and column faces of the empirical variogram for the residuals for the Local 651 of the Paraná experiment referring to models 19 (panels (a) and (c)) and 20 (panels (b) and (d)). The panels (a) and (b) are for the column direction and the others are for the row direction.	94
C.6	Dispersion plot of the genetic EBLUPs values of the Model 20 and the traditional non-spatial analysis model for this design (without spatial dependence) for Local 651 of the Paraná. The cut-offs for the 14 best test lines (7 % upper) in each Model are indicated by the dotted line.	95
C.7	Plot of the empirical semi-variogram for the residuals for the Local 851 of the Paraná experiment referring to Model 15.	98
C.8	Dispersion plot of the genetic EBLUPs values of the Model 15 and the traditional non-spatial analysis model for this design (without spatial dependence) for Local 851 of the Paraná. The cut-offs for the 14 best test lines (7 % upper) in each Model are indicated by the dotted line.	99
C.9	Plot of the empirical semi-variogram for the residuals for the Local 852 of the Paraná experiment referring to Model 15.	101
C.10	Dispersion plot of the genetic EBLUPs values of the Model 15 and the traditional non-spatial analysis model for this design (without spatial dependence) for Local 852 of the Paraná. The cut-offs for the 14 best test lines (7 % upper) in each Model are indicated by the dotted line.	102
C.11	Dispersion plots of the genetic EBLUPs from model selected for each Paraná Local. Panels the relationship between pairs of Locals (a) 651 and 851; (b) 651 and 852; and (c) 851 and 852. The cut-offs for the 14 best test lines in each Local are indicated by the dotted line.	103
D.1	Boxplot of TCH for each group of lines for Locals of Goiás.	110
D.2	Heat map relating to TCH for Local 3 of Goiás.	110
D.3	Heat map relating to TCH for Local 521 of Goiás.	111
D.4	Heat map relating to TCH for Local 533 of Goiás.	111
D.5	Heat map relating to fall down for Local 3 of Goiás.	112
D.6	Plots of the empirical semi-variogram for the residuals for the Local 3 of the Goiás experiment referring to models 19 and 21.	114

D.7	Dispersion plot of the genetic EBLUPs values of the Model 19 and the traditional non-spatial analysis model for this design (without spatial dependence) for Goiás Local 3. The cut-offs for the 15 best test lines (7 % upper) in each model are indicated by the dotted line.	115
D.8	Plots of the empirical semi-variogram for the residuals for Local 521 of Goiás experiment referring to Models (a) 20 and (b) 23.	118
D.9	Dispersion plot of the genetic EBLUPs values of the Model 20 and the traditional non-spatial analysis model for this design (without spatial dependence) for Goiás Local 521. The cut-offs for the 15 best test lines (7 % upper) in each model are indicated by the dotted line.	119
D.10	Plots of the empirical semi-variogram for residuals from Model 15 for Local 533 of the Goiás experiment.	121
D.11	Dispersion plot of the genetic EBLUPs values of the Model 20 and the traditional non-spatial analysis model for this design (without spatial dependence) for Goiás Local 533. The cut-offs for the 15 best test lines (7 % upper) in each model are indicated by the dotted line.	122
D.12	Dispersion plots for the genetic EBLUPs values from model selected for each Local in Goiás. Panels the relationship between pairs of Locals (a) 3 and 521; (b) 3 and 533; and (c) 521 and 533. The cut-offs for the 15 best test lines in each Local are indicated by the dotted line.	123
E.1	Boxplot of TCH for each group of lines for Locals of Ribeirão.	128
E.2	Heat map relating to TCH for Local 20 of Ribeirão.	128
E.3	Heat map relating to TCH for Local 72 of Ribeirão.	129
E.4	Heat map relating to TCH for Local 140 of Ribeirão.	129
E.5	Plots of the column and row faces of the empirical variogram for the residuals for the Local 20 of the Ribeirão experiment for models 2 (panels (a) and (c)), and 14 (panels (b) and (d)). The panels (a) and (b) are for the column direction and the others for row.	132
E.6	Dispersion plot of the genetic EBLUPs values of the Model 14 and the traditional non-spatial analysis model for this design (without spatial dependence) for Ribeirão Local 20. The cut-offs for the 30 best test lines (7 % upper) in each model are indicated by the dotted line.	133
E.7	Plot of the empirical semi-variogram for the residuals for Local 140 of Ribeirão experiment referring to Model 14.	136
E.8	Dispersion plot of the genetic EBLUPs values of the Model 14 and the traditional non-spatial analysis model for this design (without spatial dependence) for Ribeirão Local 140. The cut-offs for the 30 best test lines (7 % upper) in each model are indicated by the dotted line.	137
E.9	Dispersion plot of the genetic EBLUPs from model selected for Ribeirão Locals 20 and 140. The cut-offs for the 30 best test lines in each Local are indicated by the dotted line.	138
F.1	Boxplot of TCH for each group of lines for Locals of Piracicaba.	142
F.2	Heat map relating to TCH for Local 54 of Piracicaba.	142
F.3	Heat map relating to TCH for Local 58 of Piracicaba.	143
F.4	Heat map relating to TCH for Local 76 of Piracicaba.	143
F.5	Plots of the row and column faces of the empirical variogram for the residuals for the Local 54 of Piracicaba experiment for models 8 (panels (a) and (c)) and 18 (panels (b) and (d)). The panels (a) and (b) are column direction and the others are row direction.	145

F.6	Dispersion plot of the genetic EBLUPs values of the Model 18 and the traditional non-spatial analysis model for this design (without spatial dependence) for Piracicaba Local 54. The cut-offs for the 30 best test lines (7 % upper) in each model are indicated by the dotted line.	147
F.7	Plot of the empirical semi-variogram for the residuals of the Local 58 of Piracicaba. This refers to Model 19.	149
F.8	Dispersion plot of the genetic EBLUPs values of the Model 19 and the traditional non-spatial analysis model for this design (without spatial dependence) for Piracicaba Local 58. The cut-offs for the 30 best test lines (7 % upper) in each model are indicated by the dotted line.	151
F.9	Plots of the column and row faces of the empirical variogram for the residuals for Local 54 of Piracicaba experiment for models 18 (panels (a) and (c)) and 19 (panels (b) and (d)). The panels (a) and (b) are column direction and the others are row direction.	153
F.10	Dispersion plot of the genetic EBLUPs values of the Model 19 and the traditional non-spatial analysis model for this design (without spatial dependence) for Piracicaba Local 76. The cut-offs for the 30 best test lines (7 % upper) in each model are indicated by the dotted line.	155
F.11	Dispersion plots for the genetic EBLUPs values from selected model for each Piracicaba Local. Panels refer the relation between Local (a) 54 and 58; (b) 54 and 76; (c) 58 and 76. The cut-offs for the 30 best test lines in each Local are indicated by the dotted line.	155
G.1	Boxplot of TCH in each area for Local 101 of Araçatuba.	160
G.2	Heat maps relating to TCH in Local 101 of Araçatuba. The panel refers to (a) area 1; (b) area 3 and (c) area 2. The names PAD1, PAD2, PAD3 and INTERC are the check plots and the test lines are labeled alpha-numerically (AR plus number).	160
G.3	Plots of the row and column faces of the empirical variogram for the residuals for Local 101 area 1 of Araçatuba experiment for models 4 (panels (a) and (c)) and 15 (panels (b) and (d)). The panels (a) and (b) are column direction and the others are row direction.	163
G.4	Boxplot of TCH in each area for Local 130 of Araçatuba.	166
G.5	Heat maps relating to TCH in Local 130 of Araçatuba. The panel refers to (a) area 1 and (b) area 2. The names PAD1, PAD2, PAD3 and INTERC are the check plots and the test lines are labeled alpha-numerically (AR plus number). “x” represents the empty plots.	167
G.6	Plots of the column and row faces of the empirical variogram for the residuals for Local 130 area 2 of Araçatuba experiment for Model 8. The panels (a) is column direction and (b) is row direction.	170
G.7	Boxplot of TCH in each area for Local 551 of Araçatuba.	172
G.8	Heat maps relating to TCH in the Local 551 of Araçatuba. The panel (a) refers to area 1 and (b) area 2. The names PAD1, PAD2, PAD3 and INTERC are the check plots and the test lines are labeled alpha-numerically (AR plus number). “x” represents the empty plots.	172
G.9	Plots of the column and row faces of the empirical variogram for the residuals for Local 551 area 1 of Araçatuba experiment for Model 3. The panels (a) is column direction and (b) is row direction.	173

LIST OF TABLES

2.1	Summary of the analyzed experiments.	28
2.2	Some possible covariance structures for direct genetic effects over cut or sites (<i>i</i>).	30
2.3	Summary of the selected models for all experiments. All the models have a fixed check varieties effects and random effects for test lines denominated as the direct genetic effects (G). The other effects are represented with the symbols: Cext (co-variate, external effect in the 3 east columns); H (harvest); C (column); R (row); N (neighbour genetic); spl(.) (spline in some direction indicated in parentheses) and lin(.) (linear trend in some direction indicated in parentheses). The structures are US(.) - unstructured; AR(1) - first-order autoregressive, Band(.) - banded correlation, the number indicates the order of correlation, SAR(2) - constrained autoregressive and Id - identity.	31
2.4	REML estimates of variance parameters from each fitted model of the individual analysis in relation to the residual component (σ^2). $\gamma = \sigma^2/\sigma^2$; given that each letter represent: <i>c</i> -column; <i>r</i> - row; <i>g</i> - direct genetic; <i>n</i> - neighbour genetic; <i>gn</i> - genetic covariance; <i>u</i> - nugget. ϕ are the local residual parameters related to the correlations. Observe Table 2.3 to check if ϕ is the direct correlation (banded models) or not (autoregressive models).	32
2.5	Correlations (a) between the direct genetic EBLUPs from each fitted model of the individual analysis; (b) between the direct genetic effects from the joint analysis.	33
2.6	Number of convergence cases in each model and simulation scenario for single site experiments, given 1000 data sets.	36
2.7	Number of convergence cases in each model and scenario for separate versus joint analyses, given 1000 data sets.	41
3.1	Convergence number of the lodging models algorithms in each scenario and percentage of plots lodged, given 1000 simulated data sets.	51
4.1	Numbers of check lines and duplicated test lines for five spatially optimized designs. Design 1 has the same numbers of check lines as Design 0. All designs were on grid of 500 plots and had 414 test lines.	57
4.2	Relative efficiencies of spatially optimized designs to unreplicated designs with systematic checks. In parenthesis are the PEV values for Design 0.	60
4.3	Number of estimation algorithm converged for each design-scenario combination, given 1000 data sets for each.	61
B.1	Descriptive analysis of the groups of sugarcane carried out in Paraná in the first and second cut.	75
B.2	Summary of the fitted models to the Paraná first cut data with restricted log likelihood (Log-lik.) and the <i>p</i> -value of the REML ratio test. All models include a random direct genetic effects for the test lines (G), fixed check varieties effects and co-variate. The other effects are represented with the symbols: H (harvest); Co(cone); S (sowing); C (column); R (row); N (neighbour genetic); spl(.) (spline in some direction indicated in parentheses) and lin(.) (linear trend in some direction indicated in parentheses). The structures can be: US(.)- unstructured; RR(.)- reduced rank; AR(1) - first-order autoregressive; Band(.) - banded correlation, the number indicates the order of correlation; and Id - identity.	78
B.3	REML variance parameters estimates from fitted Model 18 to the experiment of Paraná, first cut.	79

B.4	Wald F tests for fixed terms for Model 18. Df means degrees of freedom and DenDF is the approximate denominator degrees of freedom.	80
B.5	Estimates of the fixed effects for Model 18 and their respective standard errors.	80
B.6	The 30 best test lines with predicted values (pred. value) and standard errors (stand. error) from Model 18.	80
B.7	Summary of the fitted models to the Paraná second cut data with restricted log likelihood (Log-lik.) and the p -value of the REML ratio test. All models include a random direct genetic effects for the test lines (G), fixed check varieties effects and co-variate. The other effects are represented with the symbols: H (harvest); Co (cone); S (sowing); C (column); R (row); N (neighbour genetic); spl(.) (spline in some direction indicated in parentheses) and lin(.) (linear trend in some direction indicated in parentheses). The structures can be: US(.) - unstructured; RR(.) - reduced rank; sar(2) - constrained autoregressive; AR(.) - autoregressive; Band(.) - banded correlation; and Id - identity. The number indicates the order of correlation.	82
B.8	REML estimates of variance parameters from fitted Model 14 for the experiment of Paraná, second cut.	83
B.9	Wald F tests for fixed terms for Model 14. Df means degrees of freedom and DenDF is the approximate denominator degrees of freedom.	84
B.10	Estimates of the fixed effects for Model 14 and their respective standard errors.	84
B.11	The 30 best test lines with predicted values (pred. value) and respective standard errors (stand. error) for Model 14.	84
B.12	Summary of the fitted models to joint analysis of the Paraná experiments in the first (1) and second (2) cut, with restricted log likelihood (Log-lik.); p -value of the REML ratio test and AIC. All the models have the same fixed; local and global effects modeled in the individual analysis plus the cut fixed effect.	86
B.13	REML estimates of variance parameters from fitted Model 4 to the joint analysis of the first-cut (1) and second-cut (2) of the experiment carried out in the Paraná.	86
B.14	Wald F tests for fixed terms from Model 4. Df means degrees of freedom and DenDF is the approximate denominator degrees of freedom.	87
B.15	Predicted values and standard errors of the 30 best test lines of the Model 4 for each cut. .	87
C.1	Descriptive analysis of the groups of clones carried out in Paraná.	89
C.2	Summary of the fitted models to Local 651 with restricted log-likelihood (log-lik.) and the p -value of the REML ratio test. All models include a random direct genetic effects for the test lines (G) and fixed check varieties effects. The other effects are represented with the symbols: H (harvest); Co (cone); S (sowing); C (column); R (row); N (neighbour genetic); spl(.) (spline in some direction indicated in parentheses) and lin(.) (linear trend in some direction indicated in parentheses). The structures can be: US(.) - unstructured; AR(1) - first-order autoregressive; Band(.) - banded correlation, the number indicates the order of correlation; and Id - identity.	93
C.3	REML estimates of variance parameters for fitted Model 20 for the experiment in the Local 651 of the Paraná.	94
C.4	Wald F tests for fixed terms for Model 20. Df means degrees of freedom and DenDF is the approximate denominator degrees of freedom.	94
C.5	Estimates of the fixed effects for Model 20 and their respective standard errors.	95
C.6	The 14 best test lines with predicted values and respective standard errors for Model 20. . .	95

C.7	Summary of the fitted models to Local 851 with restricted log-likelihood (log-lik.) and the p -value of the REML ratio test. All models include a random direct genetic effects for the test lines (G) and fixed check varieties effects. The other effects are represented with the symbols: H (harvest); Co (cone); S (sowing); C (column); R (row); N (neighbour genetic); spl(.) (spline in some direction indicated in parentheses) and lin(.) (linear trend in some direction indicated in parentheses). The structures can be: RR(.) - reduced rank; AR(1) - first-order autoregressive; Band(.) - banded correlation, the number indicates the order of correlation; and Id - identity.	97
C.8	REML estimates of variance parameters for fitted Model 15 for the experiment in the Local 851 of the Paraná.	98
C.9	Wald F tests for fixed terms from Model 15. Df means degrees of freedom and DenDF is the approximate denominator degrees of freedom.	98
C.10	Estimates of the fixed effects for Model 15 and their respective standard errors.	98
C.11	The 14 best test lines with predicted values and respective standard errors for Model 15.	99
C.12	Summary of the fitted models to Local 852 with restricted log-likelihood (log-lik.) and the p -value of the REML ratio test. All models include a random direct genetic effects for the test lines (G) and fixed check varieties effects. The other effects are represented with the symbols: H (harvest); Co (cone); S (sowing); C (column); R (row); N (neighbour genetic); spl(.) (spline in some direction indicated in parentheses) and lin(.) (linear trend in some direction indicated in parentheses). The structures can be: US(.) - unstructured; RR(.) - reduced rank; AR(.) - autoregressive structure; Band(.) - banded correlation, the number indicates the order of correlation; sar(2) constrained autoregressive; and Id - identity.	100
C.13	REML estimates of variance parameters for Model 15 to the experiment in the Local 852 of the Paraná.	101
C.14	Wald F tests for fixed terms for Model 15. Df means degrees of freedom and DenDF is the approximate denominator degrees of freedom.	101
C.15	Estimates of the fixed effects for Model 15 and their respective standard errors.	102
C.16	The 14 best test lines with predicted values and respective standard errors for Model 15.	102
C.17	Summary of the fitted models to joint analysis of the Paraná experiments carried out in the Locals 851 (2) and 852 (3), with REML log (log-lik.) and the p -value of the REML ratio test. All the models have the same effects as modeled in the individual analysis plus the fixed Local effect. Here σ_{gi}^2 is the genetic variance at the i th Local and $\rho_{g_{ii'}}$ is the genetic correlation between the i th and i' th Locals.	104
C.18	Summary of the models fitted in joint analysis of the Paraná experiments carried out in the Locals 561 (1), 851 (2) and 852 (3), with REML ratio tests (log-lik.) and the p -values of the REML ratio tests. All the models have the same effects as modeled in the individual analysis plus the fixed Local effect. Here σ_{gi}^2 is the genetic variance at the i th Local, $\rho_{ii'}$ is the genetic correlation between the i th and i' th Locals and σ_{g*} is the genetic covariance between two locals which have been hypothesis to be equal for two or more pairs of Locals.	104
C.19	REML estimates of the variance parameters from fitted Model 3 in the joint analysis of the Locals in Paraná.	105
C.20	Wald F test for fixed terms of the Model 3. Df means degrees of freedom and DenDF is the approximate denominator degrees of freedom.	105
C.21	Estimates of the fixed effects from Model 3 and their respective standard errors.	106
C.22	Predicted values (pred.) and standard error (error) of the 14 best test lines from Model 3 for each Local.	106

D.1	Descriptive analysis of TCH for the groups of clones carried out in Goiás.	109
D.2	Summary of the fitted models to Local 3 of the Goiás experiment with restricted log-likelihood (log-lik.) and the p -value of the REML ratio test. All models include a random direct genetic effect for the test lines (G) and fixed check varieties effect. The other effects are represented with the letters: H (harvest); Co (cone); S (sowing); C (column); R (row); N (neighbour genetic); spl(.) (spline in some direction indicated in parentheses) and lin(.) (linear trend in some direction indicated in parentheses). The structures can be: US(.) - unstructured; AR(.) - autoregressive; sar(2) - constrained autoregressive; Band(.) - banded correlation, the number indicates the order of correlation; and Id - identity.	113
D.3	REML estimates of variance parameters from fitted Model 19 to the experiment in the Local 3 of the Goiás.	114
D.4	Wald F tests for fixed terms for Model 19. Df means degrees of freedom and DenDF is the approximate denominator degrees of freedom.	114
D.5	Estimates of the fixed effects for Model 19 and their respective standard errors.	115
D.6	The 15 best test lines with predicted values and respective standard errors from Model 19.	115
D.7	Summary of the fitted models to Local 521 of the Goiás experiment with restricted log-likelihood (log-lik.) and the p -value of the REML ratio test. All the models include a random direct genetic effects for the test lines (G) and fixed check varieties effects. The other effects are represented with the letters: H (harvest); Co(cone); S (sowing); C (column); R (row); N (neighbour genetic); spl(.) (spline in some direction indicated in parentheses) and lin(.) (linear trend in some direction indicated in parentheses). The structures can be: US(.) - unstructured; RR(.) - reduced rank; AR(.) - autoregressive; Band(.) - banded correlation, the number indicates the order of correlation; sar(2) - constrained autoregressive; and Id - identity.	117
D.8	REML estimates of variance parameters for Model 20 fitted to the data from Local 521 from the Goiás experiment.	118
D.9	Wald F tests for fixed terms from Model 20. Df means degrees of freedom and DenDF is the approximate denominator degrees of freedom.	118
D.10	Estimates of the fixed effects for Model 20 and their respective standard errors.	118
D.11	The 15 best test lines with predicted values and respective standard errors from Model 20.	119
D.12	Summary of the fitted models to Local 533 of the Goiás experiment with restricted log-likelihood (log-lik.) and the p -value of the REML ratio test. All the models include a random direct genetic effects for the test lines (G) and fixed check varieties effects. The other effects are represented with the letters: H (harvest); Co (cone); S (sowing); C (column); R (row); N (neighbour genetic); spl(.) (spline in some direction indicated in parentheses) and lin(.) (linear trend in some direction indicated in parentheses). The structures can be: RR(.) - reduced rank; AR(.) - first-order autoregressive; sar(2) - constrained autoregressive; Band(.) - banded correlation, the number indicates the order of correlation; and Id - identity.	120
D.13	REML estimates of variance parameters from fitted Model 15 for the experiment in the Local 533 of the Goiás.	121
D.14	Wald F tests for fixed terms from Model 15. Df means degrees of freedom and DenDF is the approximate denominator degrees of freedom.	121
D.15	Estimates of the fixed effects from Model 15 and their respective standard errors.	122
D.16	The 15 best test lines with predicted values and respective standard errors for Model 15.	122

D.17	Summary of the fitted models to joint analysis of the Goiás experiments carried out in the Locals 3 (1) and 533 (3), with REML test (log-lik.) and the p -value of the REML ratio test. All the models have the same effects as modeled in the individual analysis plus the fixed Local effects. Here σ_{gi}^2 is the genetic variance at the i th Local, $\sigma_{gii'}$ is the genetic covariance between the i th and i' th Locals which have been hypothesized to be equal for two or more pairs of Locals.	124
D.18	REML estimates of variance parameters for fitted Model 3 to the joint analysis of the Local 3 (1) and 533 (3) of the experiment carried out in Goiás.	124
D.19	Wald F tests for the fixed terms in Model 3. Df means degrees of freedom and DenDF is the approximate denominator degrees of freedom.	124
D.20	The 15 best test lines with predicted values and respective standard error from Model 3 for each Local.	125
E.1	Descriptive analysis of the groups of clones carried out in Ribeirão.	127
E.2	Summary of the fitted models for Local 20 with restricted log-likelihood (log-lik.) and the p -value for the REML ratio test. All the models include a random direct genetic effects for the test lines (G) and fixed checks varieties effects. The other effects are represented with the symbols: H (harvest); Co (cone); S (sowing); C (column); R (row); N (neighbour genetic); spl(.) (spline in some direction indicated in parentheses) and lin(.) (linear trend in some direction indicated in parentheses). The structures can be: US(.) - unstructured; AR(1) - autoregressive; Band(.) - banded correlation, the number indicates the order of correlation; and Id - identity.	131
E.3	REML estimates of variance parameters for fitted Model 14 for the experiment in Local 20 from Ribeirão.	132
E.4	Wald F test for fixed terms for Model 14. Df means degrees of freedom and DenDF is the approximate denominator degrees of freedom.	132
E.5	Estimates of the fixed effects for Model 14 and their respective standard errors.	133
E.6	The 30 best test lines with predicted values and respective standard errors for Model 14.	133
E.7	Estimates of the variance parameters and standard error for the model with first-order autoregressive structure in row and column direction, and direct genetic, row and column random effects. The log-likelihood of the model is -1851.92.	134
E.8	Summary of the models fitted for Local 140 with restricted log-likelihood (log-lik.) and the p -value for the REML ratio test. All the models include a random direct genetic effects for the test lines (G) and fixed checks varieties effects. The other effects are represented with the symbols: H (harvest); Co (cone); S (sowing); C (column); R (row); N (neighbour genetic); spl(.) (spline in some direction indicated in parentheses) and lin(.) (linear trend in some direction indicated in parentheses). The structures can be: US(.) - unstructured; AR(1) - first-order autoregressive; Band(.) - banded correlation, the number indicates the order of correlation; and Id - identity.	135
E.9	REML estimates of variance parameters for fitted Model 14 to the experiment in the Local 140 of the Ribeirão.	136
E.10	Wald F test of fixed terms from Model 14. Df means degrees of freedom and DenDF is the approximate denominator degrees of freedom.	136
E.11	Estimates of the fixed effects for Model 14 and their respective standard errors.	137
E.12	The 30 best test lines with predicted values (pred. value) and respective standard errors (stand. error) for Model 14.	137

E.13	Summary of the models fitted to joint analysis for the Ribeirão experiments carried out for the Local 20 (1) and 140 (3), with REML log (log-lik.) and the p -value of the REML ratio test. All the models have the same fixed; local and global effects modeled in the individual analysis plus the local fixed effects.	138
E.14	REML estimates of variance parameters for fitted Model 3 for the joint analysis of Local 20 (1) and 140 (3) of the experiment carried out in Ribeirão.	138
E.15	Wald F test for fixed terms of the Model 3. Df means degrees of freedom and DenDF is the approximate denominator degrees of freedom.	139
E.16	Predicted values and standard errors of the 30 best test lines of the Model 3 for each local.	139
F.1	Descriptive analysis of the groups of clones carried out in Piracicaba.	141
F.2	Summary of the fitted models to Local 54 of Piracicaba with restricted log-likelihood (log-lik.) and the p -value of the REML ratio test. All models include a random direct genetic effects for the test lines (G) and fixed checks varieties effects. The other effects are represented with the symbols: H (harvest); Co (cone); S (sowing); C (column); R (row); N (neighbour genetic); spl(.) (spline in some direction indicated in parentheses) and lin(.) (linear trend in some direction indicated in parentheses). The structures can be: US(.) - unstructured; RR(.) - reduced rank; AR(1) - first-order autoregressive; Band(.) - banded correlation, the number indicates the order of correlation; and Id - identity.	144
F.3	REML estimates of variance parameters for fitted Model 18 for the experiment of Piracicaba, Local 54.	145
F.4	Wald F test for fixed factors for Model 18. Df means degrees of freedom and DenDF is the approximate denominator degrees of freedom.	146
F.5	Estimates of the fixed effects for Model 18 and their respective standard errors.	146
F.6	The 30 best test lines with predicted values (pred. value) and respective standard errors (stand. error) for Model 18.	146
F.7	Summary of the fitted models to Local 58 of Piracicaba with restricted log-likelihood (log-lik.) and the p -value of the REML ratio test. All models have a random direct genetic effects for the test lines (G) and fixed check varieties effects. The other effects are represented with the symbols: H (harvest); Co(cone); S (sowing); C (column); R (row); N (neighbour genetic); spl(.) (spline in some direction indicated in parentheses) and lin(.) (linear trend in some direction indicated in parentheses). The structures can be: US(.) - unstructured; RR(.) - reduced rank. AR(.) - autoregressive; Band(.) - banded correlation, the number indicates the order of correlation; sar(2) - constrained autoregressive 3; and Id - identity.	148
F.8	REML estimates of variance parameters for fitted Model 19 for the experiment of Piracicaba, Local 58.	149
F.9	Wald F test for fixed terms from Model 19. Df means degrees of freedom and DenDF is the approximate denominator degrees of freedom.	150
F.10	Estimates of the fixed effects for Model 19 and their respective standard errors.	150
F.11	The 30 best test lines with predicted values (pred. value) and respective standard errors (stand. error) for Model 19.	150

F.12	Summary of the fitted models to Local 76 of Piracicaba with restricted log-likelihood (log-lik.) and the p -value of the REML ratio test. All models include a random direct genetic effects for the test lines (G) and fixed checks varieties effects. The other effects are represented with the symbols: H (harvest); Co (cone); S (sowing); C (column); R (row); N (neighbour genetic); spl(.) (spline in some direction indicated in parentheses) and lin(.) (linear trend in some direction indicated in parentheses). The structures can be: US(.) - unstructured; RR(.) - reduced rank; AR(.) - autoregressive; Band(.) - banded correlation, the number indicates the order of correlation; sar(2) - constrained autoregressive; and Id - identity.	152
F.13	REML estimates of variance parameters for fitted Model 19 for the experiment of Piracicaba, Local 76.	153
F.14	Wald F test for fixed terms from Model 19. Df means degrees of freedom and DenDF is the approximate denominator degrees of freedom.	154
F.15	Estimates of the fixed effects from Model 19 and their respective standard errors.	154
F.16	The 30 best test lines with predicted values (pred. value) and respective standard errors (stand. error) for Model 19.	154
F.17	Summary of the fitted models to joint analysis of the Piracicaba experiments carried out in the Local 54 (1), 58 (2) and 76 (3), with REML log (log-lik.) and the p -value of the REML ratio test. All the models have the same effects modeled in the individual analysis plus the fixed Local effects. Here σ_{gi}^2 is the genetic variance at the i th Local, σ_{gij} is the genetic correlation between the i th and j th Locals and σ_{g*} is the genetic correlation between two locals which has been hypothesis to be equal for two or more pairs of Locals.	156
F.18	REML estimates of the variance parameters for fitted Model 5 in the joint analysis of the Locals in Piracicaba.	156
F.19	Predicted values (pred.) and standard errors (error) of the 30 best test lines from Model 5 for each Local.	157
G.1	Descriptive analysis of the groups of clones carried out in Local 101 of Araçatuba.	159
G.2	Summary of the fitted models to Local 101 area 1 with restricted log-likelihood (log-lik.) and the p -value of the REML ratio test. All the models include a random direct genetic effects for the test lines (G) and fixed checks varieties effects. The other effects are represented with the symbols: H (harvest); Co (cone); S (sowing); C (column); R (row); N (neighbour genetic); spl(.) (spline in some direction indicated in parentheses) and lin(.) (linear trend in some direction indicated in parentheses). The structures can be: US(.) - unstructured; RR(.) - reduced rank; AR(1) - autoregressive; Band(.) - banded correlation, the number indicates the order of correlation; and Id - identity.	162
G.3	Wald F test for fixed effects for Model 15. Df means degrees of freedom and DenDF is the approximate denominator degrees of freedom.	163
G.4	Estimates of the fixed effects and their respective standard errors for Model 15.	164
G.5	The 15 best test lines with predicted values (pred. values) and respective standard errors (stand. errors) for Model 15.	164
G.6	REML estimates of variance parameters from models 1 and 2 for the experiment in Local 101 of Araçatuba in area 2. The values in brackets are the standard errors of the estimates.	165
G.7	REML estimates of variance parameters from models 1 and 2 for the experiment in Local 101 of Araçatuba in area 3. The values in brackets are the standard errors of the estimates.	165
G.8	Descriptive analysis of the groups of clones carried out in Local 130 of Araçatuba.	166

G.9	REML estimates of variance parameters for models 1 and 2 for the experiment in Local 130 of Araçatuba in area 1. The values in brackets are the standard errors of the estimates. . . .	168
G.10	Summary of the models fitted to area 2 of the Local 130 with restricted log-likelihood (log-lik.) and the p -value of the REML ratio test. All the models include a random direct genetic effects for the test lines (G) and fixed checks varieties effects. The other effects are represented with the symbols: H (harvest); Co (cone); S (sowing); C (column); R (row); N (neighbour genetic); spl(.) (spline in some direction indicated in parentheses) and lin(.) (linear trend in some direction indicated in parentheses). The structures can be: US(.) - unstructured; AR(1) - first-order autoregressive; Band(.) - banded correlation, the number indicates the order of correlation; and Id - identity.	169
G.11	REML estimates of variance parameters for fitted Model 8 to the experiment in the Local 130 of the Araçatuba.	170
G.12	Wald F tests for the fixed effects for Model 8. Df means degrees of freedom and DenDF is the approximate denominator degrees of freedom.	170
G.13	Estimated fixed parameters and their respective standard errors for Model 8.	170
G.14	Predicted values and standard errors of the test lines for Model 8.	171
G.15	Descriptive analysis of the groups of clones carried out in Local 551 of Araçatuba.	171
G.16	Summary of the models fitted to Local 551 area 1 of the Araçatuba experiment with restricted log (log-lik.). All the models include a random direct genetic effect for the test lines, fixed effects for checks and there is one model with linear trend in column direction (lin(Col)). AR(1) is the first-order autoregressive structure and Id is the identity structure.	173
G.17	REML estimates of variance parameters for models 1, 2 and 3 for the experiment in Local 551 of Araçatuba in area 1. The values in brackets are the standard errors (std.error) of the estimates.	173
G.18	Summary of the models fitted to Local 551 area 2 of the Araçatuba experiment with restricted log-likelihood (log-lik.). All the models include a random direct genetic effects for the test lines, fixed effects for checks and there is one model with linear trend in column direction (lin(Col)). AR(1) is the first-order autoregressive structure; Band(3) is the third-order banded correlation and Id is the identity structure.	174
G.19	REML estimates of variance parameters from models 1 and 3 for the experiment in Local 551 of Araçatuba in area 2. The values in brackets are the standard errors (std.error) of the estimates.	174

1 INTRODUCTION

Important food and bioenergy source, the sugarcane can be a significant component of the economy of many countries in the tropics and subtropics, mainly because of three attributes: high productivity; inputs efficient agriculture; and local processing into value-added products (such as sugar, molasses, ethanol and energy) (MOORE *et al.*, 2013).

Brazil is the largest producer of sugarcane, being responsible for more than 45% of worldwide production (FAOSTAT, 2013). This is one of Brazil's most important trade commodities, although it is used mainly to get ethanol fuel, considered to be a sustainable biofuels economy. For 2016/17 the Companhia Nacional de Abastecimento (CONAB, 2016) estimates a yield of 690.98 million tonnes, an increase of 3.8% relative to the previous years, with a cultivated area of 9,073.7 thousand hectares.

The outlook for agribusiness is favorable due to the petroleum price and its attractiveness as a renewable energy source, which is better for environmental protection. Moreover, the world consumption of sugar and ethanol grows at significant rates. In Brazil, the cultivation of sugarcane continues to grow in areas of traditional plants and dozens of new projects are being carried out in culture expansion regions (DINARDO-MIRANDA *et al.*, 2008).

The total area by percentage of sugarcane in Brazil is higher in the southeast and central-west where around 65.8% of the crop is produced. As shown in Figure 1.1, by Federation Unit, São Paulo (SP) is the biggest producer with 55.2%; followed by Goiás (GO) with 10.7%; Minas Gerais (MG) with 10.1%; Mato Grosso do Sul (MS) with 7.0%; Paraná (PR) with 6.7%; Alagoas (AL) with 2.6%; Mato Grosso (MT) with 2.1% and Pernambuco (PE) with 2.0%. These eight states are responsible for 96.4% of the national production (CONAB, 2016).

Figure 1.1. Map of Brazil with the estimated percentage of sugarcane production in April/2016.

To increase annual yield, breeding companies work every day pursuing lower costs for production and better crop yield. The release and recommendation of new commercial varieties requires several breeding trials with potential genotypes obtained from crossings of known varieties with highly performing established ones. At the initial stage of the breeding program there are many genotypes and little material for planting in experiments, which are usually in large fields. The experiments are often highly unbalanced and involve substantial sources of non-genetic variation and correlation. This is similar to the situation described by STRINGER (2006), who reports experiments with 100 to 1232 clones.

In Brazil, the main sugarcane breeding programs belong to: RIDESA - Rede Interuniversitária para o Desenvolvimento do Setor Sucroenergético; CTC - Centro de Tecnologia Canavieira; and the IAC - Instituto Agrônomo de Campinas. With these breeding programs the sugar industry is provided with

the best varieties to cover a wide environmental range, adapted to specific conditions, either more adverse or favourable.

1.1 CTC company

The CTC is one of the biggest research centres of sugarcane in the world, and holds a very large germplasm bank. Their efforts cover all processes in the sugarcane chain, investing in genetic improvement, biotechnology and new technologies for the sugar energy sector (CTC, 2013).

The development of a new variety of sugar cane is divided into the following steps: I) crossings; II) phase 1, where the best seedlings are planted in the field and evaluated as individual plants, and the best 5 to 10% are selected; III) these clones are planted in phase 2 in 2-row plots, and evaluated for agronomic performance. The best 10% of clones are selected to plant in phase 3. IV) phase 3 clones are planted in 3 locations, but without replication in each location. V) The best 7% of clones from phase 3 are planted in multilocation trials in phase 4 where they are evaluated for potential release as commercial varieties.

The research takes around eight years and requires large areas to accommodate the number of genotypes tested. The results come from trials set up in more than one environment and year. The main production area of Brazil can be divided into six parts due to edaphoclimatic conditions, as Figure 1.2 illustrates.

Figure 1.2. Map with the six regions where CTC carries out the experiments. On the left is the complete map of Brazil and right the specific CTC areas. Source: CTC (2013)

In phase 1 the test lines are compared with commercial varieties (checks), and a percentage of the best new materials are selected. The clones are evaluated as experiments by families. Within and between plots of the same family are different individuals.

For phases 2 and 3 the unreplicated designs with systematic checks (KEMPTON, 1984; CLARKE and STEFANOVA, 2011) are used. These designs have been used commonly for plant breeding trials and they include a large number of test lines and commercial varieties. The new test lines normally are not replicated, because of the available area and small amount of clonal material or seeds, but checks are replicated what allows to capture the residual effect. The advantage of evaluating the treatments with only a small number of replications is a significant economic gain because they need a smaller area and lower production resources. However, decreasing the number of experimental units implies a less accurate analysis, with a smaller number of residual degrees of freedom, when compared with designs in which all lines are replicated.

In these early stages of selection trials (phases 1 to 3), it has been an usual practice to have a visual analysis to be made by a sugarcane expert technician, who observes a series of plant development

characteristics, giving scores for height, colour, number of tillering, presence of disease, lodging, etc. To improve the visual accuracy, a “calibration” process is carried out using checks. For this, in phase 3, three conventional checks (equally replicated and placed normally in neighbouring row plots) are used plus one special check allocated systematically on a diagonal grid throughout the trial, so that in the same column there are always eight or fewer test lines between checks. For this special check is allocated an area of around 11% of the experimental area.

The idea behind this systematic arrangement of the check varieties was that it would contribute to capture better the spatial effects and facilitate the visual evaluation on the selection of the better lines. However, more recently CTC decided to move to a selection of lines based on results given by mixed model analysis.

1.2 Outline of thesis

The purpose of this thesis is to (I) develop the quantitative evaluation of phase 3 experiments based on the fitting of appropriate mixed models, and (II) investigate how the design of the experiments for this phase can be improved.

The thesis is organized as follows. Chapter 2 describes and analyzes several CTC experiments using mixed models that account for the local, global and extraneous variation. When an experiment involves multiple sites or cuts, both individual and joint analyses are conducted. The experimental results were used to develop a simulation study to assess the capacity of the design being used by CTC to fit the models required to describe the genetic and non-genetic variation in their breeding experiments. The accuracy of individual and joint analyses are also compared.

After, the lodging effect is evaluated in Chapter 3 using simulated data and Chapter 4 presents six different designs and their prediction error variance (PEV) optimality criteria measure. At the end, Chapter 5 has general conclusions and recommendations for future research.

2 INDIVIDUAL AND JOINT ANALYSIS OF SUGARCANE EXPERIMENTS TO SELECT TEST LINES

One purpose of the breeding programs is the selection of the best test lines. With the experimental results, the accuracy of selection can be improved by using the best fitted model. Finding the best model is not easy, especially in large experiments which assess more than one hundred lines without replication, mainly because of few replicated check plots and the number of parameters in the complex variance structure that needs to be assumed. The main objective of this chapter is to discuss the models fitted to the data from 21 trials of sugarcane which can be grouped into two types: (i) one longitudinal data set (two cuts) and (ii) five regional groups of experiment (each group is a region with three sites). The single and joint analyses were carried out using mixed models that allowed for genetic and/or residual competition and spatial correlation. In the single-site analyses, evidence was found of competition effects at the residual or genetic level. The correlations between the genetic EBLUPs from the individual analysis were not very strong (less than 0.55), resulting in a different group of selected test lines in each cut or site. Another important observation is that the banded correlations were a more realistic structure than the autoregressive process to model the local effects. The joint analysis, conducted based on the mixed models selected for the individual cuts or sites, showed significant direct genetic correlation between cuts or sites within the same region; the direct correlations were more than double the measure of the individual analysis from EBLUPs. Two simulation studies were conducted. The first verified that, using the type of design employed in the experiments, it is possible to estimate with accuracy the competition effect from individual analysis. The second assessed the individual and joint analysis from three sites with different local effects. In this second study, both the banded and autoregressive structures in the row direction were used in the residual model for the same data set. The results suggested that the competition genetic effects from individual analysis were captured, but the convergence rates were between 47.0% and 68.7%; the difference in estimated parameters and selection between the true simulated model (with competition effect) and the simplest (only check and test lines effects) was not large; the joint analysis achieved better precision in the estimation of the genetic merit of the test lines; and there was no obvious difference between the models fitted with autoregressive or banded structure. In particular, the simulation study identified that the group of selected true best test lines of this experiment can be 55% or less.

Keywords: Assertiveness of selection; Autoregressive correlation; Banded correlation; Competition effect; Mixed models; Simulation study.

2.1 Introduction

Test line selection from breeding programs can depend on the model chosen and it is necessary to identify the effects present in the experiment to ensure the maximum precision in selecting genotypes, especially in large experiments, because of the residual variability. Indeed, as well as the model, the selection precision will be better.

Using mixed linear models, GILMOUR *et al.* (1997) identify three major components of spatial variation in plot errors from field experiments: local, global and extraneous variation. Local variation reflects short distance spatial effects. Global variation ranges across the entire field. Extraneous variation is associated with trial management. They used a separable first-order autoregressive process ($AR(1) \times AR(1)$) in the row and column directions to model the local variation. They affirmed that a more accurate assessment of the presence of global and extraneous variation or outliers can be obtained when using this two-dimensional spatial procedure rather than assuming independent plot errors.

Global variation is modelled using polynomial functions of the spatial coordinates in row and/or

column direction and/or smoothing splines. Extraneous variation can arise from the practice of serpentine harvesting, i.e., harvesting rows of plots in alternating directions, inaccurate plot trimming methods, leading to plots of unequal length, or the use of multiplot seeders (GILMOUR *et al.*, 1997; STEFANOVA *et al.*, 2009).

Then, the autoregressive models for residual variation have become common for researchers (STRINGER and CULLIS, 2002; ATKIN *et al.*, 2009; HUNT *et al.*, 2013; LIU *et al.*, 2015). The variation may occur because of management practices impacting on the experiment, non-stationary spatial trend occurring across the field, and neighbouring plots being more similar than those further apart related to soil fertility or moisture levels (DE FAVERI *et al.*, 2015). SMITH *et al.* (2005) presented an overview of mixed model approaches to analyze the crop cultivar breeding and evaluation trials, merging the spatial and randomization models.

For sugarcane experiments, the plots are generally long and it is to be expected that the correlation between plots sharing the longest boundary (here different rows) will be greater than those sharing the shorter boundary (here different columns). STRINGER *et al.* (2011) found that the most striking feature was the occurrence of a negative residual correlation between neighbouring plots in different rows. This is evidence of competition, as would be a correlation value in row direction that is small, compared to the value in column direction.

Several researchers (BESAG and KEMPTON, 1986; STRINGER and CULLIS, 2002; STRINGER *et al.*, 2011; SILVA and KERR, 2013; HUNT *et al.*, 2013) argued that the interplot competition between plants of different varieties or treatments is one effect that can seriously bias the assessment of varietal performance and thus reduce the accuracy of the genetic prediction.

SILVA and KERR (2013), with a simulated study on genetic forestry trials, found out that ranking genotypes on the basis of predicted breeding values using models that ignore genetic competition does not correlate well with a ranking based on predicted total breeding values, which are a function of both direct and competitive additive effects. This happens because the variances estimated for genetic and residual effects tended to be biased under models that ignored genetic competition.

STRINGER *et al.* (2011) extends GILMOUR *et al.* (1997) to include the Random Treatment Interference Model (R-TIM), which incorporates direct and neighbour genetic effects. They concluded, using autoregressive models, that the competition at the residual level can be modelled by second-order (AR(2)) and if there are both effects, trend and competition effects, the third-order, or the constrained version of this model (SAR(2), notation in BUTLER *et al.* (2009)) can be adequate. STRINGER *et al.* (2011) showed a disagreement between classical approaches and competition models in terms of selection, given that a correlation between predicted values in one trial was 0.76 and in another (presented in supplementary material) was 0.95.

As some agricultural experiments are conducted in more than one harvest and/or site, the joint analysis can be used. CULLIS *et al.* (2000), assessing several experiments from Australia crop variety evaluation programs in different years and locations, advised that the joint analysis is better than individual. However, they did not discuss competition and there are few papers which do this with real data from groups of experiments and/or longitudinal data. Another relevant issue not found in the works quoted is to compare and show how much the joint analysis improves in heritability estimation and assertiveness selection, with this kind of experiment.

Hence, the aim in this chapter is to find which was the best model to select the most productive test lines of sugarcane from five regions of Brazil, where there are two harvests or three sites each, in a systematic design with up to 21% of check plots (commercial varieties). With the chosen individual models, the joint analysis is investigated, evaluating the genetic correlation among sites or year of harvest.

Additionally, simulations are performed to establish that the parameters of the complex models being proposed for sugarcane experiments can be estimated using designs such as those typically used

by “Centro de Tecnologia Canavieira” (CTC) in their experiments, in Brazil. For this simulation study it was used the gain in precision of estimated variance components, heritability or selection gain, and assertiveness of selection (based on 30 best test lines, around the top 7%), of the models from individual and joint analysis.

The chapter is arranged as follows. First, the experiments are described in Section 2.2. Section 2.3 presents the general model for these data, including an extension of competition effects of STRINGER *et al.* (2011) and correlation structures for the residual variation. Section 2.4 gives the results of the individual and joint analysis of all experiments. Section 2.5 and Section 2.6 describe and show the results of the simulation studies, respectively. Some concluding remarks are made in Section 2.7.

2.2 Material

The experimental data were provided by the CTC, Piracicaba - Brazil, in a partnership with the University of São Paulo and are summarized in Table 2.1. The experiments were carried out in five regions in the southeastern and central-west regions of Brazil (see Figure 1.2). These areas were chosen by the company based on climatic and soil conditions. Each region had three sites, called Locals which were numbered, with the same sugarcane lines to test. In the Paraná region there was an exception with one site only that was assessed in two consecutive years. Because of space problems, Araçatuba had multiple experimental areas at each Local.

All experiments had 79% or more of the area planted with new sugarcane lines, denominated test lines. These lines had only one plot in each site due to insufficient material for replication and the large area that would be necessary to accommodate more plots.

The aim of the company project was to select the best seven percent of the test lines in each area. It does this by comparing them with four commercial varieties, which were called checks and were replicated such that they occupied at most 21% of the plots. One special check, nominated interspersed plot, was planted in almost 11% of the plots and it was allocated systematically on a diagonal grid throughout the trial. The other three checks were equally replicated and each replicate was spread out in three neighbouring row plots. The idea behind this systematic arrangement of the check varieties was that it would contribute to capture better the spatial effects and facilitate the visual evaluation and selection of the better lines, because the same column always had eight or fewer test lines between checks.

Table 2.1. Summary of the analyzed experiments.

Region	Site	Number of			Number of plots	
		harvests	columns $\times$ rows	test lines	interspersed	total
Paraná	—	2	26×19	406	55	488
	Local 651	1	10×24	200	26	238
Paraná	Local 851	1	10×24	200	28	240
	Local 852	1	10×24	200	28	240
	Local 20	1	19×28	429	57	516
Ribeirão	Local 72	1	21×25	429	58	517
	Local 140	1	14×38	429	59	519
	Local 54	1	14×37	422	57	506
Piracicaba	Local 58	1	14×37	422	57	506
	Local 76	1	15×35	422	57	506
	Local 3	1	13×20	212	30	260
Goiás	Local 521	1	13×20	212	30	260
	Local 533	1	13×20	212	30	260
		1	3×18	42	6	51
	Local 101	1	15×21	252	35	302
		1	3×15	31	5	39
Araçatuba		1	26×12	259	35	309
	Local 130	1	5×15	62	10	75
		1	9×21	109	19	137
	Local 551	1	11×24	209	29	250

The plots were 12m long, double-furrows with 0.9m between furrows within the plot and 1.5m spacing between different plots. There was 1m of space between columns. Therefore, the plots which were row-neighbours (i.e., within the same column) shared the longest plot boundary.

The unnamed Paraná site was the first experiment studied. It was planted in 2013, there were two cuts, with the first harvest being obtained in 2014 and the second in 2015. The other trials were planted in 2014 and harvested in 2015.

The yields in tonnes of cane per hectare (TCH) from each cut or Local were analyzed individually. Afterwards, the results were used to perform the joint analysis for the experiments from a region. Other variables were measured by CTC, and one problem, related by the company's breeder, was that some plots exhibited lodging. However, they did not have the information about the direction in which plants fell down, only the score of each plot that rated the amount of lodging. The impact of plant lodging will be simulated and reported in Chapter 3.

2.3 Methodology

To provide reliable predictions of test line performance across harvest or environments in analyzing Brazilian sugarcane experiments, we propose to follow GILMOUR *et al.* (1997) and to use mixed models that incorporate terms for local, global and extraneous variation. In addition, as STRINGER *et al.* (2011) and HUNT *et al.* (2013), terms for genetic and residual competition will be included. Then, the general form of the model for $n \times 1$ vector of yields, y (assumed ordered as rows within columns), where n is the total number of plots, can be written as:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\tau} + \mathbf{Z}_g\mathbf{u}_g + \mathbf{Z}_o\mathbf{u}_o + \boldsymbol{\varepsilon}; \quad (2.1)$$

where $\mathbf{X}$ is the incidence matrix associated with vector $\boldsymbol{\tau}$ of fixed effects, $\mathbf{Z}_g$ is the incidence matrix associated with the vector $\mathbf{u}_g$ of random genetic effects of the test lines, $\mathbf{Z}_o$ is the incidence matrix associated with the vector $\mathbf{u}_o$ of non-genetic random effects, and $\boldsymbol{\varepsilon}$ is the vector of residual effects.

We assume that the joint distribution of $(\underset{\sim}{\mathbf{u}}_{\mathbf{g}}, \underset{\sim}{\mathbf{u}}_{\mathbf{o}}, \underset{\sim}{\boldsymbol{\varepsilon}})$ is Gaussian with zero mean and variance matrix

$$\sigma^2 \begin{bmatrix} \mathbf{G}_{\mathbf{g}}(\gamma_{\mathbf{g}}) & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_{\mathbf{o}}(\gamma_{\mathbf{o}}) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{R}(\phi) \end{bmatrix}.$$

For a single experiment, the effects include the global mean, commercial varieties (checks) and linear trend, if any, that is, the effects of the checks, which are generally standard released varieties, are considered fixed, as in SANTOS *et al.* (2002); PIEPHO *et al.* (2008); PETERNELLI *et al.* (2009); PASTINA *et al.* (2012); ALI *et al.* (2013). For the genetic effects of m test lines from a single experiment, we consider (i) random genetic effects only for which $\mathbf{G}_{\mathbf{g}}(\gamma_{\mathbf{g}}) = \gamma_{\mathbf{g}} \mathbf{I}_{\mathbf{m}}$; and (ii) the Random-effects Treatment Interference Model (R-TIM) (Stringer et al, 2011) in which $\underset{\sim}{\mathbf{u}}_{\mathbf{g}}^t = (\underset{\sim}{\mathbf{u}}_{\mathbf{gd}}^t, \underset{\sim}{\mathbf{u}}_{\mathbf{gn}}^t)$ contains the direct and neighbour genetic effects and

$$\mathbf{G}_{\mathbf{g}}(\gamma_{\mathbf{g}}) = \begin{bmatrix} \gamma_{gd} & \gamma_{gdn} \\ \gamma_{gdn} & \gamma_{gn} \end{bmatrix} \otimes \mathbf{I}_{\mathbf{m}};$$

where $\sigma_{gd}^2 = \gamma_{gd}\sigma^2$, $\sigma_{gn}^2 = \gamma_{gn}\sigma^2$ and $\sigma_{gdn} = \gamma_{gdn}\sigma^2$ are the variance components for the direct and neighbour genetic effects, and covariance component between these effects, respectively.

For this model the incidence matrix for the genetic effects is $\mathbf{Z}_{\mathbf{g}} = [\mathbf{Z}_{\mathbf{gd}} \quad \mathbf{N}_{\mathbf{g}} \mathbf{Z}_{\mathbf{gd}}]$, where $\mathbf{Z}_{\mathbf{gd}}$ is the incidence matrix associated with the direct genetic effect and $\mathbf{N}_{\mathbf{g}}$ is a first-order neighbour incidence matrix. For between-row genetic competition, $\mathbf{N}_{\mathbf{g}} = \mathbf{I}_{\mathbf{c}} \otimes \mathbf{N}_{\mathbf{r}}$, where $\mathbf{N}_{\mathbf{r}}$ is the within-row neighbour incidence matrix.

A model for the local spatial variation can be written as $\underset{\sim}{\boldsymbol{\varepsilon}} = \underset{\sim}{\boldsymbol{\eta}} + \underset{\sim}{\boldsymbol{\zeta}}$, where $\underset{\sim}{\boldsymbol{\eta}}$ is a vector of technical or measurement error (nugget effect) and $\underset{\sim}{\boldsymbol{\zeta}}$ is a vector that represents a spatially dependent process, which is usually modeled using a separable process in the row and column directions such as $\mathbf{R} = (\sum_c \otimes \sum_r) + \gamma_{\psi} \mathbf{I}_{\mathbf{n}}$, with $\sum_c$ and $\sum_r$ representing the spatial correlation matrix in the column (c) and row (r) direction (GILMOUR *et al.*, 1997; STEFANOVA *et al.*, 2009; STRINGER *et al.*, 2011).

In terms of the starting model for the model fitting process, the approach of SMITH *et al.* (2005) is followed and a model that includes all the terms that experience tell us are likely to occur in sugarcane experiments. It includes an R-TIM model, an $\text{AR}(1) \times \text{AR}(1)$ spatial model and random row and column effects. Autoregressive models are used initially because often close neighbours are more highly correlated than far neighbours. Here, they will also be compared with banded correlation structures as a check that the pattern in the bands of observed correlations conforms to the more restrictive patterns assumed for alternative correlation structures. The genetic and non-genetic effects are assessed, in order, as follows:

- (i) Neighbour genetic effects, by replacing the R-TIM with independent genetic effects only;
- (ii) Local effects in the column and row directions, by replacing the $\text{AR}(1) \times \text{AR}(1)$ spatial model with $\text{AR}(1) \times \text{Band}(3)$ in the model selected in (i);
- (iii) Management practice effects, such as harvesting direction; sowing direction; cone seeder, by including fixed effects for these in the selected model in (ii);
- (iv) Global variation, by splitting the column and row global effects in the model selected in (iii) into cubic smoothing spline and non-smooth deviations;
- (v) Retest the neighbour genetic effect in the context of model (iv).

The models were compared formally, using a REML ratio test, for selecting between nested random models, or Wald F tests, for the fixed effects, assuming a 5% significance level. Each selected model is assessed using diagnostic plots, including plots of faces of the empirical sample variogram and/or the empirical semi-variogram (STEFANOVA *et al.*, 2009).

The models from the individual analysis are used to formulate the model for the joint analysis, but with the inclusion of cut or site main effects and the interaction of cut or site with the genetic effects.

Beginning with the model for independent direct genetic effects (id), the covariance structures shown in Table 2.2 are tested.

Table 2.2. Some possible covariance structures for direct genetic effects over cut or sites (i).

	Structures	parameters
idh()	heterogeneous identity	γ_{gi}
corg()	general correlation	γ_{gi} and ρ_{gi}
cor()	uniform correlation	γ_g and ρ_g
id()	identity	γ_g

All computations were done in the R software (R CORE TEAM, 2014) using the ASReml-R (BUTLER *et al.*, 2009) and asremlPlus (BRIEN, 2016) packages.

2.4 Data analysis results

Each individual analysis showed the peculiarities of each site within the region. In Table 2.3, the selected individual model is shown and Table 2.4 displays the REML estimates of variance parameters for these models. In most trials there were no significant global or local effects in the column direction. However, in Piracicaba Local 58 and Goiás Local 3 there were negative correlations in this direction which STRINGER (2006) also found in sugarcane experiments. However, competition does not seem a plausible biological explanation for such correlation in the column direction, in contrast to the row direction. We believe that, in our case, it perhaps was because material from different plots become mixed up as a result of lodging so that separation of the material from different plots at harvest became difficult. Also, the correlation in row direction, when it was significant, normally did not conform to an autoregressive process as suggested by some authors (GILMOUR *et al.*, 1997; STRINGER *et al.*, 2011). We compared autoregressive with banded models and out of 11 experiments, eight were modeled more realistically using a banded correlation structure. In all of these 11 experiments there were significant row correlations and because, except for Paraná Local 651, the first neighbour correlation was negative or smaller than the second neighbour correlation, it was concluded that there was competition at the residual level. For Paraná Local 651 the magnitude of the first neighbour correlation was similar to the column correlation so that it is concluded that there is not competition at the residual level at this site.

Only in Goiás Local 521 there was evidence of significant genetic competition, with $\gamma_{gd} = 3.44$. This value was more than three times the ratios found in the other two sites at the same region. Consequently, we believe that this is most likely a spurious result.

In the Araçatuba experiments, the direct genetic effects were close to zero or the standard errors (s.e) for the estimated components were high when compared with the magnitude of the parameters (Local 101-1: s.e genetic= 111.61 and s.e residual= 72.66. Local 130-2: s.e genetic= 123.19 and s.e residual= 105.53). Therefore, in this case, two explanations were proposed. First, in every Local there was at least one experiment with less than 10 rows, which makes it difficult to estimate spatial dependence, and/or less than 100 plots which limits the ability to estimate variances. Second, because even the larger experiments performed poorly, it could be that the small areas are inherently more variable or there were greater difficulties encountered in carrying out the operations for the experiment. The improvement of this type of trials will require more uniform areas and/or more careful performance of the operations during the experiment.

One problem found using ASReml-R is that the structure Band() $\times$ Band() is not available and, for this reason, in one direction we still used the autoregressive process. Another is that, for the models using spl(), the response variance matrix cannot be written with this effect to obtain the plot of the empirical variogram using asremlPlus. Hence, in this case, the empirical semi-variogram, from ASReml-R, was used.

Table 2.3. Summary of the selected models for all experiments. All the models have a fixed check varieties effects and random effects for test lines denominated as the direct genetic effects (G). The other effects are represented with the symbols: Cext (co-variate, external effect in the 3 east columns); H (harvest); C (column); R (row); N (neighbour genetic); spl(.) (spline in some direction indicated in parentheses) and lin(.) (linear trend in some direction indicated in parentheses). The structures are US(.) - unstructured; AR(1) - first-order autoregressive, Band(.) - banded correlation, the number indicates the order of correlation, SAR(2) - constrained autoregressive and Id - identity.

Experiments		Fixed effects	Random effects	Residual variation	
		Global/extraneous	Genetic	Local	
				column $\times$ row	
Paraná	1st. cut	Cext	Row	G	Id $\times$ Band(4)
	2nd. cut	Cext+H+lin(C)+lin(R)	spl(C)+spl(R)	G	Id $\times$ Band(3)
	651		nugget	G	AR(1) $\times$ AR(1)
	851	lin(C)	spl(C)+R	G	Id $\times$ Id
	852	lin(C)+lin(R)	spl(R)	G	Id $\times$ Band(2)
Ribeirão	20	lin(C)+lin(R)	R	G	Id $\times$ Band(1)
	72	direct genetic effects ($\tilde{\sigma}_g^2=14.27$) less than its standard error (79.92).			
	140	lin(C)+lin(R)	spl(R)	G	AR(1) $\times$ Band(1)
Piracicaba	54	lin(C)	R	G	Id $\times$ Band(1)
	58	lin(R)	spl(R)+R	G	Band(2) $\times$ SAR(2)
	76		R	G	Id $\times$ Band(2)
	3	H+lin(R)	C+spl(R)	G	AR(1) $\times$ Band(2)
Goiás	521	lin(C)+lin(R)	C+spl(R)	US(G:N)	Id $\times$ Id
	533	lin(R)	spl(R)	G	Id $\times$ AR(1)
	101 -1	lin(C)		G	Id $\times$ Id
Araçatuba	101 -2	direct genetic effects close to zero			
	101 -3	direct genetic effects close to zero			
	130 -1	direct genetic effects close to zero			
	130 -2		G		Id $\times$ Id
	551 -1	direct genetic effects ($\tilde{\sigma}_g^2=95.40$) less than its standard error (141.74).			
	551 -2	direct genetic effects close to zero			

At the same region, the joint analyses presented a significant estimated correlation between direct genetic effects in all cases. Table 2.5 compares the correlations between direct genetic effects from the joint analyses with the calculated correlations of the direct genetic EBLUPs from the individual analyses. In all cases the correlations from the joint analyses are around twice, or more, than those from the corresponding individual analysis. For Paraná first and second cuts, the 7% best test lines selected using the direct genetic effects from the joint analysis were the same in both cuts, because the correlation between the direct genetic effects was approximately one. For the experiments at different sites within a region, the selected group of test lines using the genetic effects from the joint analysis were rather more similar than when the genetic effects from the individual analyses were used. Cullis et al. (2000), in examining the efficiency of Australian crop variety evaluation programs, concluded that the effect of increasing the number of locations is much less than that of increasing the number of years.

Table 2.5. Correlations (a) between the direct genetic EBLUPs from each fitted model of the individual analysis; (b) between the direct genetic effects from the joint analysis.

Region	Cuts	(a)			(b)		
		1st,2nd			1st,2nd		
Paraná		0.54	—	—	1.00	—	—
	Sites	1st,2nd	1st,3rd	2nd,3rd	1st,2nd	1st,3rd	2nd,3rd
Paraná		0.37	0.33	0.49	0.84	0.84	0.91
Ribeirão		—	0.40	—	—	0.82	—
Piracicaba		0.31	0.30	0.28	0.84	0.84	0.84
Goiás		0.17	-0.02	0.03	—	0.39	—

2.5 Simulation studies

Two simulation studies were conducted. The first involved a single-site experiment that employed an unreplicated design with systematic checks and involved a model for simulation that always included genetic competition. It sought to investigate the effect on the selection of test lines using individual analyses of (i) the magnitude of the direct genetic variation, (ii) the presence or absence of non-genetic competition, and (iii) the model fitted. The second simulation study involved a three-site experiment that used an unreplicated design with systematic checks at each site, but did not include competition at the genetic level. In this case, the study investigated the effect on the selection of test lines of (i) the magnitude of between-sites genetic correlation, (ii) unequal versus equal site variances, (iii) fitting autoregressive versus banded correlation structures, and (iv) individual versus joint analyses of the sites. It was also of interest in these studies to see how well the variance components are estimated with the type of unreplicated design used by CTC.

In all cases, 1000 data sets were simulated for a design grid of 25 columns by 20 rows, totaling 500 plots. An unreplicated design of the type employed by CTC was used to assign 418 lines to this grid. The 418 lines consisted of 414 unreplicated test lines and four check lines; three of the check lines were replicated 10 times each and the fourth had 56 replicates laid out diagonally every eight columns and eight rows; the test lines were randomized to the plots remaining after the check lines had been assigned.

2.5.1 Single site experiments

The model for simulation had: (i) global column (γ_c) and row (γ_r) variation; (ii) direct (γ_{gd}) and neighbour (γ_{gn}) genetic variation, as well as covariance (γ_{gdn}) between these two sources of variation; and (iii) residual correlation with AR(1) in the column (ϕ_c) direction and Band(1) in the row (ϕ_r) direction.

Four different scenarios for simulations were conducted. For all scenarios $\gamma_c = \gamma_r = 0.3$, $\gamma_{gdn} = 0.3$, $\gamma_{gn} = 0.1$, $\phi_c = 0.5$ and the scale parameter, σ^2 , was set at a value of 1.0. The values of γ_{gd} and ϕ_r were varied to yield the different simulation scenarios:

- a. $\gamma_{gd} = 1$ and $\phi_r = -0.3$;
- b. $\gamma_{gd} = 1$ and $\phi_r = 0.3$;
- c. $\gamma_{gd} = 1.5$ and $\phi_r = -0.3$;
- d. $\gamma_{gd} = 1.5$ and $\phi_r = 0.3$.

Thus the simulations differed in the magnitude of the direct genetic variation and in whether or not there was competition at the residual level.

In generating the data for each scenario a single set of direct and neighbour genetic effects was generated for that using the values of the genetic variance parameters for the simulation. This set of effects constituted the true genetic effects for the simulation. They were added to each of the 1000 simulated data sets obtained for a model with all the non-genetic parameters included.

For each simulation, the data were analyzed using four models:

Model 1 only random column, row and direct genetic terms;

Model 2 Model 1 plus neighbour genetic and direct-neighbour genetic covariance terms;

Model 3 Model 1 plus residual correlation;

Model 4 Model 2 plus residual correlation.

Clearly, Model 1 is the simplest model, while Model 4 is the same as the simulation model.

For each simulation and each analysis model, the following information was collected:

- the number of converged analyses;
- the correlation between the direct genetic EBLUPs and the true (generated) direct genetic effects, at the same model and between pairs of models;
- the selection gain measured as the average of the true direct genetic effects for the 30 test lines (around 7%) with the largest estimated direct genetic EBLUPs;
- the assertiveness of selection or average similarity where similarity is measured as the percentage of the 30 test lines with the largest direct genetic EBLUPs that are in the true 30 best test lines; the true 30 best test lines are the 30 test lines with the highest true (generated) direct genetic effects;
- the estimates of the variance parameters.

2.5.2 Separate versus joint analyses

This study was based on an experiment with three sites, as in a CTC experiment in one region. Each site employed the same unreplicated design with systematic checks, although test lines were rerandomized at each site. In all simulations, all sites assumed the same direct genetic variance ($\gamma_{gd} = 0.8$, which is approximately the average for the three-site CTC experiments) and an AR(1) residual correlation in the column direction ($\phi_c = 0.2$). Further, the assumed structure for the residual row correlation ($\sum_{ri}$, $i = 1 \dots 3$) differed between sites, then $\mathbf{R} = \text{diag}(\sigma_1^2 \sum_c \otimes \sum_{r1}, \sigma_2^2 \sum_c \otimes \sum_{r2}, \sigma_3^2 \sum_c \otimes \sum_{r3})$. At Local 1 it was assumed to be SAR(2) ($\phi_{r11} = 0.6$ and $\phi_{r12} = -0.2$), at Local 2 to be AR(1) ($\phi_{r21} = 0.4$), and

at Local 3 to be AR(2) ($\phi_{r31} = -0.2$ and $\phi_{r32} = 0.1$). Hence, Locals 1 and 3 were assumed to have non-genetic competition of different types and Local 2 to not have residual competition.

Four simulation scenarios were conducted in which the sites were assumed to have either equal or unequal residual variance (σ_i^2) and one of two values for the genetic correlation between sites (ρ_{gs}). The values of the parameters used in the simulations were:

- a. $\rho_{gs} = 0.70$, $\sigma_1^2 = 1.15$, $\sigma_2^2 = 1.30$, and $\sigma_3^2 = 1.00$;
- b. $\rho_{gs} = 0.70$ and $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = 1.00$;
- c. $\rho_{gs} = 0.90$, $\sigma_1^2 = 1.15$, $\sigma_2^2 = 1.30$, and $\sigma_3^2 = 1.00$;
- d. $\rho_{gs} = 0.90$ and $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = 1.00$.

In generating the data for each scenario a set of direct genetic effects was generated for the three sites for that simulation scenario using the values of the genetic variance parameters. This set of effects constituted the true genetic effects for the simulation scenario. Then, 1000 simulated data sets for the whole experiment were obtained for a model that included all the non-genetic parameters. Finally, the set of true genetic effects was added to each of the 1000 simulated data sets.

Two joint analyses was performed on each of the 1000 data sets for each scenario. The two mixed models used included all the terms in the model for simulation, but differed in their assumed residual row correlation structures: (i) as in the simulated data, the autoregressive structures SAR(2), AR(1), and AR(2) used for Locals 1, 2 and 3, respectively; and (ii) the banded structures Band(3), Band(2) and Band(1) used for Locals 1, 2 and 3, respectively.

For each scenario and each analysis model, the following information was collected:

- the number of converged analyses;
- the correlations between the direct genetic EBLUPs and the true direct genetic effects for each site obtained from the individual and joint analyses;
- the assertiveness of selection or the average similarity for each site obtained from the individual and joint analyses, where similarity is measured as the percentage of the 30 test lines with the largest direct genetic EBLUPs that are in the true 30 best test lines; the true 30 best test lines are the 30 test lines with the highest true (generated) direct genetic effects ;
- the heritability estimates (as proposed by CULLIS *et al.* (2006));
- the estimates of the direct genetic variance parameter.

2.6 Simulation results

2.6.1 Single site experiments

The estimation process for the simplest model converged without problems for all simulated data sets. However, when the number of parameters increased convergence was not obtained in some of the cases, see Table 2.6. Note that for the fitted of the true model, Model 4, the algorithm converged between 470 to 687 times in each scenario and also Model 2 had a smaller convergence rate in some cases. Therefore, with this data, fitting the models with competition genetic effects resulted in more problems than the simplest model or the model with residual correlation (Model 3). Observe that scenario b Model 2 converged only 90 times. This number is very small and the results were not used to compare with the other models.

Table 2.6. Number of convergence cases in each model and simulation scenario for single site experiments, given 1000 data sets.

		Simulation scenario			
models		a	b	c	d
	1	1000	1000	1000	1000
	2	789	90	749	343
	3	997	682	997	771
	4	470	578	530	687

Assessing the correlation of the direct genetic EBLUPs between pairs of models in each scenario, the average correlations were superior to 0.80 (see Figure 2.1) although the correlations were better when there was negative row correlation rather than positive row correlation. This means that on average there were small differences between the models.

Figure 2.2 illustrates the correlations between true simulated direct genetic effects and the direct genetic EBLUPs. The results were around 0.65 to 0.80 which means that the estimates are not that close to the true values. Also, there was little difference in the EBLUPs among the models. Although the correlation for Model 4 was a little better in all scenarios. Assessing the selection gain, related to the average EBLUPs of the 30th best test lines in each model and scenario, the results from Model 4 were better, see Figure 2.3. Also, the best selections came from Model 4, which selected in average around 44 to 51% of the true 30 best test lines (Figure 2.4). In general, the greater the genetic variance component the greater will be the selection gain and assertiveness. Note that there are some differences between the results from scenarios a and b with c and d, because of the direct genetic effects. Moreover, these results do not represent a big difference among the models from the genetic criteria (selection gain and similarity), because of the magnitude of the parameters and maybe the convergence problems when fitting the models.

Figure 2.1. Boxplots of the correlations between direct genetic EBLUPs for pairs of models from each scenario (a, b, c and d). In (b) Model 2 converged only for 90 out of 1000 simulated data sets and hence the density plot for genetic components effects are not displayed (see Table 2.6).

Figure 2.2. Boxplots of the correlation between the true direct genetic effects and the respective EBLUP's from each model for each scenario (a, b, c and d). In (b) Model 2 converged only for 90 out of 1000 simulated data sets and hence the density plot for genetic components effects are not displayed (see Table 2.6).

Figure 2.3. Densities of the selection gain from each model for each scenario (a, b, c and d). In (b) Model 2 converged only for 90 out of 1000 simulated data sets and hence the density plot for genetic components effects are not displayed (see Table 2.6).

Figure 2.4. Percentage of average similarity of the 30 test lines selected as best compared to the true best 30 lines in each model and scenario (a, b, c and d). In scenario b, Model 2 converged only for 90 out of 1000 simulated data sets and hence the density plot for genetic components effects are not displayed (see Table 2.6).

CULLIS *et al.* (1998), with barley and wheat experiments, affirmed that models with complex variance structure can affect estimation of the genetic merit of breeding lines. However, for the simulation study, Model 4 was not strongly affected. Figure 2.5 displays the densities of the genetic estimates of the genetic parameters for all scenarios and models (exception scenario b Model 2). Each plot represents the results for a combination a scenario and a component. Note that only Models 2 and 4 assumed genetic covariance and neighbour components. It can be confirmed that with these simulated models, it is possible to obtain the competition genetic effect with this design. Model 4 had some deviation in relation to the true parameters, but it still displayed the best results, when the model algorithm converged, compared with the other models. Investigating the residual and global row and column variance parameters, as well as the row and column correlation (Figures 2.6 and 2.7), this true model presents a better fit than Models 2 and 3. However, overall Model 1 tends to give the best estimates of the row and column variance components (Figure 2.7), whereas Model 4 tends to underestimate them.

Figure 2.5. Densities of the direct genetic (first column of plots), neighbour genetic (third) components along with their covariance (second column) for each scenario and model. Each row represents a different scenario and the vertical lines represent the assumed values of the parameters. Remember that the convergence was different in each one (see Table 2.6) and only Models 2 and 4 assumed neighbour genetic effect. In scenario b Model 2 converged only for 90 out of 1000 simulated data sets and hence the density plot for genetic components effects are not displayed.

Figure 2.6. Densities of the residual component (first column of plots) and the column (second column) and row (third column) correlations for each scenario and model. Each row represents a different scenario and the vertical lines represent the assumed values of the parameters. Remember that the convergence was different in each one (see Table 2.6) and only Model 3 and 4 assumed residual correlation. In scenario b Model 2 converged only for 90 out of 1000 simulated data sets and hence the density plot for genetic components effects are not displayed.

Figure 2.7. Densities of the global components, column (in the left) and row (in the right), for each scenario and model. Each row represents a different scenario and the vertical lines represent the assumed values of the parameters. Remember that the convergence was different in each one (see Table 2.6). In scenario b Model 2 converged only for 90 out of 1000 simulated data sets and hence the density plot for genetic components effects are not displayed.

2.6.2 Separate versus joint analyses

There was no big difference between the number of simulated data set and the convergence models (Table 2.7). However, the convergence rates were slightly smaller for the banded models, mainly in the joint analysis. This was not good because better predictions were expected from the joint analysis, but the convergence can be obtained with the update of the model or by increasing the number of iterations.

Table 2.7. Number of convergence cases in each model and scenario for separate versus joint analyses, given 1000 data sets.

Scenarios		Models			
		Local 1	Local 2	Local 3	Joint
autoreg.	a	1000	1000	1000	999
	b	997	1000	1000	998
	c	998	1000	1000	999
	d	998	1000	1000	999
banded	a	999	999	1000	962
	b	999	998	999	931
	c	999	1000	1000	951
	d	999	997	1000	914

As the results from autoregressive and banded models were similar, for the following criteria we present only the results for the autoregressive models.

Examining the average genetic EBLUP correlations between pairs of Local in the individual and joint analysis (Figure 2.8(a)), those for the joint analyses are twice those for the individual analyses. Since the true values of this parameter is either 0.7 or 0.9, it is clear that the individual analyses underestimate this correlation and that the joint analyses tend to overestimate it a little. An unexpected result was the strong correlation for average genetic EBLUPs between individual and joint analysis at the same Local (Figure 2.8(b)). These were a little bit more for scenario with small genetic correlation component (0.70, scenarios a and b) than the others (0.90, scenarios c and d), and means that there were no strong differences among genetic EBLUPs from individual and joint analysis. On the other hand, the assertiveness of selection, reported in Figure 2.9, showed differences of approximately 6 to 12% among the models from individual and joint analysis at the same Local. Regarding the number of test lines, this represented a small difference, however all improvements in the selection are welcome.

Figure 2.8. Average EBLUP correlations in the individual (light symbols) and joint (dark symbols) analysis for autoregressive models in each scenario. The panels refer to comparison between (a) pairs of Locals; (b) individual and joint analysis at the same Local.

Figure 2.9. Average similarity between the true and estimated 30 best test lines (around top 7%). The dark symbols represent the result from joint analysis and the light are the individuals for autoregressive models in each scenario.

For the joint analysis from autoregressive models, the densities of the heritability for each scenario are presented in Figure 2.10. There are differences among Locals when the scenario assumed different residual variances between sites (scenarios a and c) and a little bit more heritability from simulated scenarios which set out the larger genetic correlation ($\gamma_{gs} = 0.9$ - scenarios c and d). For instance, the average heritability estimates were around 0.6 and 0.7.

Figure 2.10. Densities of the heritability for each site from the joint analysis for autoregressive models in each scenario. The panels refer to the scenarios.

The densities of the direct genetic components for all scenario did not differ between Locals and/or correlation structures from fitting autoregressive and banded correlation structures, but between

the joint and individual analysis there were differences, (see Figure 2.11). The estimated values were more accurate for the joint analysis, independently of residual variance or genetic correlation parameters.

Figure 2.11. Densities of the direct genetic components from autoregressive and banded models for each Local from scenarios (a, b, c and d). The vertical lines represent the assumed values of the parameters.

When investigating the residual variance components (Figures 2.12, 2.13, 2.14 and 2.15) it can be noted that there was no strong difference between the estimated variance components from fitting autoregressive and banded correlation structures. However, the estimates of the residual variance components showed greater spread, and so are less accurate, for the individual analyses than for the joint analyses. The difference seemed greater when the genetic correlation between sites is higher (scenarios c and d).

Figure 2.12. Densities of the residual variance components from autoregressive and banded models for each Local from scenario a. The vertical lines represent the assumed values of the parameters.

Figure 2.13. Densities of the residual variance components from autoregressive and banded models for each Local from scenario b. The vertical lines represent the assumed values of the parameters.

Figure 2.14. Densities of the residual variance components from autoregressive and banded models for each Local from scenario c. The vertical lines represent the assumed values of the parameters.

Figure 2.15. Densities of the residual variance components from autoregressive and banded models for each Local from scenario d. The vertical lines represent the assumed values of the parameters.

2.7 Conclusions

Like SMITH *et al.* (2005), the strategy employed for selecting a mixed model for the data from a single site was to start with a model that includes all the terms that experience tells us are likely to occur in the experiment. Then diagnostic plots were used to assess model adequacy and the need to vary the model. Also, we advocate the fitting of banded correlation structures with about three bands as a diagnostic technique for checking that the correlation structure conforms to the assumed pattern for autoregressive models such as AR(1), AR(2) and SAR(2). If it was found that the autoregressive

models do not describe the pattern then banded correlation models were used. In addition to modelling the correlation structure, the R-TIM model for genetic competition was considered and models for global and extraneous variation were tested. A joint analysis of the sites within a region was conducted, the model for the spatial variation being a combination of those selected for the individual analyses.

Of the 21 experiments investigated, seven had no significant direct genetic effects (component close to zero and/or standard error larger than the estimate). For the others, in 11 there was significant competition, but at the residual level rather than at the genetic level. Each individual trial had its own peculiarity as far as the residual variance structure is concerned and so it was impossible to determine one specific model for the residual variation that applied to all the experiments. Only one experiment had the competition at the genetic level significant. It would appear that there is more general competition in Brazilian experiments rather than it being genotype specific.

The test lines selected within a region differed between sites and the genetic EBLUP correlations were less than 0.54. However, the genetic correlations were significant in the joint analyses and these correlations were stronger than in the individual analyses. Hence, using the joint analyses, better accuracy in the selection of the test lines was obtained. Higher predicted values were found with lower standard errors because the observation of genotypes at multiple sites provides replication of them. Thus a joint analysis is preferable to individual analyses whenever it is possible, whether it is based on different sites or more than one harvest from the same trial. As CULLIS *et al.* (2000) have already shown, using models suggested by GILMOUR *et al.* (1997), the effect of increasing the number of locations is much less efficient than increasing the number of years for crops such as barley and wheat. However, in our case, the Paraná trial with two cuts resulted in an estimated correlation between the direct genetic effects (ρ_{gd}), obtained from the joint analysis, that was close to 1.00 (see Table 2.5). It remains to establish whether this results in more efficient selection of lines.

The simulation study for single sites showed that, when there is genetic competition, it is difficult to fit a model that incorporates a term for it. When there is also residual competition, such that the row autocorrelation is negative, the true model will fit in only about 50% of cases, but a model without genetic competition will fit almost surely. When there is no residual competition, a model with genetic competition will fit around 60% of the time, but the only model that is guaranteed to fit is one with no genetic competition term and no residual correlation. The simulation study demonstrated clearly that a strategy of fitting a single model to the data from single sites or cuts, even if it is the true model, is not a viable strategy and that detecting genetic competition in an experiment using the CTC type of design is fraught. Hence, it cannot be stated conclusively that genetic competition does not occur in Brazilian sugarcane experiments.

Fortunately, it would appear that the estimation of the direct genetic effects is little affected by the model selected and the presence or not of residual competition, although it is affected by the magnitude of the genetic variance. While, in general, Model 4 gives the best correlation between true and estimated genetic effects, selection gain and average similarity, the differences between it and the other models are small.

The results of the simulation study for separate versus joint analyses shows there were no problems in fitting either single-site or joint models when there is no genetic competition and that better estimates of the genetic effects come from joint analyses. There was little effect of heterogeneous variances, but there were differences in the results that depended on the magnitude of the genetic variance. For the models simulated, there was little difference between fitting autocorrelation and banded correlation structures.

Given the difficulties in fitting models that include both genetic and residual competition and that only around 45 to 55% of the true best test lines were selected even for the joint analyses, more research into improved design and management practices is necessary, to see if model fitting can be

improved and to increase the direct genetic variance so as to achieve better selection results.

3 NEGATIVE RESIDUAL CORRELATION IN SUGARCANE EXPERIMENTS

The negative residual correlation in the row direction in sugarcane experiments is associated with competition effects between plots in different rows and is taken to result from the narrowness of the rows. On the other hand, negative correlation between columns is generally not observed. However, when such correlation occurs in an experiment then the question arises as to what might have caused it. It was conjectured that one possibility is that lodging, which occurs in dry weather, could lead to negative correlation in the column direction. Our aim in this chapter is to show, with a simulated study, that this is the case. Then, careful harvest processing is necessary to minimize the effect of lodging on the recorded results and avoid confusing it with competition effects.

Keywords: Competition, Harvest process, Lodging, Simulation study

3.1 Introduction

Sugarcane is generally large, perennial, tropical or subtropical grasses that evolved under conditions of high sunlight, high temperatures, and large quantities of water (MOORE *et al.*, 2013). During the production year, the plant can be around 2 meters high. It is not surprising, that under climatic effects (mainly strong winds and/or dry weather), the sugarcane can lose its erectness (SINGH *et al.*, 2002; VAN HEERDEN *et al.*, 2015).

These effects were investigated by some authors (SINGH *et al.*, 2002; VAN HEERDEN *et al.*, 2011, 2015), which have shown a decrease in yield under some conditions of lodging. As a result, SINGH *et al.* (2002) affirmed that the impact of lodging on cane yield tends to be masked given that the better grown crops of sugarcane are more likely to lodge than poorly grown crops. VAN HEERDEN *et al.* (2015) declared that the lodging close to harvest had no seriously deleterious effects on yields.

STRINGER *et al.* (2011), working with models assuming competition at the residual level, showed that this biological effect results in negative or smaller correlation for the first neighbour than the second in one direction, because the plots are long and thin. Hence the idea of competition in both directions is not expected. Moreover, STRINGER and CULLIS (2002) and STRINGER (2006) found significant negative correlation for both directions working with autoregressive models in some sugarcane experiments. The magnitude of the correlation was smaller for the direction that shared a smaller boundary and STRINGER and CULLIS (2002) affirmed that a special plant-breeding program, which results in a higher competition effect, was the dry one and a particular trial there did not receive irrigation. Then, both competition and lodging effects can be related to dry weather, which causes plant stress.

We noted, in some breeding experiments conducted by the “Centro de Tecnologia Canavieira” (CTC - Brazil), that sometimes the lodged sugarcane fell down over adjacent plots, causing serious problems in separating the material from different plots at harvest. Given this and that the process is mechanical, it is possible that part of lodged plot yield was ascribed to its neighbouring plot. It was also noted that in the experiments there was significant negative residual correlation in the column direction. We suspect that, even in the absence of competition, lodging could lead to negative correlation in the column direction. Thus, our aim in this chapter is to prove that the negative residual correlation in column direction can be associated with inaccurate mechanical harvest that might occur when plants are lodged in the row direction.

The chapter is arranged as follows. First, we describe the motivating example (Section 3.2). Section 3.3 presents the material and methods. Section 3.4 gives the results and section 3.5 some concluding remarks.

3.2 Motivating example

Our motivation came from a partnership between the CTC and the University of São Paulo (Brazil) where a series of sugarcane breeding experiments were provided to facilitate the development of new models for the company and to understand the “real” experimental effects. One particular experiment, on which the lodging score was recorded, had 5.4% of plots with score 9 and 10.7% with score 8. These represent high levels of lodging and it means that in these plots most of the plant were either prostrate or nearly so.

The selected model for this experiment included harvest, row spline and random column effects, with negative residual correlation in both directions. This was not expected given that the plot dimensions were 12m long, double-furrows with 0.9m between furrows within the plot and 1.5m between different plots, and 1m of space between columns.

Then, if a plot with 2m high cane became extensively lodged, it can be expected that 1m of some stalks came into the neighbouring plot. It has been noted that this can result in the plot borders becoming obscured and making it difficult to determine during the harvesting the plot to which some of the cane belongs. This is particularly a problem for the column borders.

Other experiments, with the same design structure, also resulted in negative correlation in both directions, however lodging scores were not available.

3.3 Material and methods

We begin by assuming that, at the harvest, lodging results in a loss of tonnes of cane per hectare (TCH) are lost to a neighbouring plot in the same row. Two cases were assumed, one with a 5% loss of TCH and the second with 15%. In addition, four percentages of lodged plots for the experiments were studied: 0%, 1%, 5% and 10%.

The design array has 25 columns by 20 rows, totaling 500 plots, with 414 unreplicated test lines and four checks. Three checks were replicated 10 times each and one 56 times in a diagonal grid, each eight columns or rows, respecting the similar characteristics of the original design from CTC. In practice, the checks (commercial varieties) are stable and erect, so the lodging only occurs with test line plots.

The model for the data vector, $\mathbf{Y}$, is:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\tau} + \mathbf{Z}_g\mathbf{u}_g + \boldsymbol{\varepsilon}; \quad (3.1)$$

given that: $\mathbf{X}$ is the incidence matrix associated with vector $\boldsymbol{\tau}$ of fixed effects; $\mathbf{Z}_g$ is the incidence matrix associated with the vector $\mathbf{u}_g$ of random genetic effects of the test lines; $\boldsymbol{\varepsilon}$ is the vector of residual effects.

We assume that the joint distribution of $(\mathbf{u}_g, \boldsymbol{\varepsilon})$ is Gaussian with zero mean and variance matrix

$$\sigma^2 \begin{bmatrix} \mathbf{G}_g(\gamma_g) & \mathbf{0} \\ \mathbf{0} & \mathbf{R}(\phi) \end{bmatrix} = \begin{bmatrix} \mathbf{G}_g & \mathbf{0} \\ \mathbf{0} & \mathbf{R} \end{bmatrix}.$$

where, $\mathbf{G}_g(\gamma_g) = \gamma_g \mathbf{I}_m$ and $\mathbf{R} = \sum_c \otimes \sum_r = \text{AR}(1) \otimes \text{Band}(1)$, assuming $\mathbf{Y}$ are in field order with rows within columns. $\gamma_g = \sigma_g^2 / \sigma^2$, $\mathbf{I}_m$ is the identity matrix of order m , such as m is the number of test lines, $\text{AR}(1)$ indicates a first-order autoregressive correlation matrix and $\text{Band}(1)$ the first-order banded correlation.

For all simulations $\gamma_g = 1$, $\rho_r = -0.3$ and the scale parameter, σ^2 , was set at a value of 1.0. Two types of data sets are provided assuming: (a) $\rho_c = 0.5$ and (b) $\rho_c = 0$. For each, 1000 data sets were generated. The data set combined with the percentage of lodged plants result in the following four scenarios.

- 1) $\rho_c = 0.5$ and 5% of TCH lost;

- 2) $\rho_c = 0.5$ and 15% of TCH lost;
- 3) $\rho_c = 0$ and 5% of TCH lost;
- 4) $\rho_c = 0$ and 15% of TCH lost;

The density plots of the residual variance parameters were used to show how the lodging affects the estimation. All computations were done in the R software (R CORE TEAM, 2014) using the ASReml-R package (BUTLER *et al.*, 2009).

3.4 Results

In most of the cases the algorithm converged (Table 3.1). An exception was for 1% lodging, a loss of 15% and $\rho_c = 0.5$, when nearly half of the plots failed to converge.

However, observing the densities of the residual components (Figure 3.1), the residual variance increases as the percentage of plots lodged increases, but only for a 5% loss of TCH. In general, the column correlation becomes increasingly more negative as the percentage of plots that are lodged increases. Note that the row correlation is also affected by the lodging of plots.

Table 3.1. Convergence number of the lodging models algorithms in each scenario and percentage of plots lodged, given 1000 simulated data sets.

				Percentage of plots lodged			
				0%	1%	5%	10%
scenario	1	$\rho_c = 0.5$ and 5% of TCH lost	986	943	1000	1000	1000
	2	$\rho_c = 0.5$ and 15% of TCH lost		552	984	973	973
	3	$\rho_c = 0$ and 5% of TCH lost	984	982	1000	997	997
	4	$\rho_c = 0$ and 15% of TCH lost		923	989	997	997

Figure 3.1. Densities of the residual component and the column and row correlation for each scenario. The curves within a panel are for the different percentage of plots lodged. The plot of the residual components have been truncated to exclude a few estimated values that were in excess of 4.

3.5 Conclusions

In summary, the negative correlation can be due to the lodging effect in the experiment if this results in material from one plot being associated with a plot from a neighbouring column at the harvest. As the dry weather can also cause more competition between plots, the negative correlation can be related to either one or both of these effects. Hence, when there is lodging care in separating material from different plots will lead to improved results.

4 COMPARISON OF A SYSTEMATIC DESIGN WITH SOME SPATIALLY OPTIMIZED DESIGNS

This chapter considers a trial of fixed size, similar to those used by the “Centro de Tecnologia Canavieira” (CTC), a Brazilian sugarcane company. The efficiency of the type of unreplicated design that they typically employ is compared to spatially optimized unreplicated and p -rep designs with checks and a spatially optimized p -rep design. A simulation study is conducted to compare the designs as far as the genetic gain from selection is concerned. The comparison is done by using the prediction error variances (PEVs), computed from the simulated data, to compute relative efficiencies of the designs to a CTC design. Other criteria calculated to evaluate performance are the relative genetic gain, the correlation between the true and estimated genetic effects and the similarity percentage that gives the percentage of the test lines that are in the top 7% of both the true and estimated genetic effects. The results show that differences in the number of check varieties have little effect on the performance of unreplicated designs with checks. On the other hand, experiments that employ a p -rep designs are better than using unreplicated designs.

Keywords: Unreplicated design, Optimal design, p -rep design, Prediction error variances, Genetic gain, Assertiveness of selection, Mixed models, Simulation study

4.1 Introduction

Breeding companies continuously pursue lower production costs and better crop yield. To achieve this, new varieties are recommended from several breeding trials with potential genotypes; these recommended varieties become the test lines in subsequent experiments. However, at the initial phases of the breeding program there are many test lines and limited material for planting in the experiments. SMITH *et al.* (2006) pointed out that, in such circumstance, fully replicated designs are prohibitively expensive and are unnecessary from a statistical perspective.

One solution, discussed by KEMPTON (1984), is to employ unreplicated designs in which there are unreplicated test lines and replicated commercial varieties, called checks. CULLIS *et al.* (2006) call these designs grid-plot designs. As CLARKE and STEFANOVA (2011) describe, there are two types of unreplicated designs: (i) designs in which the replicates of the check are systematically allocated and (ii) augmented designs (FEDERER, 1956) in which a standard randomized design (e.g. randomized complete-block design, Latin square design) is used to allocate check lines and the test lines are added to this design.

A problem is, given there may be global, local and/or extraneous variation (GILMOUR *et al.*, 1997; STEFANOVA *et al.*, 2009), the precision with which parameters can be estimated may depend on the spatial distribution of the check lines plots (MÜLLER *et al.*, 2010). It may be that neither unreplicated designs with systematic check nor complete randomization of checks are optimal for parameter estimation. In short, the statistical methods used for design and analysis must be as accurate, efficient and informative as possible (SMITH *et al.*, 2005). To improve the precision of parameter estimation, various arrangements of the checks have been suggested (LIN and POUSHINSKY, 1983, 1985; MARTIN *et al.*, 2006; MÜLLER *et al.*, 2010; CLARKE and STEFANOVA, 2011). Of these, MARTIN *et al.* (2006) is unique in considering the use of optimality criteria to search for an optimal unreplicated design for a single site.

A more recent approach, pioneered by CULLIS *et al.* (2006), is to partially replicate the lines. They introduced p -rep designs in which $p\%$ of the lines have two replicates and the remaining lines are unreplicated; no checks are included. A search is made for an optimal design for assigning the lines to plots. CLARKE and STEFANOVA (2011) have compared unreplicated designs with p -rep designs and have concluded that the latter are superior.

4.1.1 Optimality criteria for design searches

The optimal design depends on the criterion that it is sought to optimize. For linear models, the information concerning the precision of the parameter estimates is contained in the variance-covariance matrix of parameter estimates. In general, optimal designs are those experimental layouts that optimize some function of this matrix (HOOKS *et al.*, 2009).

The most commonly used criterion for comparative experiments is the A -optimality criterion. It is the average variance of pairwise differences of effects (AVPD) and the optimal design is the one that minimizes it. It is certainly appropriate when the researcher wishes to estimate differences between specific lines and, as SMITH *et al.* (2005) outline, in this case the lines effects should be assumed to be fixed. FEDERER and RAGHAVARAO (1975) introduced three additional A -criteria that might be more appropriate in the context of breeding experiments: (i) the AVPD between test line effects, A_{tt} ; (ii) the AVPD between checks, A_{cc} (not useful in breeding experiments); and (iii) the AVPD between test lines and checks, A_{tc} . Which of these is relevant depends on the comparisons that are of most importance to the researcher.

On the other hand, SMITH *et al.* (2006) argue that the lines effects should be assumed as random if the objective of the experiment is to select the best lines so that the accuracy of the line rankings is paramount. PIEPHO *et al.* (2008) agree, but note that this is not practical when the genetic variance components cannot be accurately estimated. It is generally accepted that an optimal design for random genetic effects is one that maximizes the genetic gain due to selection from field trials (for example, KEMPTON, 1984; SARKER and SINGH, 2015; CULLIS *et al.*, 2006; CLARKE and STEFANOVA, 2011). CULLIS *et al.* (2006) and BUENO FILHO and GILMOUR (2007) demonstrate that maximizing the expected genetic gain (EGG) is equivalent to minimizing the average pairwise prediction error variance (AVPD, albeit applied to random effects). On the other hand, SMITH *et al.* (2015) conjecture, in the context of multiphase experiments that display autocorrelation, that minimizing the AVPD does not lead to a design that gives maximum genetic gain.

Another approach to assessing the accuracy of the parameter estimates is based on considering the correlation between true and observed genetic effects. FOUILLOUX and LALOË (2001) propose that the accuracy of the estimates be assessed by the prediction error variance (PEV), or some function of it. BUTLER (2013) notes that PEV is proportional to the squared correlation coefficient trace of the PEV matrix for test line effects. This leads to suggesting the sum of the PEV of each genetic effect, or, equivalently, the average of these variances (APEV), as potential criteria for assessing competing designs.

HOOKS *et al.* (2009) points out that, while the AVPD is equivalent to the APEV when the effects for each random term in a model and those for different random terms can be assumed independent; they are not when one or more random terms involve autocorrelation.

PIEPHO and WILLIAMS (2006) declared that finding a universally optimal design that capitalizes on genetic correlation structure is difficult, because of the genetic variance components are typically not accurately known, and they differ among traits. However, BLUE and BLUP are valid methods of analysis, when lines can be regarded as random.

4.1.2 Comparing different design types using simulation

A topic, related to criteria for searching for designs, is that of criteria for comparing designs of different types that have been selected as possible contenders for use in an experiment. In breeding experiments, the aim is to evaluate the genetic gain resulting from selection and so this guides the choice of criteria. Here we suppose that the selected design is optimal, according to either the AVPD or the PEV, for a particular mixed model with nominated values for its variance parameters. Then, simulation is to be used to obtain values for one or more criteria for comparing designs.

Obvious criteria for comparing designs are one or both of the criteria used in searching for an optimal design: the AVPD or the PEV. However, as CLARKE and STEFANOVA (2011) noted, a number of authors have suggested that the AVPD may not be appropriate for breeding experiments. They compared designs using the SE ratio, or the relative error, which is the standard error of the comparison between test lines and control for a design as a proportion of that standard error in a completely randomized design. CULLIS *et al.* (2006) used the relative genetic gain which is the realized genetic gain (RGG), calculated as the mean of the top $s\%$ of test line EBLUPs, to the true genetic gain, calculated as the mean of the top $s\%$ of the true test line effects. They show that the RGG is well correlated with the expected genetic gain (EGG) used by KEMPTON (1984). An alternative measure of realized genetic gain, described by PIEPHO and MÖHRING (2007) and MOEHRING *et al.* (2014), is the response to selection, calculated as the mean of the true genetic effects for the test lines with the top $s\%$ of EBLUPs. PIEPHO and MÖHRING (2007) also propose another measure of genetic gain, namely the square of the correlation between the true and predicted test line effects; this is the coefficient of determination discussed by FOUILLOUX and LALOË (2001) and others. SMITH *et al.* (2011) and MOEHRING *et al.* (2014) use the (unsquared) correlation. A related measure used by MOEHRING *et al.* (2014) is the mean square error of a difference between simulated and estimated test lines effects (MSED), i.e. the root mean square error (RMSE). Finally, the similarity percentage might also be used, it being the percentage of test lines that are in the top $s\%$ of both the true genetic effects and the EBLUPs (PETERNELLI *et al.*, 2009).

Note that many of these measures require the true genetic effects and so, in conducting a simulation for a design, the same set of genetic effects must be used for every simulated data set.

The main goal of the study reported below is to evaluate the performance of an unreplicated design with systematic checks, as used by a Brazilian sugarcane company, in comparison to spatially optimized unreplicated and p -rep designs with checks and spatially optimized p -rep designs. The optimal designs are obtained using a mixed model, with the values of its variance parameters set to values like those found for the experiments analysed in Chapter 2. A simulation study was conducted to compare the genetic gain from selection for each of the designs.

4.2 Motivating example

The motivation for this work came from a partnership between CTC and University of São Paulo (Brazil). CTC provided the data from a series of sugarcane breeding experiments to form the basis of a scientific study about the design and analysis of such experiments.

There were 21 trials available, where all had 79% or more of the area planted with new sugarcane lines, denominated test lines. These lines had only one plot in each site due to insufficient material for replication and the large area that is required. The best seven percent of the test lines were chosen by comparing them with four commercial varieties, which were called checks and were replicated such that they occupied at most 21% of the plots. One special check, nominated as the interspersed check (Check “T”), was planted in almost 11% of the plots and it was allocated systematically on diagonals throughout the trial. The other three checks (Check “A”, Check “B”, Check “C”) were equally replicated and a single replicate, made up a plot of each of these three checks, was placed in three neighbouring row plots; the replicates are spread throughout to minimize, as much as possible, the distance of the test lines from a check.

CTC believed that this distribution of check lines would better capture the environmental effects in the statistical models, and that it facilitated visual selection using them, because the same column always had eight or fewer test lines between checks. However, no objective study to confirm this has been done.

Typically, a CTC experiment would occupy a rectangular array of 20 rows (r) by 25 columns

(c). In general, individual plots in these experiments are rectangular, being narrow in the row direction. There would be, say, 56 replicates of the interspersed check and 10 replicates of each of the other checks. This would leave space for 414 test lines. The Check “T” was allocated to diagonal plots in every eighth row and eighth column; a replicate of the other three checks was placed in every second or third column. The test lines were randomized to plots not occupied by the checks. Figure 4.1 gives this design, which is referred to as Design 0.

Figure 4.1. Layout of the Design 0, where there are four commercial varieties, named “A”, “B”, “C” and “T” checks, and 414 test lines in the unnamed plots.

From the results presented in Chapter 2, it can be expected that, in such experiments, there will be residual competition in the row direction and occasionally either autocorrelation or residual competition in the column direction. Global variation can also be expected in the row direction and less frequently in the column direction.

4.3 Material and methodology

In addition to Design 0, described in Section 4.2, five spatially optimized designs were identified for investigation:

1. a spatially optimized, unreplicated design with four check lines;
2. a spatially optimized, unreplicated design with a single check line;
3. a spatially optimized, unreplicated design with three check lines;
4. a spatially optimized, p -rep design with with three check lines;
5. a spatially optimized, p -rep design.

All of the designs assumed a rectangular array of 20 rows (r) by 25 columns (c) to which 414 test lines (m) were assigned. The differences between the design were in the numbers of the checks and the duplicated test lines, as presented in Table 4.1. Given these numbers the error degrees of freedom will be approximately equal for all designs.

Table 4.1. Numbers of check lines and duplicated test lines for five spatially optimized designs. Design 1 has the same numbers of check lines as Design 0. All designs were on grid of 500 plots and had 414 test lines.

	Check “A”	Check “B”	Number of plots with		Duplicated test lines
			Check “C”	Check “T”	
Design 1	10	10	10	56	—
Design 2	—	—	—	86	—
Design 3	29	29	28	—	—
Design 4	10	10	10	—	56
Design 5	—	—	—	—	86

The layouts for the spatially optimized were obtained using the `od` package (BUTLER, 2014) and the PEV optimality criterion. The number of random exchanges in searching for a layout was set at 10000. A completely randomized design was supplied as an initial design. In order to obtain a layout, a mixed model with nominated values for its variance parameters must be supplied and this model is called the design model.

The design model for the data vector, $\mathbf{Y}$, is:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\tau} + \mathbf{Z}_g\mathbf{u}_g + \mathbf{Z}_o\mathbf{u}_o + \boldsymbol{\varepsilon}; \quad (4.1)$$

given that: $\mathbf{X}$ is the incidence matrix associated with vector $\boldsymbol{\tau}$ of fixed effects; $\mathbf{Z}_g$ is the incidence matrix associated with the vector $\mathbf{u}_g$ of random genetic effects of the test lines; $\mathbf{Z}_o$ is the incidence matrix associated with the vector $\mathbf{u}_o$ of column and row random effects; $\boldsymbol{\varepsilon}$ is the vector of residual effects.

We assume that the joint distribution of $(\mathbf{u}_g, \mathbf{u}_o, \boldsymbol{\varepsilon})$ is Gaussian with zero mean and variance matrix

$$\sigma^2 \begin{bmatrix} \mathbf{G}_g(\gamma_g) & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_o(\gamma_o) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{R}(\phi) \end{bmatrix} = \begin{bmatrix} \mathbf{G}_g & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_o & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{R} \end{bmatrix}.$$

where, assuming $\mathbf{Y}$ are in field order with rows within columns, $\mathbf{G}_g(\gamma_g) = \gamma_g \mathbf{I}_m$, $\mathbf{G}_o(\gamma_o) = \gamma_o (\mathbf{I}_c \otimes \mathbf{J}_r) + \gamma_r (\mathbf{J}_c \otimes \mathbf{I}_r)$, $\mathbf{R} = \sum_c \otimes \sum_r = \text{AR}(1) \otimes \text{AR}(1)$, $\gamma_c = \sigma_c^2 / \sigma^2$, $\mathbf{I}_*$ is an identity matrix of order $*$, $\mathbf{J}_*$ is an unitary matrix of order $*$, and $\text{AR}(1)$ indicates a first-order autoregressive correlation matrix. It is noted that the results presented in Chapter 2 indicate that a banded correlation structure is more likely in the row direction than a first-order autoregressive correlation structure. However, the latter is used because the specification of banded correlation structures is not implemented in `od`. For small values of ϕ_r , the differences between the two correlation structures are small.

The values used for the variance parameters for these models are roughly consistent with the values obtained in the analyses conducted in Chapter 2. For all models σ^2 was set to 1, γ_g to 1.0, γ_c and γ_r to 0.3, and ϕ_c to 0.5.

For the design generation, two versions of each of Designs 1–5 were obtained, one in which ϕ_r was set to 0.40 (positive residual row correlation) and the other in which ϕ_r was set to -0.25 (negative residual row correlation, representing competition at the residual level). Thus 10 layouts were obtained in all and those for the spatially optimized designs are presented in Figure 4.2.

Figure 4.2. Layouts from the optimal designs 1 (panels (a) and (b)), 2 (panels (c) and (d)), 3 (panels (e) and (f)). Panels (a), (c) and (e) used $\phi_r = 0.4$ and the other panels used $\phi_r = -0.25$. “A”, “B”, “C” and “T” are the checks and the numbers are the duplicated test lines as well the unnamed plots.

Figure 4.3. Layouts from the optimal designs 4 (panels (a) and (b)) and 5 (panels (c) and (d)). Panels (a) and (c) used $\phi_r = 0.4$ and the other panels used $\phi_r = -0.25$. “A”, “B”, “C” and “T” are the checks and the numbers are the duplicated test lines as well the unnamed plots.

A simulation study was conducted to compare the designs in terms of the genetic gain from selection achievable with them. It used a mixed model, called the simulation model. There were four scenarios for each of Designs 1–5 for which simulated data sets were obtained:

- ++: design and simulation models with positive row correlation;
- +−: design model with positive row correlation, but simulation model with negative row correlation;
- −+: design model with negative row correlation, but simulation model with positive row correlation;
- −−: design model and simulation model with negative row correlation.

There were two scenarios for Design 0 that differed in their row correlation values.

The simulation model is the same as the design model in Equation 4.1, except that $\mathbf{R}$ is of the form $\sum_c \otimes \sum_r = \text{AR}(1) \otimes \text{Band}(1)$, where $\text{Band}(1)$ indicates a first-order banded correlation matrix. The two values used for the row correlation in the first off-diagonal banded were $\phi_r = 0.3$ or $\phi_r = -0.3$. The values for the other parameters are the same as those assumed for the design model.

To construct simulated data sets, the simulation model is separated into the genetic effects, $\mathbf{u}_{\mathbf{g}}$, and the non-genetic simulation model:

$$\mathbf{Y}_{\sim}^* = \mathbf{X}_{\sim} \boldsymbol{\tau} + \mathbf{Z}_{\sim} \mathbf{u}_{\sim} + \boldsymbol{\varepsilon}_{\sim};$$

Only one set of genetic effects is generated for use in all simulations. These become the true predicted genetic effects, $\mathbf{u}_{\mathbf{g}}$. Then two simulations are performed, in each of which 1000 non-genetic data sets are generated; one using negative and the other positive row correlation. For each design and scenario combination, the i th simulated data set, $\mathbf{y}_{\mathbf{i}}$, is obtained by identifying (i) the $\mathbf{Z}_{\mathbf{g}}$ matrix for its layout and (ii) the i th non-genetic data set, say $\mathbf{y}_{\mathbf{i}}^*$, for its simulation model; then $\mathbf{y}_{\mathbf{i}} = \mathbf{Z}_{\mathbf{g}}\mathbf{u}_{\mathbf{g}} + \mathbf{y}_{\mathbf{i}}^*$.

For every simulated data set, the simulation model is fitted and, if the estimation algorithm converged, the following criteria are calculated from the results and stored: PEV; the relative genetic gain; the correlation between the true and estimated genetic effects; and the similarity percentage for the top 7% of test lines. For this example, the top 7% of test lines is equivalent to the top 30 test lines, the number that CTC selects in its breeding experiments.

PEV is calculated as:

$$\text{trace}(\mathbf{Z}_{\mathbf{g}}^t(\mathbf{G}_{\mathbf{o}} + \mathbf{R})^{-1}\mathbf{Z}_{\mathbf{g}} + \mathbf{G}_{\mathbf{g}} - \mathbf{Z}_{\mathbf{g}}^t(\mathbf{G}_{\mathbf{o}} + \mathbf{R})^{-1}\mathbf{X}(\mathbf{X}^t(\mathbf{G}_{\mathbf{o}} + \mathbf{R})^{-1}\mathbf{X})^{-1}\mathbf{X}^t(\mathbf{G}_{\mathbf{o}} + \mathbf{R})^{-1}\mathbf{Z}_{\mathbf{g}})^{-1}.$$

The other measures are obtained as described in Section 4.1.2.

For each design and scenario combination there are 1000 simulated data sets and the results of the converged algorithm are summarized as follows: the density of each criterion, except the similarity percentage is produced; for the similarity percentage, boxplots are produced. In addition, since the aim is to obtain a spatially optimized design that improves on Design 0, the relative efficiency of the i th simulated data set for a target design is calculated as the ratio of two PEVs. The numerator is the PEV for the i th data set of Design 0 for the same simulation model as the target design; the denominator is the PEV for the target design. A mean of the relative efficiencies for each design-scenario combination is calculated. Relative efficiencies greater than one indicate that the spatially optimized design is more efficient than Design 0.

4.4 Results

Table 4.2 gives the mean relative efficiencies of the systematically optimized designs when compared to Design 0. Designs 1 and 0 are equally efficient, the only difference in the designs being that Design 1 is spatially optimized. Designs 2 and 3 are slightly less efficient than Design 0. They have fewer check lines (one and three, respectively), but the same number of plots with checks. In comparison designs 4 and 5, that involve p -rep designs, are more efficient than Design 0 by around 2% and 5%, respectively.

Note that the PEVs from designs assuming $\phi_r = 0.4$ are smaller values than the same design with $\phi_r = -0.25$.

Table 4.2. Relative efficiencies of spatially optimized designs to unreplicated designs with systematic checks. In parenthesis are the PEV values for Design 0.

Design	Row residual correlation			
	$\phi_r = 0.4$		$\phi_r = -0.25$	
0	1.0000	(0.4086)	1.0000	(0.4422)
1	1.0079		1.0032	
2	0.9813		0.9855	
3	0.9791		0.9835	
4	1.0197		1.0241	
5	1.0477		1.0516	

Table 4.3 presents the number of estimation algorithm converged for each design-scenario combination. While the convergence numbers differ, the differences are not large.

Table 4.3. Number of estimation algorithm converged for each design-scenario combination, given 1000 data sets for each.

Design	Scenario			
	++	+-	-+	--
0	993	977	993	977
1	987	973	987	970
2	990	970	987	981
3	990	977	992	976
4	985	999	982	974
5	990	986	992	982

Figure 4.4 presents the densities of the correlation between the true and estimated genetic effects, the relative genetic gain and the PEV values for each scenario and design. Figure 4.5 gives boxplots of the similarity percentage. Overall, there is little difference between the densities and boxplots of the four scenarios. Further, there is little difference between the Designs in relative genetic gain. However, Designs 4 and 5 are superior in terms of the correlation between true and predicted genetic effects (larger values), PEV (smaller values) and similarity percentage (higher medians). For Designs 4 and 5, in half of the data-set analyzed, the similarity percentage is between 50 to 80%.

Figure 4.4. Densities of the correlation between the true and estimated genetic effects (first column of plots), relative genetic gain (second column) and prediction error variance - PEV (third) for each scenario and design.

Figure 4.5. Similarity percentage between true and estimated top 7% of test lines. The horizontal gray line indicates 50%. The panels represent scenarios the different scenarios.

4.5 Conclusions

For the experimental situation considered here, an experimental design that included partial replication of the test lines was superior to the others. It resulted in better correlation between the true and predicted genetic effects, better relative genetic gain, smaller PEV values and higher median values of similarity percentages. This is in accordance with CLARKE and STEFANOVA (2011).

However, if a researcher wants to be able to compare test lines with commercial checks to ensure that the test lines are better than commercial varieties then a design with both checks and partial replication of test lines is to be preferred to design with unreplicated test lines. Greater genetic gain will be achieved. This agrees with the conclusion by MOEHRING *et al.* (2014) that an augmented p -rep design is better than an augmented design for individual environments.

Of interest is that, independent of the numbers of check lines, there is no advantage in using one of the spatially optimized, unreplicated designs over the unreplicated design with systematic checks. It may be that the latter is preferred because it is practically or economically more efficient. This too is in accordance with the conclusion by CLARKE and STEFANOVA (2011) that the distribution of the checks is not important. Further, it would appear that the number of check lines has only a slight effect on performance of the unreplicated designs which agrees with the conclusions of MÜLLER *et al.* (2010) for augmented designs. The effect of the number of check varieties in the case of the p -rep design with checks (Design 4) could be investigated. However, KEMPTON (1984) and MÜLLER *et al.* (2010) suggest that a minimum of two genotypes, of similar genetic background to the test material, be used to guard against differential response of lines to soil fertility variations.

It is noted that no attempt was made in this study to find improved designs for the situation in which establishing that test lines are superior to check lines is considered to be more important than selecting the best test lines. In this case, A_{tc} -optimal designs would be important.

In this simulated study, the models to generated the optimal designs and the data sets were different. However, the results from these combinations were not different given residual row correlation positive or negative, which means that in this case does not matter if the model used to generate the design is the same to the data.

5 GENERAL CONCLUSIONS AND FUTURE WORKS

In this work, we have fitted models for 21 sugarcane trials, where some of them had similar lines but they were planted in different sites or harvest year at the same region. With this, we found that each trial had its own peculiarity as far as the residual variance structure was concerned and so it was impossible to determine one specific model for the residual variation that applied to all the experiments. The strategy to select a mixed model was to include all the terms that the experience tells us are likely to occur in the experiment. However, it was found that some data set did not have significant direct genetic effects (component close to zero and/or standard error larger than the estimate) and the banded correlation structure models were more realistically for row direction than the autoregressive models. For the competition effect, it would appear that there was more general competition in Brazilian experiments rather than it being genotype specific, because there was only one experiment with significant neighbour genetic effect while ten others had significant residual competition. One response not outcome was the high genetic correlation value obtained from the joint analyses. This was twice or more than the correlation between genetic EBLUPs from individual analyses. Hence, using the joint analyses, better accuracy in the selection of the test lines was obtained. Higher predicted values were found with lower standard errors, so that the observation of genotypes at multiple sites provides replication of them. Thus a joint analysis was preferable to individual analyses whenever it is possible, whether it is based on different sites or more than one harvest from the same trial.

The simulation study for single sites or cuts demonstrates clearly that a strategy of fitting a model, even if it is the true model, was not a viable strategy and that detecting genetic competition in an experiment using the CTC type of design was fraught. Hence, it cannot be stated conclusively that genetic competition does not occur in Brazilian sugarcane experiments. Fortunately, it would appear that the prediction of the direct genetic effects was little affected by the model selected and the presence or not of residual competition, although it was affected by the magnitude of the genetic variance. While, in general, the simulated model (Model 4) gave the best correlation between true and predicted genetic effects, selection gain and average similarity, the differences between it and the models without neighbour genetic effects and/or residual correlation are small.

The results of the simulation study for separate versus joint analyses shows there were no problems in fitting either single-site or joint models when there was no genetic competition and that better predicted values of the genetic effects come from joint analyses. There was little effect of heterogeneous variances, but there were differences in the results that depended on the magnitude of the genetic variance. For the models simulated, there was little difference between fitting autocorrelation and banded correlation structures.

In the chapter 3, it was showed that the negative correlation can be due to the lodging effect if this results in material from plot being associated with a plot from a neighbouring column at the harvest. As the dry weather can also cause more competition between plots, the negative correlation can be related to either one or both of these effects. Hence, when there is lodging care in separating material from different plots will lead to improved results.

In the end, when designs were studied, it was presented, under specific conditions, that the partial replication of the test lines was superior to the others. It resulted in better correlation between the true and predicted genetic effects, better relative genetic gain, smaller PEV values and higher median values of similarity percentages. However, the design does not allow to compare test lines with commercial checks to ensure that the test lines are better than commercial varieties. Then a design with both checks and partial replication of test lines is to be preferred to design with unreplicated test lines. Other important result was that, independent of the numbers of check lines, there is no advantage in using one of the spatially optimized unreplicated designs over the unreplicated design with systematic checks. It

may be that the latter is preferred because it is practically or economically more efficient. Further, it would appear that the number of check lines has only a slight effect on performance of the unreplicated designs.

Future works could be the development of new experiments using different designs and space in the plots. It is conjectured that double furrow used in the Brazilian sugarcane experiments plots decreases the competition effect, as this indicates an extra space in each plot. Then, we suggest new experiments with the same lines but different design and single furrow to see if there is significant neighbour genetic effects and the direct genetic variance increase to achieve better selection results. Given that the true best simulated test lines are selected with average around 45 to 55%, even for the individual and joint analyses or optimal design. Another study could be with the test power to identify the significance of the variance and covariance components.

In the software ASReml-R the structure $\text{Band()} \times \text{Band()}$ is not available, and the response variance matrix cannot be written with spline effects to obtain the plot of the empirical variogram using `asremlPlus`. Then new functions could be developed.

REFERENCES

- ALI, I., H. M. AHMED, and S. A. SHAH, 2013 Evaluation and selection of rapeseed (*brassica napus* l.) mutant lines for yield performance using augmented design. *Journal of Animal and Plant Sciences* **23**: 1125–1130.
- ATKIN, F., J. STRINGER, and M. DIETERS, 2009 Interplot competition between sugarcane families significantly affects estimates of cane yield and family selection. *Proceedings of the 2009 Conference of the Australian Society of Sugar Cane Technologists held at Ballina* **31**: 326–331.
- BESAG, J. and R. KEMPTON, 1986 Statistical analysis of field experiments using neighbouring plots. *Biometrics* pp. 231–251.
- BRIEN, C. J., 2016 *asremlPlus: Augments the Use of 'Asreml' in Fitting Mixed Models*.
- BUENO FILHO, J. S. S. and S. G. GILMOUR, 2007 Block designs for random treatment effects. *Journal of Statistical Planning and Inference* **137**: 1446–1451.
- BUTLER, D., 2014 *od: generate optimal designs*. R package version 0.75.
- BUTLER, D., B. CULLIS, A. GILMOUR, and B. GOGEL, 2009 *Asreml-r reference manual*. The State of Queensland, Department of Primary Industries and Fisheries, Brisbane .
- BUTLER, D. G., 2013 *On the optimal design of experiments under the linear mixed model*. Ph.D. thesis, School of Mathematics and Physics, Brisbane, Australia.
- CLARKE, G. P. Y. and K. T. STEFANOVA, 2011 Optimal design for early-generation plant-breeding trials with unreplicated or partially replicated test lines. *Australian & New Zealand Journal of Statistics* **53**: 461–480.
- CONAB, 2016 Companhia nacional de abastecimento - acompanhamento da safra brasileira de cana-de-açúcar, primeiro levantamento. http://www.conab.gov.br/01alaCMS/uploads/arquivos/16_04_18_14_27_15_boletim_cana_portugues_-_1o_lev_-_16.pdf, accessed 24 May of 2016 .
- CTC, 2013 Centro de tecnologia canavieira - institucional. <http://www.ctcanavieira.com.br/index.html>, accessed 24 May of 2016 .
- CULLIS, B., B. GOGEL, A. VERBYLA, and R. THOMPSON, 1998 Spatial analysis of multi-environment early generation variety trials. *Biometrics* pp. 1–18.
- CULLIS, B. R., A. SMITH, C. HUNT, and A. GILMOUR, 2000 An examination of the efficiency of australian crop variety evaluation programmes. *The Journal of Agricultural Science* **135**: 213–222.
- CULLIS, B. R., A. B. SMITH, and N. E. COOMBES, 2006 On the design of early generation variety trials with correlated data. *Journal of Agricultural, Biological, and Environmental Statistics* **11**: 381–393.
- DE FAVERI, J., A. P. VERBYLA, W. S. PITCHFORD, S. VENKATANAGAPPA, and B. R. CULLIS, 2015 Statistical methods for analysis of multi-harvest data from perennial pasture variety selection trials. *Crop and Pasture Science* **66**: 947–962.
- DINARDO-MIRANDA, L., A. VASCONCELOS, and M. LANDELL, 2008 *Cana-de-açúcar*. Instituto Agromico de Campinas, Campinas.
- FAOSTAT, F., 2013 *Agricultural commodities production. Countries by commodity*. http://faostat3.fao.org/browse/rankings/countries_by_commodity/E, accessed 18 July of 2016 .

- FEDERER, W. T., 1956 Augmented (or hoonuiaku) designs. *Hawaiian Planters Record* **55**: 191–208.
- FEDERER, W. T. and D. RAGHAVARAO, 1975 On augmented designs. *Biometrics* **31**: 29–35.
- FOUILLOUX, M.-N. and D. LALOË, 2001 A sampling method for estimating the accuracy of predicted breeding values in genetic evaluation. *Genetics Selection Evolution* **33**: 473–486.
- GILMOUR, A. R., B. R. CULLIS, and A. P. VERBYLA, 1997 Accounting for natural and extraneous variation in the analysis of field experiments. *Journal of Agricultural, Biological, and Environmental Statistics* pp. 269–293.
- HOOKE, T., D. MARX, S. KACHMAN, and J. PEDERSEN, 2009 Optimality criteria for models with random effects. *Revista Colombiana de Estadística* **32**: 17–31.
- HUNT, C. H., A. B. SMITH, D. R. JORDAN, and B. R. CULLIS, 2013 Predicting additive and non-additive genetic effects from trials where traits are affected by interplot competition. *Journal of Agricultural, Biological, and Environmental Statistics* **18**: 53–63.
- KEMPTON, R. A., 1984 The design and analysis of unreplicated field trials. *Vorträge für Pflanzenzüchtung* **7**: 219–242.
- LIN, C.-S. and G. POUSHINSKY, 1983 A modified augmented design for an early stage of plant selection involving a large number of test lines without replication. *Biometrics* pp. 553–561.
- LIN, C.-S. and G. POUSHINSKY, 1985 A modified augmented design (type 2) for rectangular plots. *Canadian journal of plant science* **65**: 743–749.
- LIU, S., G. A. CONSTABLE, B. R. CULLIS, W. N. STILLER, and P. E. REID, 2015 Benefit of spatial analysis for furrow irrigated cotton breeding trials. *Euphytica* **201**: 253–264.
- MARTIN, R., N. CHAUHAN, J. ECCLESTON, and B. CHAN, 2006 Efficient experimental designs when most treatments are unreplicated. *Linear algebra and its applications* **417**: 163–182.
- MOEHRING, J., E. R. WILLIAMS, and H.-P. PIEPHO, 2014 Efficiency of augmented p-rep designs in multi-environmental trials. *Theoretical and applied genetics* **127**: 1049–1060.
- MOORE, P. H., A. H. PATERSON, and T. TEW, 2013 Sugarcane: the crop, the plant, and domestication. *Sugarcane: Physiology, Biochemistry, and Functional Biology* pp. 1–17.
- MÜLLER, B., A. SCHÜTZENMEISTER, and H.-P. PIEPHO, 2010 Arrangement of check plots in augmented block designs when spatial analysis is used. *Plant breeding* **129**: 581–589.
- PASTINA, M., M. MALOSETTI, R. GAZAFFI, M. MOLLINARI, G. MARGARIDO, K. OLIVEIRA, L. PINTO, A. SOUZA, F. VAN EEUWIJK, and A. GARCIA, 2012 A mixed model qtl analysis for sugarcane multiple-harvest-location trial data. *Theoretical and Applied Genetics* **124**: 835–849.
- PETERNELLI, L. A., E. F. M. D. SOUZA, M. H. P. BARBOSA, and M. P. D. CARVALHO, 2009 Delineamentos aumentados no melhoramento de plantas em condições de restrições de recursos. *Ciência Rural* **39**: 2425–2430.
- PIEPHO, H., J. MÖHRING, A. MELCHINGER, and A. BÜCHSE, 2008 Blup for phenotypic selection in plant breeding and variety testing. *Euphytica* **161**: 209–228.
- PIEPHO, H. and E. WILLIAMS, 2006 A comparison of experimental designs for selection in breeding trials with nested treatment structure. *Theoretical and Applied Genetics* **113**: 1505–1513.

- PIEPHO, H. P. and J. MÖHRING, 2007 Computing heritability and selection response from unbalanced plant breeding trials. *Genetics* **177**: 1881–1888.
- R CORE TEAM, 2014 R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
- SANTOS, A. H., E. BEARZOTI, D. F. FERREIRA, and J. A. L. D. SILVA FILHO, 2002 Simulation of mixed models in augmented block design. *Scientia Agricola* **59**: 483–489.
- SARKER, A. and M. SINGH, 2015 Improving breeding efficiency through application of appropriate experimental designs and analysis models: a case of lentil (*lens culinaris medikus subsp. culinaris*) yield trials. *Field Crops Research* **179**: 26–34.
- SILVA, J. A. C. and R. J. KERR, 2013 Accounting for competition in genetic analysis, with particular emphasis on forest genetic trials. *Tree genetics & genomes* **9**: 1–17.
- SINGH, G., S. CHAPMAN, P. JACKSON, and R. LAWN, 2002 Lodging reduces sucrose accumulation of sugarcane in the wet and dry tropics. *Crop and Pasture Science* **53**: 1183–1195.
- SMITH, A., B. R. CULLIS, and R. THOMPSON, 2005 The analysis of crop cultivar breeding and evaluation trials: an overview of current mixed model approaches. *The Journal of Agricultural Science* **143**: 449–462.
- SMITH, A., P. LIM, and B. R. CULLIS, 2006 The design and analysis of multi-phase plant breeding experiments. *The Journal of Agricultural Science* **144**: 393–409.
- SMITH, A. B., D. G. BUTLER, C. R. CAVANAGH, and B. R. CULLIS, 2015 Multi-phase variety trials using both composite and individual replicate samples: a model-based design approach. *Journal of Agricultural Science* **153**: 1017–1029.
- SMITH, A. B., R. THOMPSON, D. G. BUTLER, and B. R. CULLIS, 2011 The design and analysis of variety trials using mixtures of composite and individual plot samples. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **60**: 437–455.
- STEFANOVA, K. T., A. B. SMITH, and B. R. CULLIS, 2009 Enhanced diagnostics for the spatial analysis of field trials. *Journal of agricultural, biological, and environmental statistics* **14**: 392–410.
- STRINGER, J. and B. CULLIS, 2002 Application of spatial analysis techniques to adjust for fertility trends and identify interplot competition in early stage sugarcane selection trials. *Crop and Pasture Science* **53**: 911–918.
- STRINGER, J. K., 2006 Joint modelling of spatial variability and interplot competition to improve the efficiency of plant improvement.. Ph.D. thesis, University of Queensland.
- STRINGER, J. K., B. R. CULLIS, and R. THOMPSON, 2011 Joint modeling of spatial variability and within-row interplot competition to increase the efficiency of plant improvement. *Journal of agricultural, biological, and environmental statistics* **16**: 269–281.
- VAN HEERDEN, P. *et al.*, 2011 Effects of lodging and anti-lodging chemicals on the productivity of variety n25: Preliminary findings. *Proceedings of the Annual Congress-South African Sugar Technologists' Association* pp. 169–172.
- VAN HEERDEN, P., A. SINGELS, A. PARASKEVOPOULOS, and R. ROSSLER, 2015 Negative effects of lodging on irrigated sugarcane productivity - an experimental and crop modelling assessment. *Field Crops Research* **180**: 135–142.

APPENDIX

A CORRELATION MODELS

Assuming the general matrix of dimension $n \times n$ as:

$$\begin{bmatrix} \rho_0 & \rho_1 & \rho_2 & & \rho_{n-1} \\ \rho_1 & \rho_0 & \rho_1 & & \rho_{n-2} \\ \rho_2 & \rho_1 & \rho_0 & & \rho_{n-3} \\ & & & \ddots & \\ \rho_{n-1} & \rho_{n-2} & \rho_{n-3} & & \rho_0 \end{bmatrix}$$

Given $i = 0, 1, 2, \dots, n-1$. The structures are:

Identity (id):

$$\begin{bmatrix} 1 & 0 & 0 & & 0 \\ 0 & 1 & 0 & & 0 \\ 0 & 0 & 1 & & 0 \\ & & & \ddots & \\ 0 & 0 & 0 & & 1 \end{bmatrix};$$

$$\rho_0 = 1 \text{ and } \rho_{i,i>0} = 0$$

First order autoregressive (AR(1)):

$$\begin{bmatrix} 1 & \phi_1 & \phi_1^2 & & \phi_1^{n-1} \\ \phi_1 & 1 & \phi_1 & & \phi_1^{n-2} \\ \phi_1^2 & \phi_1 & 1 & & \phi_1^{n-3} \\ & & & \ddots & \\ \phi_1^{n-1} & \phi_1^{n-2} & \phi_1^{n-3} & & 1 \end{bmatrix};$$

$$\rho_i = \phi_1^i; |\phi_1| < 1$$

Second order autoregressive (AR(2)):

$$\begin{bmatrix} 1 & \frac{\phi_1}{(1-\phi_2)} & \frac{\phi_1^2}{(1-\phi_2)} + \phi_2 & & \rho_{n-1} \\ \frac{\phi_1}{(1-\phi_2)} & 1 & \frac{\phi_1}{(1-\phi_2)} & & \rho_{n-2} \\ \frac{\phi_1^2}{(1-\phi_2)} + \phi_2 & \frac{\phi_1}{(1-\phi_2)} & 1 & & \rho_{n-3} \\ & & & \ddots & \\ \rho_{n-1} & \rho_{n-2} & \rho_{n-3} & & 1 \end{bmatrix};$$

$$\rho_0 = 1, \rho_1 = \frac{\phi_1}{(1-\phi_2)} \text{ and } \rho_{i,i>1} = \phi_1 \rho_{i-1} + \phi_2 \rho_{i-2}, |\phi_1 \pm \phi_2| < 1, |\phi_1| < 1, |\phi_2| < 1$$

Third order autoregressive (AR(3)):

$$\begin{cases} \rho_0 = 1 \\ \rho_1 = \frac{\phi_1 + \phi_2 \phi_3}{\omega} \\ \rho_2 = \frac{\phi_1(\phi_1 + \phi_3) + \phi_2(1 - \phi_2)}{\omega} \\ \rho_{i,i>2} = \phi_1 \rho_{i-1} + \phi_2 \rho_{i-2} + \phi_3 \rho_{i-3} \\ \omega = 1 - \phi_2 - \phi_3(\phi_1 + \phi_3) \end{cases}$$

$$|\phi_1| < (1 - \phi_2), |\phi_2| < 1, |\phi_3| < 1$$

Constrained autoregressive 3 (sar(2)):

$$\text{as for AR(3) where: } \begin{cases} \phi_1 = \gamma_1 + 2\gamma_2 \\ \phi_2 = -\gamma_2(2\gamma_1 + \gamma_2) \\ \phi_3 = \gamma_1\gamma_2^2 \end{cases}$$

first-order banded (Band(1)):

$$\begin{bmatrix} 1 & \rho_1 & 0 & & 0 \\ \rho_1 & 1 & \rho_1 & & 0 \\ 0 & \rho_1 & 1 & & 0 \\ & & & \ddots & \\ 0 & 0 & 0 & & 1 \end{bmatrix};$$

$$\rho_0 = 1, \rho_{i,i>1} = 0.$$

Second-order banded (Band(2)):

$$\begin{bmatrix} 1 & \rho_1 & \rho_2 & & 0 \\ \rho_1 & 1 & \rho_1 & & 0 \\ \rho_2 & \rho_1 & 1 & & 0 \\ & & & \ddots & \\ 0 & 0 & 0 & & 1 \end{bmatrix};$$

$$\rho_0 = 1, \rho_{i,i>2} = 0.$$

Unstructured general covariance (us(n)):

$$\begin{bmatrix} \phi_{11} & \phi_{12} & \phi_{13} & & \phi_{1n} \\ \phi_{12} & \phi_{22} & \phi_{23} & & \phi_{2n} \\ \phi_{13} & \phi_{23} & \phi_{33} & & \phi_{3n} \\ & & & \ddots & \\ \phi_{1n} & \phi_{2n} & \phi_{3n} & & \phi_{nn} \end{bmatrix};$$

Factor analytic order k (fa(k)):

$$\begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_k \end{bmatrix} \begin{bmatrix} \gamma_1 & \gamma_2 & \cdots & \gamma_k \end{bmatrix} + \begin{bmatrix} \psi_1 & 0 & 0 \\ 0 & \psi_2 & 0 \\ & & \ddots \\ 0 & 0 & \psi_k \end{bmatrix} = \Gamma\Gamma' + \Psi;$$

Γ contains covariance factors and Ψ contains specific variance.

Reduced rank (RR(2)):

$$\sigma \begin{bmatrix} 1 & \rho \\ \rho & \rho^2 \end{bmatrix};$$

B PARANÁ - LONGITUDINAL DATA

This experiment assessed 406 test lines with 4 commercial varieties nominated “PAD1”, “PAD2”, “PAD3” and “Inter”. It was planted in 2013 in a rectangular array of 26 columns by 19 rows and the yield in TCH was obtained in 2014 as first cut and 2015 in a second cut. The results were organized in subsections where B.1 presents the descriptive analysis of the cuts, B.2 shows the individual analysis for each one and B.3 the joint analysis. Finally, we make concluding remarks in subsection B.4.

B.1 Descriptive analysis

A summary about TCH in the first and second cut is presented in Table B.1. The average of TCH decrease in the first to second cut and this may be due to lower environmental effect. The variances in the sugarcane check plots is smaller for the second cut when compared to the first and the largest variance (1588.89) was found in “PAD2” variety. For the plots with test lines there are lower average of yield and the variance increases when comparing the cuts. Looking at Figure B.1, the boxplots of TCH of the groups of clones in each cut reveal some possible outliers. Note that one of the “PAD2” variety replicates presented one plot with a super yield for the first cut.

Table B.1. Descriptive analysis of the groups of sugarcane carried out in Paraná in the first and second cut.

	Groups of clones	Minimum	Maximum	Mean	Variance
First cut	PAD1	59.50	131.00	96.57	547.98
	PAD2	59.50	192.00	89.29	1588.89
	PAD3	68.50	110.10	85.50	261.89
	Interspersed	53.60	138.40	92.57	330.91
	Test lines (Reg)	20.80	142.90	74.88	414.69
Second cut	PAD1	55.10	104.20	82.34	279.19
	PAD2	40.20	89.30	68.29	250.90
	PAD3	56.50	83.30	69.43	89.03
	Interspersed	20.80	105.70	80.96	237.13
	Test lines (Reg)	8.90	141.40	65.80	448.62

Figure B.1. Boxplot of TCH for each group of lines for Paraná cuts. Panel (a) refers to first cut and (b) second.

Figures B.2 and B.3 display the yield for each plot in the first and second cut, respectively. The names “PAD1”, “PAD2”, “PAD3” and “INTER” are the checks, the alpha-numerical labels (PR plus number) are the test lines, “x” are empty plots. There is at 3 last column yielded less than the

others because an interference effect extern. Then, a co-variate for the difference between these and other columns was included (namely in the analysis, Cext). Also, it is possible to observe a high yield in the “PAD2” plot (first column and second row). This same plot presented a regular yield in the second cut. When discussing this with the company, they informed that this measure for first cut was incorrect, and hence was removed from the analysis.

Figure B.2. Heat map for TCH in the first cut of the Paraná.

Figure B.3. Heat map for TCH in the second cut of the Paraná.

B.2 Individual analysis

For the first cut, it was observed one wrong observation of the standard check (with TCH was 192) which was omitted. Then, the experiment was analyzed with 486 and 487 plots in the first and second cut, respectively.

A summary of the sequence of models fitted to Paraná first cut data is presented in Table B.2. Our analysis began with the model denominated random effects Treatment Interference Model (R-TIM), used by Stringer et al. (2011). This model includes random column and row effects, and unconstrained direct and neighbour genetic effects with an AR(1) correlation structure in row and column direction. However, the genetic effects can be written as a special case, denominated reduced rank. From the REML ratio test, we conclude that the neighbour genetic effect was not significant. Then, the local and global effects were assessed, respectively. Looking at the empirical variograms of the residuals, Figure B.4, we observed that for Model 4, the semivariances of the residuals in both the column and row direction were within the expected values, although, for some models in the row direction, some semivariances were close to the envelope limits. Finally, the selected model was Model 18. This model contains random row and direct genetic effects, fixed checks and co-variate effects and strange local correlation in row direction modelled with a fourth-order banded structure. The REML estimates of variance parameters from the selected model are shown in Table B.3.

Table B.2. Summary of the fitted models to the Paraná first cut data with restricted log likelihood (Log-lik.) and the p -value of the REML ratio test. All models include a random direct genetic effects for the test lines (G), fixed check varieties effects and co-variate. The other effects are represented with the symbols: H (harvest); Co(cone); S (sowing); C (column); R (row); N (neighbour genetic); spl(.) (spline in some direction indicated in parentheses) and lin(.) (linear trend in some direction indicated in parentheses). The structures can be: US(.)- unstructured; RR(.)- reduced rank; AR(1) - first-order autoregressive; Band(.) - banded correlation, the number indicates the order of correlation; and Id - identity.

Model	Effects				Log-lik.	Test	p -value
	Fixed	Global/extraneous	Genetic	Local (column $\times$ row)			
1		C + R	US(G:N)	AR(1) $\times$ AR(1)	-1650.75		
2		C + R	RR(G:N)	AR(1) $\times$ AR(1)	-1650.75		
3		C + R	G	AR(1) $\times$ AR(1)	-1651.37	M2 vs M3	0.26
4		C + R	G	AR(1) $\times$ Band(3)	-1647.81		
5		C + R	G	AR(1) $\times$ Band(4)	-1645.74	M5 vs M4	0.04
6		C + R	G	Id $\times$ Band(4)	-1646.12	M5 vs M6	0.36
7	H	C + R	G	Id $\times$ Band(4)	-1644.57	not significant	
8	Co	C + R	G	Id $\times$ Band(4)	-1645.14	not significant	
9	S	C + R	G	Id $\times$ Band(4)	-1644.62	not significant	
10	lin(C)+lin(R)	spl(C)+C+spl(R)+R	G	Id $\times$ Band(4)	-1647.77		
11	lin(C)+lin(R)	spl(C)+spl(R)+R	G	Id $\times$ Band(4)	-1647.97	M10 vs M11	0.26
12	lin(C)+lin(R)	spl(C)+spl(R)	G	Id $\times$ Band(4)	-1651.74	M11 vs M12	0.00
13	lin(C)+lin(R)	spl(R)+R	G	Id $\times$ Band(4)	-1648.09	M11 vs M13	0.32
14	lin(C)+lin(R)	spl(R)	G	Id $\times$ Band(4)	-1651.88	M13 vs M14	0.00
15	lin(C)+lin(R)	R	G	Id $\times$ Band(4)	-1648.26	M13 vs M15	0.27
16	lin(C)	R	G	Id $\times$ Band(4)	-1647.26	not significant	
17	lin(R)	R	G	Id $\times$ Band(4)	-1647.47	not significant	
18		R	G	Id $\times$ Band(4)	-1646.46		
19		R	RR(G:N)	Id $\times$ Band(4)	-1646.22	M19 vs M18	0.49

Figure B.4. Plots of the row and column faces of the empirical variograms for the residuals referring to Models 4 (panels (a) and (d)); 6 (panels (b) and (e)) and 18 (panels (c) and (f)). Panels (a), (b) and (c) refer to the column direction and panels (d), (e) and (f) refer to the row direction.

Observe in Table B.3 that the direct genetic component is approximately half of the residual component and even through the random row component is very small when compared with the residual component (ratio of 0.09), it is significant. The residual correlation in row direction between the second, third and fourth neighbours are similar, approximately 0.18. These were not expected and look like they reflect a machinery or procedural effect, but we do not have information from the company about this. There is a competition effect at the residual level because the correlation between the first neighbours in row direction is close to zero (0.02) and the second is approximately 8 times more.

Table B.3. REML variance parameters estimates from fitted Model 18 to the experiment of Paraná, first cut.

Variance parameters		Ratios (γ)	Estimates	Std.errors
Row	$(\tilde{\sigma}_r^2)$	0.09	20.50	11.23
Direct genetic	$(\tilde{\sigma}_g^2)$	0.52	123.16	43.67
Residual	$(\tilde{\sigma}^2)$	1.00	238.87	37.28
Spatial	$(\tilde{\rho}_{r_1})$	—	0.02	0.08
Spatial	$(\tilde{\rho}_{r_2})$	—	0.17	0.08
Spatial	$(\tilde{\rho}_{r_3})$	—	0.18	0.07
Spatial	$(\tilde{\rho}_{r_4})$	—	0.19	0.07

Table B.4 shows the Wald F tests for fixed terms from the selected model (Model 18). These showed no evidence to affirm that there was a significant difference between the group of test lines and checks (Control) or among the checks (Control:Check). The estimated fixed effects from this model are in Table B.5.

Table B.4. Wald F tests for fixed terms for Model 18. Df means degrees of freedom and DenDF is the approximate denominator degrees of freedom.

Fixed terms	Df	denDF	F test	p-value
(Intercept)	1.00	473.00	2315.00	0.00
Control	1.00	473.00	2.02	0.16
Cext	1.00	473.00	36.33	0.00
Control:Check	3.00	473.00	1.56	0.20

Table B.5. Estimates of the fixed effects for Model 18 and their respective standard errors.

Effects	Estimates	Standard error
INTERC	0.00	—
PAD1	4.76	5.41
PAD2	-10.08	5.60
PAD3	-1.82	5.28
co-variate (Cext)-1 level	0.00	—
co-variate (Cext)-2 level	-20.65	3.45
overall Checks	0.00	—
overall Test lines	-16.73	11.30
overall mean (μ)	93.60	11.34

The 30 best test lines from Model 18 and their respective predicted values and standard errors are in Table B.6.

Table B.6. The 30 best test lines with predicted values (pred. value) and standard errors (stand. error) from Model 18.

Test line	Pred. value	Stand. error	Test line	Pred. value	Stand. error
PR171	91.08	9.07	PR377	78.18	9.08
PR279	84.20	9.10	PR327	78.07	9.07
PR295	83.78	9.04	PR90	78.00	9.13
PR217	83.60	9.05	PR344	77.55	9.04
PR14	82.77	9.18	PR5	77.34	9.05
PR164	80.32	9.06	PR91	77.31	9.14
PR304	80.20	9.07	PR82	77.29	9.15
PR331	79.83	9.16	PR361	77.15	9.05
PR383	79.63	9.13	PR190	77.11	9.06
PR247	79.25	9.09	PR113	77.06	9.18
PR157	79.21	8.98	PR7	77.06	9.08
PR96	79.16	9.06	PR143	77.05	9.05
PR71	78.99	9.13	PR118	76.92	9.09
PR246	78.65	9.18	PR133	76.91	9.05
PR250	78.27	9.08	PR72	76.65	9.06

The correlation of EBLUPs when comparing the selected model (Model 18) and the traditional model with only global effects (row, column and direct genetic) is high, approximately 0.967. Although, the group of test lines selected by each model is different (see Figure B.5).

Figure B.5. Dispersion plot of the genetic EBLUPs values of the Model 18 and the traditional non-spatial analysis model for this design (without spatial dependence) for 1st cut of Paraná. The cut-offs for the 30 best test lines (7 % upper) in each Model are indicated by the dotted line.

For the second cut, the same procedure in first cut was done. However, the genetic structure effect can not be written as reduced rank. The neighbour genetic effect is close to zero and using the REML test it was verified that it is not significant. Figure B.6 presents the empirical variograms. When investigating alternative models, harvest and splines effects were significant. In the end, the best model found was Model 14 with: fixed harvest, check and linear trend effects in row and column; random direct genetic and splines effects in row and column; and third-order banded structure in row direction. The empirical semi-variogram for this model shows an improvement over that for Model 3 in Figure B.6 and indicates that the model is adequate.

Table B.7. Summary of the fitted models to the Paraná second cut data with restricted log likelihood (Log-lik.) and the p -value of the REML ratio test. All models include a random direct genetic effects for the test lines (G), fixed check varieties effects and co-variate. The other effects are represented with the symbols: H (harvest); Co (cone); S (sowing); C (column); R (row); N (neighbour genetic); spl(.) (spline in some direction indicated in parentheses) and lin(.) (linear trend in some direction indicated in parentheses). The structures can be: US(.) - unstructured; RR(.) - reduced rank; sar(2) - constrained autoregressive; AR(.) - autoregressive; Band(.) - banded correlation; and Id - identity. The number indicates the order of correlation.

Model	Fixed	Global/extraneous	Effects		Local	Log-lik. (column $\times$ row)	Test	p -value
			Random	Genetic				
1		C + R		US(G:N)	AR(1) $\times$ AR(1)	-1692.29		
2		C + R		G	AR(1) $\times$ AR(1)	-1692.69	M1 vs M2	0.51
3		C + R		G	AR(1) $\times$ Band(3)	-1690.18		
4		C + R		G	AR(1) $\times$ Band(2)	-1692.57	M3 vs M4	0.03
5		C + R		G	AR(1) $\times$ Band(4)	not converged		
6		C + R		G	AR(1) $\times$ sar(2)	not converged		
7		C + R		G	AR(2) $\times$ Band(3)	-1689.33	M7 vs M3	0.19
8		C + R		G	Id $\times$ Band(3)	-1690.18	M3 vs M8	0.96
9	H	C + R		G	Id $\times$ Band(3)	-1687.68	significant	
10	Co	C + R		G	Id $\times$ Band(3)	-1689.22	not significant	
11	S	C + R		G	Id $\times$ Band(3)	-1689.15	not significant	
12	H+lin(C)+lin(R)	spl(C)+C+spl(R)+R		G	Id $\times$ Band(3)	-1686.90		
13	H+lin(C)+lin(R)	spl(C)+spl(R)+C		G	Id $\times$ Band(3)	-1686.90	M12 vs M13	0.50
14	H+lin(C)+lin(R)	spl(C)+spl(R)		G	Id $\times$ Band(3)	-1686.99	M13 vs M14	0.33
15	H+lin(C)+lin(R)	spl(C)		G	Id $\times$ Band(3)	-1690.39	M14 vs M15	0.00
16	H+lin(C)+lin(R)	spl(R)		G	Id $\times$ Band(3)	-1690.91	M14 vs M16	0.00
17	H+lin(C)+lin(R)	spl(C)+spl(R)		US(G:N)	Id $\times$ Band(3)	-1685.98	M17 vs M14	0.26

Figure B.6. Panels (a) and (b) refer respectively to the plots of the column and row faces of the empirical variograms for the residuals referring to Model 3. Panel (c) is the semi-variogram of the Model 14.

In Table B.8 can be seen the REML estimates of variance parameters of the Model 14. The direct genetic component for the second cut is almost twice those for the first cut. With Model 14, the direct genetic component is 1.31 times the residual component, this means better test line selection for the second cut than first. Note that the correlations in row direction do not represent autoregressive process. The correlation for the second band is close to zero.

Table B.8. REML estimates of variance parameters from fitted Model 14 for the experiment of Paraná, second cut.

Variance parameters		Ratios (γ)	Estimates	Std. errors
spl(Column)		0.67	125.77	102.91
spl(Row)		0.15	28.21	25.69
Direct genetic	$(\tilde{\sigma}_g^2)$	1.31	245.05	42.42
Residual	$(\tilde{\sigma}^2)$	1.00	187.27	30.51
Spatial	$(\tilde{\rho}_{r1})$	—	-0.19	0.09
Spatial	$(\tilde{\rho}_{r2})$	—	0.03	0.09
Spatial	$(\tilde{\rho}_{r3})$	—	0.19	0.09

Table B.9 shows the Wald F tests for fixed terms. Also, there was no evidence of significant difference between the group of test lines and checks (Control), but within checks this is not true (Control:Check). The estimated fixed terms from this model are in Table B.10.

Table B.9. Wald F tests for fixed terms for Model 14. Df means degrees of freedom and DenDF is the approximate denominator degrees of freedom.

Fixed terms	Df	denDF	F test	p-value
(Intercept)	1.00	471.00	4013.00	0.00
Control	1.00	471.00	0.70	0.40
Cext	1.00	471.00	8.84	0.00
Harvest	1.00	471.00	6.82	0.01
lin(Col)	1.00	471.00	2.66	0.10
lin(Row)	1.00	471.00	3.47	0.06
Control:Check	3.00	471.00	2.84	0.04

Table B.10. Estimates of the fixed effects for Model 14 and their respective standard errors.

Effects	Estimates	Standard errors
INTERC	0.00	—
PAD1	1.67	4.87
PAD2	-9.76	4.84
PAD3	-10.49	4.81
lin(Row)	0.26	0.15
lin(Col)	-0.37	0.21
Harvest - 1 level	0.00	—
Harvest - 2 level	-4.52	1.78
co-variate (Cext)-1 level	0.00	—
co-variate (Cext)-2 level	-1.20	7.19
overall Checks	0.00	—
overall Test lines	-15.32	15.79
overall mean (μ)	85.29	15.98

In Table B.11 are the 30 best test lines from Model 14 and their respective predicted values and standard errors.

Table B.11. The 30 best test lines with predicted values (pred. value) and respective standard errors (stand. error) for Model 14.

Test line	Pred. value	Stand. error	Test line	Pred. value	Stand. error
PR96	108.59	10.97	PR204	90.13	11.07
PR56	103.58	11.09	PR150	90.03	10.95
PR358	98.64	11.17	PR1	89.77	11.30
PR331	95.03	11.20	PR400	89.75	10.80
PR52	94.65	11.25	PR207	89.50	11.53
PR279	94.40	11.19	PR6	89.41	11.10
PR232	94.15	11.56	PR318	89.21	11.36
PR90	93.44	10.98	PR233	89.10	11.30
PR386	93.26	11.06	PR122	88.99	10.83
PR98	92.88	10.84	PR342	88.77	11.12
PR133	92.77	10.97	PR170	88.46	11.31
PR9	92.53	11.21	PR14	88.26	11.36
PR327	91.92	11.19	PR143	88.20	11.20
PR171	90.85	10.98	PR189	87.71	10.85
PR309	90.38	11.10	PR109	87.40	11.10

The correlation between the EBLUPs from selected model and traditional also is high, 0.975. However the group of test lines selected is different in each model. See this relation in Figure B.7.

Figure B.7. Dispersion plot of the genetic EBLUPs values of the Model 14 and the traditional non-spatial analysis model for this design (without spatial dependence) for 2nd cut of Paraná. The cut-offs for the 30 best test lines (7 % upper) in each Model are indicated by the dotted line.

B.3 Joint analysis

Comparing the predicted values from selected model, for first and second cut, the correlation is 0.542. Then, in the joint analysis is expected a significant correlation between the direct genetic effects. Figure B.8 displays the predicted values from the selected models. Note that there are less than half test lines selected in common between the cuts.

Figure B.8. Dispersion plot of the predicted values from selected model in each cut of Paraná. The cut-offs for the 30 best test lines (7% upper) in each Model are indicated by the dotted line.

The joint analysis started with the same effects as the individual analysis to which was added the fixed cut effect. Then, as summarized in Table B.12, models that assumed different covariance matrices for the genetic effects for the different cuts were compared. However, the models using unstructured general covariance matrix (us) and covariance matrices with heterogeneous variances (corh) had some convergence problems, because they estimate the correlation between genetic effects to be one. Then, the factor analytic model (fa(2)) was used to get a converged model and to verify the estimated value for the correlation. See in Table B.12 the models tested with REML log, the p -value of the REML ratio test and the AIC.

Table B.12. Summary of the fitted models to joint analysis of the Paraná experiments in the first (1) and second (2) cut, with restricted log likelihood (Log-lik.); p -value of the REML ratio test and AIC. All the models have the same fixed; local and global effects modeled in the individual analysis plus the cut fixed effect.

Model	Genetic effects	Constraint	Log-lik.	Test	p-value	AIC
1	$\begin{bmatrix} \sigma_{g1}^2 & 0 \\ 0 & \sigma_{g2}^2 \end{bmatrix}$		-3333.46			6694.92
2	$\begin{bmatrix} \sigma_g^2 & 0 \\ 0 & \sigma_g^2 \end{bmatrix}$	$\sigma_g^2 = \sigma_{g1}^2 = \sigma_{g2}^2$	-3335.41	M1 vs M2	0.02	6696.82
3	$\begin{bmatrix} \sigma_{g1}^2 & \rho_g \sigma_{g1} \sigma_{g2} \\ \rho_g \sigma_{g1} \sigma_{g2} & \sigma_{g2}^2 \end{bmatrix}$		-3264.52			6559.04
4	$\begin{bmatrix} \sigma_{g1}^2 & \sigma_{g1} \sigma_{g2} \\ \sigma_{g1} \sigma_{g2} & \sigma_{g2}^2 \end{bmatrix}$	$\rho_g=1$	-3464.51	M2 vs M3	1.00	6557.03

Using the AIC, Model 4 is chosen as the best to explain the effects in these experiments. This indicates that the correlation between the first and second cut is 1, then the test lines effects are the same, indicating that the best genotypes are the same for both cuts. Table B.13 displays the REML estimates of the variance parameters and their respective standard errors from fitted Model 4. As, the factor analytic model has been assumed the estimated genetic covariance matrix is:

$$\begin{bmatrix} \text{fa1} \\ \text{fa2} \end{bmatrix} \cdot \begin{bmatrix} \text{fa1} & \text{fa2} \end{bmatrix} + \begin{bmatrix} \text{var1} & 0 \\ 0 & \text{var2} \end{bmatrix} = \begin{bmatrix} 154.15 & 204.45 \\ 204.45 & 271.17 \end{bmatrix}.$$

Table B.13. REML estimates of variance parameters from fitted Model 4 to the joint analysis of the first-cut (1) and second-cut (2) of the experiment carried out in the Paraná.

Variance parameters		Estimate	Standard errors
Cut-1:Row	$(\tilde{\sigma}_{r1}^2)$	20.58	10.34
Cut-2:spl(Column)		38.05	42.59
Cut-2:spl(Row)		25.31	22.79
fa(Cut-1):Genetic	var1	0.00	—
fa(Cut-2):Genetic	var2	0.00	—
fa(Cut-1):Genetic	fa1	12.42	1.06
fa(Cut-2):Genetic	fa2	16.47	1.12
Cut-1:variance	$(\tilde{\sigma}_1^2)$	204.42	20.34
Spatial:row direction	$(\tilde{\rho}_{r11})$	0.05	0.07
Spatial:row direction	$(\tilde{\rho}_{r12})$	0.14	0.06
Spatial:row direction	$(\tilde{\rho}_{r13})$	0.16	0.06
Spatial:row direction	$(\tilde{\rho}_{r14})$	0.12	0.06
Cut-2:variance	$(\tilde{\sigma}_2^2)$	172.10	24.20
Spatial:row direction	$(\tilde{\rho}_{r21})$	-0.16	0.08
Spatial:row direction	$(\tilde{\rho}_{r22})$	0.02	0.08
Spatial:row direction	$(\tilde{\rho}_{r23})$	0.22	0.08

Table B.14 contains the Wald-F test for fixed terms of the Model 4. Note there are evidences of significant difference between the cuts, checks and test lines (Control) and within checks (Control:Check). The predicted values and standard errors of the 30 best test lines from Model 4 for each cut are in Table B.15. It can be observed that in the joint analysis the precision in the selection of the test lines is better, given that the predicted values are highest and standard errors are smaller, when compared with individual analysis (Tables B.6 and B.11). However, the correlation between the genetic predicted values from individual and joint analysis is high, 0.787 and 0.934 for first and second cut, respectively. This relationship can be observed in Figure B.9.

Table B.14. Wald F tests for fixed terms from Model 4. Df means degrees of freedom and DenDF is the approximate denominator degrees of freedom.

Fixed terms	Df	denDF	F test	p-value
(Intercept)	1.00	944.00	4766.00	0.00
Cut	1.00	944.00	46.71	0.00
Control	1.00	944.00	10.27	0.00
Cext	1.00	944.00	42.65	0.00
Control:Check	3.00	944.00	5.66	0.00
Cut:Control	1.00	944.00	0.34	0.56
Cut:Cext	1.00	944.00	3.48	0.06
Cut-2:harvest	1.00	944.00	5.04	0.03
Cut-2:lin(Col)	1.00	944.00	0.37	0.54
Cut-2:lin(Row)	1.00	944.00	2.74	0.10
Cut:Control:Check	3.00	944.00	0.62	0.60

Table B.15. Predicted values and standard errors of the 30 best test lines of the Model 4 for each cut.

First cut			Second cut		
Test line	Predicted value	Standard error	Test line	Predicted value	Standard error
PR96	100.55	7.01	PR96	109.82	9.57
PR171	96.25	7.02	PR171	104.11	9.62
PR279	92.73	7.02	PR279	99.44	9.77
PR331	91.65	7.11	PR331	98.02	9.79
PR295	91.15	6.96	PR295	97.36	9.64
PR14	90.06	7.12	PR14	95.91	9.94
PR90	89.78	6.99	PR90	95.53	9.62
PR56	89.63	6.91	PR56	95.34	9.66
PR358	89.18	7.01	PR358	94.74	9.74
PR327	88.76	6.97	PR327	94.18	9.82
PR217	88.67	7.06	PR217	94.06	9.87
PR133	88.03	7.01	PR133	93.21	9.54
PR204	87.77	6.97	PR204	92.87	9.67
PR232	87.49	7.07	PR232	92.49	10.05
PR143	87.27	7.05	PR143	92.20	9.77
PR386	86.68	6.99	PR386	91.42	9.66
PR52	86.45	7.04	PR52	91.12	9.84
PR98	86.23	6.95	PR98	90.82	9.47
PR6	85.30	7.02	PR6	89.60	9.70
PR150	85.22	6.94	PR150	89.48	9.55
PR207	84.47	7.06	PR207	88.50	10.01
PR197	84.28	7.07	PR197	88.25	9.71
PR298	83.94	6.94	PR298	87.79	9.70
PR9	83.86	7.01	PR9	87.68	9.81
PR71	83.68	7.04	PR71	87.44	9.72
PR109	83.53	7.04	PR109	87.24	9.71
PR208	83.49	6.97	PR208	87.19	9.33
PR309	83.43	6.95	PR309	87.11	9.68
PR304	83.20	7.01	PR304	86.81	9.54
PR113	83.03	7.07	PR113	86.58	9.80

Figure B.9. Predicted values from individual (indiv) and joint analysis of Paraná cuts. The left panel is for the first cut and the second is to the right.

B.4 Conclusions

For the first cut a model with fourth-order banded correlation structure in row direction was chosen. It is unusual for there to be as many as four bands and unexpected that because the correlations in the second to fourth bands are the similar magnitude, at around 0.18. This looks like an external effect of some machine or procedure. However the CTC company had no idea about what it could be. There is some evidence of competition at the residual level, because the row correlation for the first band is close to zero (0.02) and the second (0.17) is significant. The genetic variance is small, being approximately half of the residual variance.

For the second cut, the genetic variance is higher than the residual variance and so test-line selection will be more successful. The ratio between genetic and residual components is 1.31, more than the double of the first-cut. The residual correlation in row direction was modeled with third-order banded correlation structure and it shows more competition than in the first cut, because the first-order correlation is negative and significant. The correlations were -0.19, 0.03 and 0.19 for first to third order, respectively. The chosen model assumed spline effects in row and column directions and also the harvest effect was significant, differing from the first-cut that assumed a global row effect.

Comparing the best test lines selected in each cut from the individual analyses, there are only nine of the selected in common between the two cuts and the correlation between all the genetic predicted values was 0.542. This relation is stronger in the joint analysis, which found that genetic variance differed between the two cuts and the correlation between them was one in each cut and correlation one between these. Hence, using the joint analysis, it can be concluded that while the test-line effects differed between the two cuts, but that the ranking of the test lines is similar for the two cuts.

C PARANÁ - GROUP OF EXPERIMENTS

There were three sites, nominated: Local 651, 851 and 852. In one assessed the same 200 test lines (N) with 4 commercial varieties nominated “PAD2”, “PAD7”, “PAD8” and “Interc”. Planted in 2014 in a rectangular array of 10 columns by 24 rows and the yield in TCH was obtained in 2015, in a unique cut. The results were organized in subsections where C.1 presents the descriptive analysis of the sites, C.2 shows the individual analyses for each site and C.3 gives the joint analysis. The concluding remarks are in subsections C.4.

C.1 Descriptive analysis

The summary of the TCH for the groups of clones by each site is presented in Table C.1. It is possible to observe that the average yield is smaller in Local 651 than the others. The Local 851 resulted in higher average yield. Also the “PAD2” variance is higher, which is unexpected because it is a commercial variety and it should be stable. When the boxplot, Figure C.1, is observed, the outlier values and the distribution of the data can be verified. Only one plot was considered outlier within the plots of each commercial varieties and this was from Local 651.

Table C.1. Descriptive analysis of the groups of clones carried out in Paraná.

	Groups of clones	Minimum	Maximum	Mean	Variance
Local 651	Interspersed	26.80	95.20	72.23	287.98
	PAD2	43.20	89.30	76.27	489.70
	PAD7	67.00	86.30	77.37	71.91
	PAD8	40.20	75.90	51.32	272.71
	Test lines (N)	13.40	122.00	61.59	386.62
Local 851	Interspersed	91.70	138.10	115.35	187.37
	PAD2	78.60	183.30	130.05	1975.11
	PAD7	116.70	160.70	141.07	370.23
	PAD8	115.50	158.30	134.22	315.05
	Test lines (N)	72.60	221.40	133.32	636.74
Local 852	Interspersed	67.60	122.40	99.34	200.77
	PAD2	86.70	107.10	100.10	84.14
	PAD7	113.50	135.20	122.12	94.71
	PAD8	72.70	128.80	108.42	603.07
	Test lines (N)	51.00	167.10	101.79	469.47

Figure C.1. Boxplot of TCH for each group of lines for Locals of Paraná.

The Figures C.2, C.3 and C.4 show the heat maps relating to TCH in each site, and some trends can be observed in Local 651 (Figure C.2) and 851 (Figure C.3). It seems that there is one specific and unusual environmental effect in these sites, not only test treatment effect (test lines and checks). However the company were unaware of problems that could affect the results. The names “PAD2”, “PAD7”, “PAD8” and “INTERC” are the check plots, the test lines are labeled alpha-numerically (PR plus number) and “x” represents the empty plots.

Figure C.2. Heat map relating to TCH for Local 651 of Paraná.

Figure C.3. Heat map relating to TCH for Local 851 of Paraná.

Figure C.4. Heat map relating to TCH for Local 852 of Paraná.

C.2 Individual analysis

For Local 651 a summary of the sequence of fitted models is presented in Table C.2. Analysis commenced with the model denominated random effects Treatment Interference Model (R-TIM), used by Stringer et al. (2011). This contains random column, row and direct and neighbour genetic effects and first-order autoregressive variance model for the residuals in the row and column direction. Using REML tests and residual graphics new models were developed and in the end Model 20 was chosen to explain the data. Hence, in the chosen model, there are random direct and nugget genetic effects and first-order autoregressive structure for the residuals. The log-likelihood did not show strong difference among the models. However, in Figure C.5, it is clear that the spatial dependence, related to Model 20, fitted better.

Here, previous studies with banded structure were done, but it was observed no competition at the residual level. In this Local, the AR(1) structure appropriately described the spatial dependence. The magnitude of the correlation in the column and row directions is similar. See the REML estimates in Table C.3.

Table C.2. Summary of the fitted models to Local 651 with restricted log-likelihood (log-lik.) and the p -value of the REML ratio test. All models include a random direct genetic effects for the test lines (G) and fixed check varieties effects. The other effects are represented with the symbols: H (harvest); Co (cone); S (sowing); C (column); R (row); N (neighbour genetic); spl(.) (spline in some direction indicated in parentheses) and lin(.) (linear trend in some direction indicated in parentheses). The structures can be: US(.) - unstructured; AR(1) - first-order autoregressive; Band(.) - banded correlation, the number indicates the order of correlation; and Id - identity.

Model	Effects		Log-lik.	Test	p -value
	Global/extraneous	Genetic			
	Fixed	Random	Local (column $\times$ row)		
1		C + R	US(G:N)	AR(1) $\times$ AR(1)	-796.70
2			US(G:N)	AR(1) $\times$ AR(1)	-796.70
3			G	AR(1) $\times$ AR(1)	-797.89
4		C + R	G	AR(1) $\times$ Band(3)	-797.49
5			G	AR(1) $\times$ Band(3)	-797.49
6			G	AR(1) $\times$ Band(2)	-797.67
7			G	AR(1) $\times$ Band(1)	-798.50
8			G	AR(1) $\times$ Id	-801.91
9			G	Id $\times$ Band(1)	-806.18
10			G	AR(1) $\times$ AR(1)	-797.89
11	H		G	AR(1) $\times$ AR(1)	-795.63
12	Co		G	AR(1) $\times$ AR(1)	-795.24
13	S		G	AR(1) $\times$ AR(1)	-796.82
14	lin(C)+lin(R)	spl(C)+spl(R)	G	AR(1) $\times$ AR(1)	-796.77
15	lin(C)+lin(R)	spl(R)	G	AR(1) $\times$ AR(1)	-796.78
16	lin(C)+lin(R)		G	AR(1) $\times$ AR(1)	-797.99
17	lin(C)		G	AR(1) $\times$ AR(1)	-798.30
18	lin(R)		G	AR(1) $\times$ AR(1)	-797.57
19			G	AR(1) $\times$ AR(1)	-797.89
20		units	G	AR(1) $\times$ AR(1)	-795.36
21		units	US(G:N)	AR(1) $\times$ AR(1)	-795.27
				M1 vs M2	1.00
				M2 vs M3	0.21
				M4 vs M5	1.00
				M5 vs M6	0.55
				M6 vs M7	0.19
				M7 vs M8	0.01
				M7 vs M8	0.00
				not significant	
				not significant	
				not significant	
				M14 vs M15	1.00
				M15 vs M16	0.06
				not significant	
				not significant	
				M20 vs M19	0.01
				M21 vs M20	0.80

Figure C.5. Plots of the row and column faces of the empirical variogram for the residuals for the Local 651 of the Paraná experiment referring to models 19 (panels (a) and (c)) and 20 (panels (b) and (d)). The panels (a) and (b) are for the column direction and the others are for the row direction.

Table C.3. REML estimates of variance parameters for fitted Model 20 for the experiment in the Local 651 of the Paraná.

Variance parameters	Ratios (γ)	Estimates	Standard errors
Direct genetic (σ_g^2)	0.78	100.31	62.71
Nugget (σ_u^2)	1.15	147.68	61.37
Residual (σ^2)	1.00	127.92	45.92
Spatial ($\tilde{\rho}_c$)	—	0.67	0.16
Spatial ($\tilde{\rho}_r$)	—	0.56	0.18

Note that the direct genetic component is smaller than the residual (ratio of 0.78) and the nugget component is high (ratio of 1.15). Table C.4 presents the Wald F-tests for fixed terms. These shown there were no evidence of significant difference between the group of checks and test lines (Control) or among checks (Control:Check). The coefficients and standard errors of the factors levels from Model 20 are in Table C.5.

Table C.4. Wald F tests for fixed terms for Model 20. Df means degrees of freedom and DenDF is the approximate denominator degrees of freedom.

Fixed terms	Df	denDF	F test	p-value
(Intercept)	1	6.80	500.80	0.00
Control	1	6.10	0.76	0.42
Control:Check	3	30.40	2.49	0.08

Table C.5. Estimates of the fixed effects for Model 20 and their respective standard errors.

Effects	Estimates	Standard errors
INTERC	0.00	—
PAD2	6.37	8.03
PAD7	4.04	7.94
PAD8	-17.65	7.98
overall Test lines	0.00	—
overall Checks	9.61	10.47
overall mean (μ)	61.68	2.77

The top 7% of the test lines for Local 651 of the Paraná experiment, along with their predicted values and standard errors, are presented in Table C.6. In Figure C.6 can be observed the relation between the genetic EBLUPs from Model 20 and the traditional non-spatial analysis, the latter having only fixed checks varieties and random direct genetic, column and row effects. The correlation between them is high, 0.911, but the group of selected test lines was not the same.

Table C.6. The 14 best test lines with predicted values and respective standard errors for Model 20.

Test line	Predicted value	Standard error
PR58	78.25	8.71
PR65	75.98	8.71
PR156	75.17	8.70
PR61	75.02	8.71
PR136	73.94	8.71
PR46	73.67	8.70
PR151	72.81	8.70
PR123	72.65	8.70
PR143	72.49	8.72
PR135	71.83	8.70
PR169	71.75	8.70
PR174	71.65	8.70
PR7	71.39	8.71
PR116	70.93	8.70

Figure C.6. Dispersion plot of the genetic EBLUPs values of the Model 20 and the traditional non-spatial analysis model for this design (without spatial dependence) for Local 651 of the Paraná. The cut-offs for the 14 best test lines (7 % upper) in each Model are indicated by the dotted line.

For Local 851, we started with the same procedure. However, the neighbour genetic effects can be written as a special case, reduced rank, but using REML test (Table C.7) it was not significant.

Then testing local and global effects and using residual graphics, the better fit was found for Model 15, which has random spline effects in column direction and row and direct genetic effects. The model does not show significant competition in the residual or genetic level and when the empirical semi-variogram in Figure C.7 is observed there is any strong effect to be modeled, only some extraneous variability not detected by the selected model.

Table C.7. Summary of the fitted models to Local 851 with restricted log-likelihood (log-lik.) and the p -value of the REML ratio test. All models include a random direct genetic effects for the test lines (G) and fixed check varieties effects. The other effects are represented with the symbols: H (harvest); Co (cone); S (sowing); C (column); R (row); N (neighbour genetic); spl(.) (spline in some direction indicated in parentheses) and lin(.) (linear trend in some direction indicated in parentheses). The structures can be: RR(.) - reduced rank; AR(1) - first-order autoregressive; Band(.) - banded correlation, the number indicates the order of correlation; and Id - identity.

Model	Effects					Test	p -value
	Fixed	Global/extraneous	Random	Genetic	Local (column $\times$ row)		
1		C + R		RR(G:N)	AR(1) $\times$ AR(1)		
2		C + R		G	AR(1) $\times$ AR(1)	M1 vs M2	0.15
3		C + R		G	AR(1) $\times$ Band(3)		
4		C + R		G	AR(1) $\times$ Band(2)	M3 vs M4	0.14
5		C + R		G	AR(1) $\times$ Band(1)	M4 vs M5	0.78
6		C + R		G	AR(1) $\times$ Id	M5 vs M6	0.39
7		C + R		G	Id $\times$ Id	M6 vs M7	0.95
8	H			G	Id $\times$ Id	not significant	
9	Co			G	Id $\times$ Id	not significant	
10	S			G	Id $\times$ Id	not significant	
11	lin(C)+lin(R)	spl(C)+C+spl(R)+R		G	Id $\times$ Id		
12	lin(C)+lin(R)	spl(C)+R		G	Id $\times$ Id	M11 vs M12	1.00
13	lin(C)+lin(R)	R		G	Id $\times$ Id	M12 vs M13	0.01
14	lin(C)+lin(R)	spl(C)		G	Id $\times$ Id	M13 vs M14	0.00
15	lin(C)	spl(C)+R		G	Id $\times$ Id		
16	lin(C)	spl(C)+R		RR(G:N)	Id $\times$ Id	M16 vs M15	0.95

Figure C.7. Plot of the empirical semi-variogram for the residuals for the Local 851 of the Paraná experiment referring to Model 15.

Table C.8 presents the REML estimates of variance parameters from Model 15, where it can be observed that for this model the direct genetic component is 1.1 times the residual component. The spline column component is very small, with the standard error being greater than the estimated value; however this term was significant using REML ratio test.

Table C.8. REML estimates of variance parameters for fitted Model 15 for the experiment in the Local 851 of the Paraná.

Variance parameters	Ratios (γ)	Estimates	Standard errors
spl(Column)	0.04	8.43	11.09
Row (σ_r^2)	0.48	115.48	46.88
Direct genetic (σ_g^2)	1.10	261.76	85.36
Residual (σ^2)	1.00	238.74	64.88

Table C.9 presents the Wald F-tests for fixed terms of the Model 15 and Table C.10 shows the estimated fixed effects. As it can be seen, there is no evidence of a significant difference between the check and test lines groups (Control) or among the checks (Control:Check).

Table C.9. Wald F tests for fixed terms from Model 15. Df means degrees of freedom and DenDF is the approximate denominator degrees of freedom.

Fixed terms	Df	denDF	F test	p-value
(Intercept)	1.00	28.10	2441.00	0.00
Control	1.00	20.60	0.71	0.41
lin(Col)	1.00	179.60	1.10	0.29
Control:Check	3.00	32.70	3.95	0.02

Table C.10. Estimates of the fixed effects for Model 15 and their respective standard errors.

Effects	Estimates	Standard errors
INTERC	0.00	—
PAD2	20.91	8.74
PAD7	19.29	8.77
PAD8	18.05	8.74
lin(Col)	-0.54	0.47
overall test lines	0.00	—
overall checks	-19.28	16.52
overall mean (μ)	136.51	3.74

In Table C.11 there are the 14 best test lines from Model 15 with their respective predicted values and standard errors. Figure C.8 compares the genetic EBLUPs from Model 15 and the traditional non-spatial analysis, the latter including fixed check varieties and random direct genetic, column and row effects. The correlation between them is 0.931 which means there is a high similarity between the results.

Table C.11. The 14 best test lines with predicted values and respective standard errors for Model 15.

Test line	Predicted values	Standard.errors
PR135	163.33	11.81
PR139	154.52	11.97
PR53	154.20	12.14
PR180	152.44	12.17
PR156	152.05	11.85
PR49	151.95	11.98
PR65	151.76	12.15
PR23	151.70	12.18
PR173	151.05	12.04
PR186	149.95	11.90
PR184	149.95	11.88
PR25	148.76	12.16
PR174	148.60	12.03
PR12	146.65	11.92

Figure C.8. Dispersion plot of the genetic EBLUPs values of the Model 15 and the traditional non-spatial analysis model for this design (without spatial dependence) for Local 851 of the Paraná. The cut-offs for the 14 best test lines (7 % upper) in each Model are indicated by the dotted line.

For Local 852, Table C.12 shows some fitted models starting with the model R-TIM, with random column, row and genetic (direct, neighbour and covariance) effects and first-order autoregressive structure in the row and column direction. The random row, column, genetic and neighbour covariance effects are very small components. They do not represent significant effects as can be observed in Table C.12 and it is not a special case of reduced rank. However, a negative correlation was found in the row direction which indicates competition among the plots. The model with third-order banded structure has problems of convergence. Hence, it is difficult to analyze the correlation levels among the plots and to decide which is the better structure to explain the data. Using the empirical semi-variogram for the residuals, it was possible to compare the models and to choose Model 15 as the best to explain the data. Figure C.9 shows the empirical semi-variogram for the residuals from the model 15.

Table C.12. Summary of the fitted models to Local 852 with restricted log-likelihood (log-lik.) and the p -value of the REML ratio test. All models include a random direct genetic effects for the test lines (G) and fixed check varieties effects. The other effects are represented with the symbols: H (harvest); Co (cone); S (sowing); C (column); R (row); N (neighbour genetic); spl(.) (spline in some direction indicated in parentheses) and lin(.) (linear trend in some direction indicated in parentheses). The structures can be: US(.) - unstructured; RR(.) - reduced rank; AR(.) - autoregressive structure; Band(.) - banded correlation, the number indicates the order of correlation; sar(2) constrained autoregressive; and Id - identity.

Model	Effects				Log-lik.	Test	p -value
	Fixed	Global/extraneous	Random	Genetic	Local (column $\times$ row)		
1		C + R		US(G:N)	AR(1) $\times$ AR(1)	-825.74	
2		C + R		G	AR(1) $\times$ AR(1)	-826.31	M1 vs M2 0.42
3		C + R		G	AR(1) $\times$ Band(3)	not converged-singularity	
4		C + R		G	AR(1) $\times$ Band(2)	-824.45	
5		C		G	AR(1) $\times$ Band(2)	-824.45	M4 vs M5 1.00
6		C		G	AR(1) $\times$ Band(1)	-828.14	M5 vs M6 0.00
7		C		G	AR(1) $\times$ sar(2)	-824.24	
8		C		G	AR(1) $\times$ AR(2)	-825.53	
9		C		G	Id $\times$ Band(1)	-828.14	M5 vs M9 0.14
10	H	C		G	Id $\times$ Band(2)	-824.05	not significant
11	Co	C		G	Id $\times$ Band(2)	-824.59	not significant
12	S	C		G	Id $\times$ Band(2)	-824.25	not significant
13	lin(C)+lin(R)	spl(C)+C+spl(R)		G	Id $\times$ Band(2)	-822.67	
14	lin(C)+lin(R)	C+spl(R)		G	Id $\times$ Band(2)	-822.67	M13 vs M14 1.00
15	lin(C)+lin(R)	spl(R)		G	Id $\times$ Band(2)	-822.76	M14 vs M15 0.33
16	lin(C)+lin(R)			G	Id $\times$ Band(2)	-824.87	M15 vs M16 0.02
17	lin(C)+lin(R)	spl(R)		RR(G:N)	Id $\times$ Band(2)	-821.51	M17 vs M15 0.11

Figure C.9. Plot of the empirical semi-variogram for the residuals for the Local 852 of the Paraná experiment referring to Model 15.

Table C.13 presents the REML estimates of variance parameters from selected model. The direct genetic component is 1.3 times the residual component and there is a negative correlation in row direction shown that have the competition at the residual level. The spline row component is very small, with standard error grater than the component, although this term was significant using REML test.

Table C.13. REML estimates of variance parameters for Model 15 to the experiment in the Local 852 of the Paraná.

Variance parameters		Ratios (γ)	Estimates	Standard errors
spl(Row)		0.03	6.03	8.17
Direct genetic	$(\tilde{\sigma}_g^2)$	1.30	252.75	57.22
Residual	$(\tilde{\sigma}^2)$	1.00	195.12	43.15
Spatial	$(\tilde{\rho}_{r_1})$	—	-0.40	0.13
Spatial	$(\tilde{\rho}_{r_2})$	—	0.29	0.12

The Wald F tests for fixed terms are present in Table C.14. They show that there were no evidence to assume difference between the group of test lines and checks (Control), but that there was among the checks (Control:Check). The estimated fixed effects from this model are in Table C.15.

Table C.14. Wald F tests for fixed terms for Model 15. Df means degrees of freedom and DenDF is the approximate denominator degrees of freedom.

Fixed terms	Df	denDF	F test	p-value
(Intercept)	1.00	228.00	5169.00	0.00
Control	1.00	228.00	0.03	0.87
lin(Col)	1.00	228.00	9.98	0.00
lin(Row)	1.00	228.00	0.03	0.87
Control:Check	3.00	228.00	4.95	0.00

Table C.15. Estimates of the fixed effects for Model 15 and their respective standard errors.

Effects	Estimates	Standard errors
INTERC	0.00	—
PAD2	-0.79	6.72
PAD7	19.60	6.93
PAD8	12.52	6.82
lin(Col)	1.21	0.45
overall test lines	0.00	—
overall checks	-1.69	16.65
overall mean (μ)	95.04	2.93

In Table C.16 are the 14 best test lines selecting using Model 15, along with their respective predicted values and standard errors. Comparing, in Figure C.10, the genetic EBLUPs from Model 15 and the traditional non-spatial analysis, that contain fixed checks and random direct genetic, row and column effects, there is a linear trend and the correlation is 0.938. However, note that the group of selected test lines is not the same.

Table C.16. The 14 best test lines with predicted values and respective standard errors for Model 15.

Test line	Predicted values	Standard errors
PR50	139.26	9.83
PR66	135.03	10.15
PR173	131.05	10.14
PR86	126.50	10.02
PR136	123.85	10.06
PR64	122.68	10.02
PR151	122.24	10.01
PR122	122.12	9.79
PR49	121.21	10.27
PR37	119.69	10.64
PR177	118.14	10.29
PR119	118.11	9.98
PR35	117.57	10.02
PR100	116.97	9.75

Figure C.10. Dispersion plot of the genetic EBLUPs values of the Model 15 and the traditional non-spatial analysis model for this design (without spatial dependence) for Local 852 of the Paraná. The cut-offs for the 14 best test lines (7 % upper) in each Model are indicated by the dotted line.

C.3 Joint analysis

Based on the separate analysis for each site, it was observed that the genetic EBLUPs are very different. Figure C.11 presents the relation of these values between pairs of Locals; the correlation is medium, among 0.330 to 0.493. This means that there is a substantial Local effect on the test lines.

Figure C.11. Dispersion plots of the genetic EBLUPs from model selected for each Paraná Local. Panels the relationship between pairs of Locals (a) 651 and 851; (b) 651 and 852; and (c) 851 and 852. The cut-offs for the 14 best test lines in each Local are indicated by the dotted line.

Then, the joint analysis of all three sites was performed in order to investigate and to compare the magnitudes of the genetic variances at each Local and to assess the genetic covariance between Locals. However, there were some convergence problems. Hence, firstly only Local 851 and 852 were jointly analyzed because their genetic components were closer in the individual analyses. Table C.17 presents the models and the REML test for the joint analysis between Local 851 and 852. Using the REML test, it is possible to conclude that the genetic components are similar and there is a correlation of 0.9 between Locals 851 and 852 (Model 3). These results were used in the joint analysis with all sites.

Table C.17. Summary of the fitted models to joint analysis of the Paraná experiments carried out in the Locals 851 (2) and 852 (3), with REML log (log-lik.) and the p -value of the REML ratio test. All the models have the same effects as modeled in the individual analysis plus the fixed Local effect. Here σ_{gi}^2 is the genetic variance at the i th Local and $\rho_{gii'}$ is the genetic correlation between the i th and i' th Locals.

Model	Genetic effects	Constraint	log-lik.	Test	p-value
1	$\begin{bmatrix} \sigma_{g2}^2 & 0 \\ 0 & \sigma_{g3}^2 \end{bmatrix}$		-1690.63		
2	$\begin{bmatrix} \sigma_g^2 & 0 \\ 0 & \sigma_g^2 \end{bmatrix}$	$\sigma_g^2 = \sigma_{g2}^2 = \sigma_{g3}^2$	-1680.64	M1 vs M2	0.46
3	$\begin{bmatrix} \sigma_g^2 & \sigma_{gg} \\ \sigma_{gg} & \sigma_g^2 \end{bmatrix}$	$\sigma_{gg} = \rho_g \sigma_g^2$	-1651.92	M3 vs M2	0.00

For the joint analysis, there are convergence problems with some models, mainly when unstructured variance model is used for genetic effects. Therefore, the results from the models in Table C.17 were used as initial parameters in fitting some models. The tested models are presented in Table C.18. Within the possible models, Model 3 was considered as the best to explain the data. It has the same environmental parameters of the individual analysis, plus the Local fixed effect and covariance between genetic and Local effect. The genetic variances for Local 851 and 852 are considered equal, as are the covariances between the genetic effect for all Locals. Table C.19 presents the REML estimates of variance components of the Model 3. Note that the direct genetic component for Local 651 was less than half when compared to the ones for Local 851 and 852. However site 651 had the smaller residual component.

Table C.18. Summary of the models fitted in joint analysis of the Paraná experiments carried out in the Locals 561 (1), 851 (2) and 852 (3), with REML ratio tests (log-lik.) and the p -values of the REML ratio tests. All the models have the same effects as modeled in the individual analysis plus the fixed Local effect. Here σ_{gi}^2 is the genetic variance at the i th Local, $\rho_{ii'}$ is the genetic correlation between the i th and i' th Locals and σ_{g*} is the genetic covariance between two locals which have been hypothesis to be equal for two or more pairs of Locals.

Model	Genetic effects	Constrained	Log-lik.	Test	p-value
1	$\begin{bmatrix} \sigma_{g1}^2 & 0 & 0 \\ 0 & \sigma_{g2}^2 & 0 \\ 0 & 0 & \sigma_{g3}^2 \end{bmatrix}$		-2476.00		
2	$\begin{bmatrix} \sigma_{g1}^2 & \sigma_{g12} & \sigma_{g13} \\ \sigma_{g12} & \sigma_{g2}^2 & \sigma_{g23} \\ \sigma_{g13} & \sigma_{g23} & \sigma_{g3}^2 \end{bmatrix}$	$\sigma_{gii'} = \rho_g \sigma_{gi} \sigma_{gi'}$	-2427.85	M2 vs M1	0.00
3	$\begin{bmatrix} \sigma_{g1}^2 & \sigma_{g1*} & \sigma_{g1*} \\ \sigma_{g1*} & \sigma_g^2 & \sigma_{gg} \\ \sigma_{g1*} & \sigma_{gg} & \sigma_g^2 \end{bmatrix}$	$\sigma_{g1*} = \rho_{g1} \sigma_{g1} \sigma_g$ $\sigma_{gg} = \rho_{g23} \sigma_{g2}^2$	-2428.59	M2 vs M3	0.35

Table C.19. REML estimates of the variance parameters from fitted Model 3 in the joint analysis of the Locals in Paraná.

Variance parameters		Estimates	Standard errors
Local651:Genetic	$(\tilde{\sigma}_{g1}^2)$	91.35	60.34
Local851-852:Genetic	$(\tilde{\sigma}_g^2)$	261.58	41.38
cov(Local 651-...):New	$(\tilde{\sigma}_{g1*})$	130.79	20.69
cov(Local 851-852):New	$(\tilde{\sigma}_{gg})$	239.13	37.62
Local 851:spl(Col)		5.63	7.60
Local 851:Row	$(\tilde{\sigma}_{r2}^2)$	108.75	41.82
Local 852:spl(Row)		5.54	7.10
Local 561: Nugget	$(\tilde{\sigma}_{u1}^2)$	154.53	58.62
Local 561: Residual	$(\tilde{\sigma}_1^2)$	120.37	40.23
Spatial:Column	$(\tilde{\rho}_{1c}^2)$	0.72	0.14
Spatial:Row	$(\tilde{\rho}_{1r}^2)$	0.52	0.17
Local 851:Residual	$(\tilde{\sigma}_2^2)$	230.43	39.21
Local 852: Residual	$(\tilde{\sigma}_3^2)$	187.52	32.39
Spatial:Row	$(\tilde{\rho}_{3r1}^2)$	-0.32	0.10
Spatial:Row	$(\tilde{\rho}_{3r2}^2)$	0.11	0.10

Table C.20 presents the Wald F-tests for fixed terms. Observe that there is the significant difference between the Locals and between check and test lines group. The coefficients and standard errors of these fixed effects from Model 3 are in Table C.21.

Table C.20. Wald F test for fixed terms of the Model 3. Df means degrees of freedom and DenDF is the approximate denominator degrees of freedom.

Fixed terms	Df	denDF	F test	p-value
(Intercept)	1.00	683.00	5972.00	0.00
Site	2.00	683.00	192.20	0.00
Control	1.00	683.00	2.37	0.12
Control:Check	3.00	683.00	5.44	0.00
Site:Control	2.00	683.00	3.19	0.04
at(Site, 851):lin(Col)	1.00	683.00	0.00	0.98
at(Site, 852):lin(Col)	1.00	683.00	12.90	0.00
at(Site, 852):lin(Row)	1.00	683.00	0.59	0.44
Site:Control:Check	6.00	683.00	2.93	0.01

Table C.21. Estimates of the fixed effects from Model 3 and their respective standard errors.

Effects	Estimates	Standard errors
overall mean (μ)	61.89	2.71
INTERC	0.00	—
PAD2	5.11	7.96
PAD7	4.08	7.89
PAD8	-18.28	7.91
Checks	9.89	10.03
Local 651	0.00	—
INTERC	0.00	—
PAD2	0.00	—
PAD7	0.00	—
PAD8	0.00	—
Checks	0.00	—
Local 851	74.29	4.33
INTERC	0.00	—
PAD2	17.25	11.65
PAD7	16.08	11.61
PAD8	36.51	11.61
lin(Col)	-0.49	0.42
Checks	-29.15	10.50
Local 852	34.11	3.78
INTERC	0.00	—
PAD2	-7.19	10.67
PAD7	16.50	10.71
PAD8	28.04	10.63
lin(Row)	-0.09	0.13
lin(Col)	1.23	0.34
Checks	-10.69	10.39

Comparing the predicted values and their standard errors from Model 3 for the 14 best test lines from the different Locals, given in Table C.22, it is observed that (i) these results have small standard errors when compared to the separate analysis of the sites, and (II) a large number of clones in common among the Locals within the top 14 test lines.

Table C.22. Predicted values (pred.) and standard error (error) of the 14 best test lines from Model 3 for each Local.

Local 651			Local 851			Local 852		
Test lines	Pred.	Error	Test lines	Pred.	Error	Test lines	Pred.	Error
PR135	80.04	6.57	PR173	164.68	9.44	PR173	133.91	8.81
PR173	79.00	6.52	PR135	164.22	9.38	PR50	132.54	8.65
PR156	78.89	6.53	PR50	160.71	9.43	PR66	128.74	8.87
PR151	78.71	6.53	PR156	158.36	9.35	PR135	127.77	8.99
PR136	77.63	6.51	PR151	157.88	9.34	PR151	127.24	8.79
PR50	77.47	6.53	PR66	156.72	9.51	PR136	124.01	8.78
PR61	75.73	6.54	PR53	155.28	9.49	PR156	123.99	8.68
PR37	75.62	6.58	PR136	153.88	9.33	PR37	121.72	9.02
PR174	75.36	6.54	PR139	153.70	9.40	PR49	121.23	8.88
PR53	75.32	6.54	PR49	153.63	9.43	PR23	120.62	8.68
PR7	74.89	6.52	PR23	153.47	9.49	PR184	119.82	8.78
PR123	74.82	6.53	PR174	152.95	9.43	PR122	119.68	8.69
PR66	74.77	6.54	PR184	152.84	9.35	PR53	119.54	8.84
PR65	74.41	6.53	PR37	152.78	9.39	PR86	119.50	8.83

C.4 Conclusions

The individual analyses resulted in different fitted models for each site. The model in Local 651 shows that there is no residual competition, because the correlations, modeled with AR(1) in both directions, are positive and similar in value. The neighbour genetic effects were not significant, neither random column or row effect, but there is a nugget effect. For Local 851 there is not significant spatial dependence neither competition at the genetic or residual level, the selected model had a significant spline term in the column direction, random row and direct genetic effects. Local 852 exhibited residual competition effects, with first-order correlation in row direction equal to -0.40 while the second-order is 0.29. The selected model had linear trend effect in column direction, spline term in the row direction and direct genetic effects. For local effect the second-order banded structure in row direction was adequate. The ratios between genetic and residual components were approximately 0.78, 1.10 and 1.30 for Locals 561, 851 and 852, respectively. This results in better selection of genotypes for Local 851 and 852 because they have larger genetic variance relative to the residual variance. When comparing the 14 best-predicted values for the test lines from each Local there were few similarities among the lines selected; there were around 2 to 5 common test lines and the correlation between the sites for all values was among 0.33 and 0.49.

However, in the joint analysis the selected model indicates the same genetic component for Local 851 and 852 (261.58) and smaller component for Local 651 (91.35). There is genetic covariance between Locals, which for 651 with 851 or 852 is 239.13, while between 851 and 852 is estimated to be 130.79. The number of common test lines in the best 14 lines is greater; more than a half is the same in all three Locals. In addition, better accuracy in the selection of the test lines is obtained, with predictions having lower standard errors.

D GOIÁS - GROUP OF EXPERIMENTS

There were three sites, nominated: Local 3, 521 and 533. The same 212 test lines (N) with 4 commercial varieties nominated “PAD1”, “PAD4”, “PAD9” and “Interc” were assessed at all sites. They were planted in 2014 in a rectangular array of 13 columns by 20 rows and the yield in TCH was obtained in 2015, in a unique cut. The results were organized in subsections where D.1 presents the descriptive analysis of the sites, D.2 shows the individual analyses for each Local, D.3 gives the joint analysis and D.4 present some conclusions.

D.1 Descriptive analysis

A summary of TCH for the groups of clones by each Local is shown in Table D.1. Considering averages yield, there were no great differences among the sites and the commercial varieties (checks). For Local 3, “PAD4” yields between 97.6 and 176.2; Local 521 and 533 also exhibit similar minimum and maximum yield values, but is for “PAD1”. When the boxplot is observed, Figure D.1, some outlier values are found for each group of clones given each site. These plots will be carefully monitored in the analyses to check whether they result in problems such as being an influential plot or an outlier. Note that the variance and average yield of the commercial varieties (checks) were different among the Locals.

Table D.1. Descriptive analysis of TCH for the groups of clones carried out in Goiás.

	Groups of clones	Minimum	Maximum	Mean	Variance
Local 3	Inerspersion	97.60	202.40	145.91	436.31
	PAD1	119.00	181.00	142.65	451.29
	PAD4	97.60	176.20	137.70	659.41
	PAD9	116.70	160.70	138.30	352.20
	Test lines (N)	67.90	264.30	125.37	637.07
Local 521	Inerspersion	95.20	157.10	132.26	297.45
	PAD1	97.60	163.10	130.35	461.09
	PAD4	107.10	140.50	121.03	180.33
	PAD9	115.50	160.70	130.55	270.69
	Test lines (N)	53.60	179.80	103.30	531.69
Local 533	Inerspersion	138.10	203.60	161.50	241.78
	PAD1	91.70	167.90	127.78	626.22
	PAD4	120.20	152.40	129.16	137.96
	PAD9	144.00	189.30	165.85	360.93
	Test lines (N)	51.20	221.40	135.36	657.31

Figure D.1. Boxplot of TCH for each group of lines for Locals of Goiás.

The Figures D.2, D.3 and D.4 show the heat maps relating to TCH in each site. The names “PAD1”, “PAD4”, “PAD9” and “INTERC” are the check plots and the test lines are labeled alpha-numerically (GS plus number). It seems there is in Local 3 a false measure in the “GS51” plot, because this had a much higher yield than expected, 264.30 TCH. Observe that its neighbour, “GS33”, had small yield, but it also was small in Local 521 and 533. Then, it was supposed that there were problems in the harvest. Maybe, “GS51” gained yield from neighbours and its value is flawed. However, this supposition was not confirmed by the CTC. For the other two sites there were no strong visual problems or discrepant measures.

Figure D.2. Heat map relating to TCH for Local 3 of Goiás.

Figure D.3. Heat map relating to TCH for Local 521 of Goiás.

Figure D.4. Heat map relating to TCH for Local 533 of Goiás.

D.2 Individual analysis

For the Local 3, the plot “GS51” was analyzed with care, because there was indication of measurement problems. The CTC send to us the lodging score of this experiment and the plot did not have the high score (classified as medium) but their neighbours “GS41” and “GS46” had. However there is no information about how much the lodging effect could affect the yield and if it may have caused serious problems in separating the material from different plots at harvest. As “GS51” had an unique information in this site and the residuals of the models did not have any problems, no plot was removed.

But, observing Figure D.5, that there the lodging score, it is possible to note that are more than 15% of plots with strong lodging (score among 8 and 9).

Figure D.5. Heat map relating to fall down for Local 3 of Goiás.

In Table D.2 a sequence of fitted models is presented. We started with the R-TIM model with random genetic, column and row effects and AR(1) in both directions. The final model was selected using the REML ratio test and the residual plots. At the beginning there is not significant competition effect at the genetic level, but a negative correlation in row direction represents the presence of competition at the residual level. In sequence, different local and global effects were tested and the chosen model was Model 19, with: fixed check varieties and harvest effects; spline and linear trend in row direction; random column and direct genetic effects; and residual structure modeled with AR(1) and Band(2) in column and row direction, respectively. Figure D.6 has the plots of empirical semi-variogram from models 19 and 21. Note that Figure D.6(a) shows a competition effect which is accounted for in Model 21, Figure D.6(b). Model 21 had small residual and direct genetic components when compared with Model 19. Hence, given that the neighbour genetic component was not significant using a REML ratio test, Model 21 was not chosen.

Table D.2. Summary of the fitted models to Local 3 of the Goiás experiment with restricted log-likelihood (log-lik.) and the p -value of the REML ratio test. All models include a random direct genetic effect for the test lines (G) and fixed check varieties effect. The other effects are represented with the letters: H (harvest); Co (cone); S (sowing); C (column); R (row); N (neighbour genetic); spl(.) (spline in some direction indicated in parentheses) and lin(.) (linear trend in some direction indicated in parentheses). The structures can be: US(.) - unstructured; AR(.) - autoregressive; sar(2) - constrained autoregressive; Band(.) - banded correlation, the number indicates the order of correlation; and Id - identity.

Model	Effects				Test	p -value
	Fixed	Global/extraneous	Random	Genetic (column $\times$ row)	Local (column $\times$ row)	
		Extraneous				
1		R + C		US(G:N)	AR(1) $\times$ AR(1)	-929.19
2		R + C		G	AR(1) $\times$ AR(1)	-930.31
3		R + C		G	AR(1) $\times$ Band(3)	not converged - singularity
4		R + C		G	AR(1) $\times$ Band(2)	-928.80
5		R + C		G	AR(1) $\times$ Band(1)	not converged - singularity
6		R + C		G	AR(1) $\times$ AR(2)	-929.70
7		R + C		G	AR(1) $\times$ AR(3)	-929.01
8		R + C		G	AR(1) $\times$ sar(2)	not converged - singularity
9		R + C		G	AR(2) $\times$ Band(2)	-927.76
10		R + C		G	Id $\times$ Band(2)	-930.77
11	Co	R + C		G	AR(1) $\times$ Band(2)	-927.32
12	H	R + C		G	AR(1) $\times$ Band(2)	-925.16
13	S	R + C		G	AR(1) $\times$ Band(2)	-926.17
14	H+lin(C)+lin(R)	spl(C)+C+spl(R)+R		G	AR(1) $\times$ Band(2)	-923.29
15	H+lin(C)+lin(R)	spl(C)+C+spl(R)		G	AR(1) $\times$ Band(2)	-923.39
16	H+lin(C)+lin(R)	spl(C)+spl(R)		G	AR(1) $\times$ Band(2)	-925.88
17	H+lin(C)+lin(R)	spl(C)+C		G	AR(1) $\times$ Band(2)	-929.77
18	H+lin(C)+lin(R)	C+spl(R)		G	AR(1) $\times$ Band(2)	-923.58
19	H+lin(R)	C+spl(R)		G	AR(1) $\times$ Band(2)	-923.12
20	H+lin(R)	C+spl(R)	US(G:N)	US(G:N)	AR(1) $\times$ Band(2)	not converged - singularity
21	H+lin(R)	C+spl(R)	US(G:N)	US(G:N)	AR(1) $\times$ AR(1)	-923.89

Figure D.6. Plots of the empirical semi-variogram for the residuals for the Local 3 of the Goiás experiment referring to models 19 and 21.

The REML estimates of the variance parameters for Model 19 are presented in Table D.3. Note that there is negative correlation in the column direction. However, in this direction there is 1 meter of space between plots and the boundary is small (2.4 meters). Then, this negative correlation in the column direction is unlikely to arise from competition between plots, as it does in the row direction. The direct genetic component is small than the residual component (ratio of 0.84) and the first-order correlation in row direction is negative and the bigger found in all experiments. The standard error for the spline is biggest than its value, although the REML ratio test is significant for this term.

Table D.3. REML estimates of variance parameters from fitted Model 19 to the experiment in the Local 3 of the Goiás.

Variance parameters		Ratios (γ)	Estimates	Standard errors
Column	$(\tilde{\sigma}_c^2)$	0.12	39.22	21.62
spl(Row)		0.02	7.62	8.80
Direct genetic	$(\tilde{\sigma}_g^2)$	0.84	269.54	61.19
Residual	$(\tilde{\sigma}^2)$	1.00	320.09	57.42
Spatial	$(\tilde{\rho}_{c_1})$	—	-0.28	0.11
Spatial	$(\tilde{\rho}_{r_1})$	—	-0.61	0.10
Spatial	$(\tilde{\rho}_{r_2})$	—	0.21	0.08

The Wald F tests for the fixed terms are in Table D.4. Note that there are no evidence of significant difference between the group of checks and test lines (Control) or among the checks (Control:Checks). Table D.5 presents the estimated fixed effects for Model 19.

Table D.4. Wald F tests for fixed terms for Model 19. Df means degrees of freedom and DenDF is the approximate denominator degrees of freedom.

Fixed terms	Df	denDF	F test	p-value
(Intercept)	1.00	246.00	3403.00	0.00
Control	1.00	246.00	0.92	0.34
colheita	1.00	246.00	5.86	0.02
lin(Row)	1.00	246.00	0.11	0.74
Control:Check	3.00	246.00	1.95	0.12

Table D.5. Estimates of the fixed effects for Model 19 and their respective standard errors.

Effects	Estimates	Standard errors
PAD1	-7.75	6.46
PAD4	-9.47	6.48
PAD9	8.60	6.63
lin(Row)	-0.09	0.16
harvest-level 1	0.00	—
harvest-level 2	8.51	3.16
overall test lines	0.00	—
overall check	17.04	16.75
overall mean (μ)	122.58	3.11

The 15 best test lines with predicted values and respective standard errors from Model 19 are presented in Table D.6. When comparing the genetic EBLUPs values from this model with the ones from the traditional non-spatial analysis, which contains fixed checks varieties and random direct genetic, column and row effects, there are some trends and the correlation between them is 0.854. See Figure D.7.

Table D.6. The 15 best test lines with predicted values and respective standard errors from Model 19.

Test lines	Predicted values	Standard errors
GS51	186.41	10.93
GS204	167.82	11.01
GS87	160.93	10.66
GS53	160.15	10.39
GS132	157.63	11.08
GS99	156.24	11.12
GS151	155.64	11.61
GS141	155.21	11.16
GS210	154.84	11.17
GS2	154.50	11.09
GS79	154.15	10.98
GS43	153.41	10.57
GS102	152.55	11.37
GS165	150.69	10.53
GS201	149.94	11.33

Figure D.7. Dispersion plot of the genetic EBLUPs values of the Model 19 and the traditional non-spatial analysis model for this design (without spatial dependence) for Goiás Local 3. The cut-offs for the 15 best test lines (7 % upper) in each model are indicated by the dotted line.

For Local 521 some models were tested and some of them are presented in Table D.7. With this data set several models had convergence problems. Then, it was difficult to decide about what effect should be true and which was confounded effect. To assess the local effect, the genetic competition and correlation genetic were removed in some models to obtain convergence, but these were not efficient for all structures tested. Using REML ratio test and residuals plots, the Model 20 was chosen as the best to fit to the data. It assumes: fixed check varieties; linear trend in the column and row directions; row spline; random direct, neighbour and direct-neighbour covariance genetic effects. See in Figure D.8 the plots of empirical semi-variogram from fitted model 20 and 23, where we did not assume the competition at the genetic level. Table D.8 presents the REML estimates of variance parameters from this fitted models.

Table D.7. Summary of the fitted models to Local 521 of the Goiás experiment with restricted log-likelihood (log-lik.) and the p -value of the REML ratio test. All the models include a random direct genetic effects for the test lines (G) and fixed check varieties effects. The other effects are represented with the letters: H (harvest); Co(cone); S (sowing); C (column); R (row); N (neighbour genetic); spl(.) (spline in some direction indicated in parentheses) and lin(.) (linear trend in some direction indicated in parentheses). The structures can be: US(.)- unstructured; RR(.)- reduced rank; AR(.) - autoregressive; Band(.) - banded correlation, the number indicates the order of correlation; sar(2) - constrained autoregressive; and Id - identity.

Model	Effects				Log-lik.	Test	p -value
	fixed	Global/extraneous	random	Genetic	Local (column $\times$ row)		
1		R + C		US(G:N)	AR(1) $\times$ AR(1)	-901.42	
2		R + C		RR(G:N)	AR(1) $\times$ AR(1)	-903.54	M2 vs M1 0.02
3		R + C		G	AR(1) $\times$ AR(1)	-904.18	M3 vs M1 0.04
4		R + C		US(G:N)	AR(1) $\times$ Band(3)	not converged - singularity	
5		R + C		US(G:N)	AR(1) $\times$ Band(2)	not converged - singularity	
6		R + C		US(G:N)	AR(1) $\times$ Band(1)	not converged - singularity	
7		R + C		US(G:N)	AR(1) $\times$ ar(2)	not converged - singularity	
8		R + C		US(G:N)	AR(1) $\times$ sar(2)	not converged - singularity	
9		R + C		G	AR(1) $\times$ Band(3)	-903.70	
10		R + C		G	AR(1) $\times$ Band(2)	not converged - singularity	
11		R + C		G	AR(1) $\times$ Band(1)	not converged - singularity	
12		R + C		G	AR(1) $\times$ AR(2)	-904.02	M12 vs M3 0.57
13		R + C		G	AR(1) $\times$ sar(2)	problems of convergence	
14		R + C		US(G:N)	AR(1) $\times$ Id	-901.62	M1 vs M14 0.52
15		R + C		US(G:N)	Id $\times$ Id	-902.40	M14 vs M15 0.21
16	Co	R + C		US(G:N)	Id $\times$ Id	-901.01	not significant
17	H	R + C		US(G:N)	Id $\times$ Id	-900.88	not significant
18	S	R + C		US(G:N)	Id $\times$ Id	-900.98	not significant
19	lin(C)+lin(R)	spl(C)+C+spl(R)+R		US(G:N)	Id $\times$ Id	-900.40	
20	lin(C)+lin(R)	C+spl(R)		US(G:N)	Id $\times$ Id	-900.40	M19 vs M20 1.00
21	lin(C)+lin(R)	C		US(G:N)	Id $\times$ Id	-902.15	M20 vs M21 0.03
22	lin(C)+lin(R)	spl(R)		US(G:N)	Id $\times$ Id	-902.28	M20 vs M22 0.02
23	lin(C)+lin(R)	C+spl(R)		G	Id $\times$ Id	-914.47	M20 vs M23 0.00
24	lin(C)+lin(R)	C+spl(R)		RR(G:N)	Id $\times$ Id	-914.47	M20 vs M24 0.04

Figure D.8. Plots of the empirical semi-variogram for the residuals for Local 521 of Goiás experiment referring to Models (a) 20 and (b) 23.

Table D.8. REML estimates of variance parameters for Model 20 fitted to the data from Local 521 from the Goiás experiment.

Variance parameters		Ratios (γ)	Estimates	Standard errors
Column	$(\tilde{\sigma}_c^2)$	0.18	15.91	12.21
spl(Row)		0.41	36.33	34.31
Direct genetic	$(\tilde{\sigma}_g^2)$	3.44	307.26	60.98
Covariance genetic	$(\tilde{\sigma}_{gn})$	-0.97	-86.31	20.46
Neighbour genetic	$(\tilde{\sigma}_n^2)$	0.67	59.76	25.39
Residual	$(\tilde{\sigma}^2)$	1.00	89.35	47.76

Table D.9 shows the Wald F tests for the fixed terms and Table D.10 the estimated fixed effects from the chosen model. Note that there are no evidences of significant difference between the group of checks with test lines (Control) or among the checks (Control:Checks).

Table D.9. Wald F tests for fixed terms from Model 20. Df means degrees of freedom and DenDF is the approximate denominator degrees of freedom.

Fixed terms	Df	denDF	F test	p-value
(Intercept)	1.00	15.60	4147.00	0.00
Control	1.00	53.30	2.23	0.14
lin(Col)	1.00	10.70	12.70	0.00
lin(Row)	1.00	57.00	0.04	0.85
Control:Check	3.00	34.20	1.72	0.18

Table D.10. Estimates of the fixed effects for Model 20 and their respective standard errors.

Effects	Estimates	Standard errors
INTERC	0.00	—
PAD1	-2.78	5.82
PAD4	-12.59	5.65
PAD9	-3.93	5.67
lin(Row)	-0.03	0.16
lin(Col)	1.40	0.39
overall test lines	0.00	—
overall checks	28.70	17.74
overall mean (μ)	93.94	3.61

The 15 best test lines from Model 20 with predicted values and respective standard errors are presented in Table D.11. Figure D.9 compares the genetic EBLUPs from Model 20 and the ones from the traditional non-spatial analysis, that contain fixed checks varieties and random direct genetic, row and column effects; there is a linear trend and the correlation between them is 0.884. However, note that the group of selected test lines is not the same.

Table D.11. The 15 best test lines with predicted values and respective standard errors from Model 20.

Test lines	Predicted values	Standard errors
GS202	146.74	10.93
GS165	143.96	11.11
GS77	142.44	10.84
GS47	133.66	10.88
GS35	130.92	11.24
GS58	130.24	9.89
GS209	128.66	10.14
GS177	128.37	9.76
GS207	127.62	9.96
GS44	127.09	11.12
GS89	126.33	11.24
GS65	125.34	11.16
GS107	124.99	11.27
GS120	124.09	11.03
GS21	122.76	11.05

Figure D.9. Dispersion plot of the genetic EBLUPs values of the Model 20 and the traditional non-spatial analysis model for this design (without spatial dependence) for Goiás Local 521. The cut-offs for the 15 best test lines (7 % upper) in each model are indicated by the dotted line.

In Local 533 several models were tested and they are presented in Table D.12. In the beginning, it was possible observe the genetic effects as a special case nominated reduced rank. However the neighbour genetic component is not significant using REML ratio test. Hence, some new models were developed with different local and global effects. Using the REML ratio test and residual plots the Model 15 is chosen that best explains the data.

Table D.12. Summary of the fitted models to Local 533 of the Goiás experiment with restricted log-likelihood (log-lik.) and the p -value of the REML ratio test. All the models include a random direct genetic effects for the test lines (G) and fixed check varieties effects. The other effects are represented with the letters: H (harvest); Co (cone); S (sowing); C (column); R (row); N (neighbour genetic); spl(.) (spline in some direction indicated in parentheses) and lin(.) (linear trend in some direction indicated in parentheses). The structures can be: RR(.) - reduced rank; AR(.) - first-order autoregressive; sar(2) - constrained autoregressive; Band(.) - banded correlation, the number indicates the order of correlation; and Id - identity.

Model	Effects				Log-lik.	Test	p -value
	Fixed	Global/extraneous	Random	Genetic	Local (column $\times$ row)		
1		R		RR(G:N)	AR(1) $\times$ AR(1)	-931.95	
2		R		G	AR(1) $\times$ AR(1)	-932.15	M2 vs M1
3		R		G	AR(1) $\times$ Band(3)	not converged - singularity	0.52
4		R		G	AR(1) $\times$ Band(2)	-933.52	
5		R		G	AR(1) $\times$ AR(2)	-932.00	M5 vs M2
6		R		G	AR(1) $\times$ sar(2)	-931.98	0.58
7		R		G	Id $\times$ AR(1)	-932.37	M2 vs M7
8	Co	R		G	Id $\times$ AR(1)	-930.79	not significant
9	H	R		G	Id $\times$ AR(1)	-930.75	not significant
10	S	R		G	Id $\times$ AR(1)	-930.99	not significant
11	lin(C)+lin(R)	spl(C)+spl(R)+R		G	Id $\times$ AR(1)	-929.94	
12	lin(C)+lin(R)	spl(C)+spl(R)		G	Id $\times$ AR(1)	-929.94	M11 vs M12
13	lin(C)+lin(R)	spl(R)		G	Id $\times$ AR(1)	-930.13	M12 vs M13
14	lin(C)+lin(R)			G	Id $\times$ AR(1)	-932.41	M13 vs M14
15	lin(R)	spl(R)		G	Id $\times$ AR(1)	-929.34	0.01
16	lin(R)	spl(R)		RR(G:N)	Id $\times$ AR(1)	-929.23	M16 vs M15
							0.64

The plots of the empirical semi-variogram for the Model 15 are presented in Figure D.10. It is observed that there is no strong trend given that the first neighbours show the flat residual effects in the graphics. The REML estimates of the variance parameters are presented in Table D.13. Here it was found more direct genetic component than residual (1.35 ratio) and also negative correlation in row direction. The standard error of row splines is a bit bigger than its component, but this term was significant when assessed using the REML ratio test.

Figure D.10. Plots of the empirical semi-variogram for residuals from Model 15 for Local 533 of the Goiás experiment.

Table D.13. REML estimates of variance parameters from fitted Model 15 for the experiment in the Local 533 of the Goiás.

Variance parameters		Ratios (γ)	Estimates	Standard errors
spl(Row)		0.10	27.39	29.06
Direct genetic	$(\tilde{\sigma}_g^2)$	1.35	353.97	69.20
Residual	$(\tilde{\sigma}^2)$	1.00	262.24	52.31
Spatial	$(\tilde{\rho}_r)$	—	-0.53	0.10

Table D.14 shows the Wald F tests for fixed terms and Table D.15 the estimated fixed effects from this model. Note that there is no evidence of significant difference between the group of checks and test lines (Control), but there is among the checks (Control:Checks).

Table D.14. Wald F tests for fixed terms from Model 15. Df means degrees of freedom and DenDF is the approximate denominator degrees of freedom.

Fixed terms	Df	denDF	F test	p-value
(Intercept)	1.00	141.80	8337.00	0.00
Control	1.00	55.30	0.94	0.34
lin(Row)	1.00	196.50	10.74	0.00
Control:Check	3.00	49.00	18.25	0.00

Table D.15. Estimates of the fixed effects from Model 15 and their respective standard errors.

Effects	Estimates	Standard errors
INTERC	0.00	—
PAD1	-36.85	6.60
PAD4	-33.11	6.39
PAD9	3.77	6.40
lin(Row)	-0.65	0.19
overall test lines	0.00	—
overall checks	26.46	19.10
overall mean (μ)	142.18	2.52

The 15 best test lines from Model 15 with predicted values and respective standard errors are presented in Table D.16. Comparing the genetic EBLUPs values from Model 15 and the traditional non-spatial analysis, that contain fixed checks varieties and random direct genetic, column and row effects, in Figure D.11, there is a linear trend and the correlation between them is 0.941. However, note that the group of selected test lines is not the same.

Table D.16. The 15 best test lines with predicted values and respective standard errors for Model 15.

Test lines	Predicted values	Standard errors
GS194	192.38	11.30
GS189	190.56	11.06
GS211	179.83	11.27
GS40	171.64	11.65
GS173	171.59	11.25
GS213	170.18	11.39
GS210	167.46	11.68
GS64	164.82	10.95
GS13	161.92	11.41
GS198	159.57	11.20
GS157	158.09	12.26
GS176	157.99	11.59
GS132	156.76	11.58
GS34	156.00	11.29
GS108	155.22	10.80

Figure D.11. Dispersion plot of the genetic EBLUPs values of the Model 20 and the traditional non-spatial analysis model for this design (without spatial dependence) for Goiás Local 533. The cut-offs for the 15 best test lines (7 % upper) in each model are indicated by the dotted line.

D.3 Joint analysis

Based on the separate analysis for each Local, it was observed that the genetic EBLUPs values are very different. Figure D.12 presents the relation of these values between pairs of Locals; the correlation is close to zero. This means that there is a Local effect on the test lines and the correlation among genetic effects should be very small.

Figure D.12. Dispersion plots for the genetic EBLUPs values from model selected for each Local in Goiás. Panels the relationship between pairs of Locals (a) 3 and 521; (b) 3 and 533; and (c) 521 and 533. The cut-offs for the 15 best test lines in each Local are indicated by the dotted line.

Hence, based on the results of the individual analyses, Local 521 is the ones that exhibits more genetic correlation with the other Locals. However, we believe that the chosen model can be biased to estimate the neighbour genetic effects. The residual component is much smaller at Local 521, almost 3 times greater than that found for the other Locals. We considered that some components as residual and neighbour genetic can be wrong estimate (underestimate and overestimate) and because of this Local 521 will not be include in the joint analysis.

Locals 3 and 533 were joint analyzed in order to investigate and to compare the magnitudes of genetic variance, assessing if there was genetic covariance between Locals. To do this the models shown in Table D.17 were tested and the results of the REML tests are also given in this Table.

Table D.17. Summary of the fitted models to joint analysis of the Goiás experiments carried out in the Locals 3 (1) and 533 (3), with REML test (log-lik.) and the p -value of the REML ratio test. All the models have the same effects as modeled in the individual analysis plus the fixed Local effects. Here σ_{gi}^2 is the genetic variance at the i th Local, $\sigma_{gii'}$ is the genetic covariance between the i th and i' th Locals which have been hypothesized to be equal for two or more pairs of Locals.

Model	Genetic effects	Log-lik.	Test	p-value
1	$\begin{bmatrix} \sigma_{g1}^2 & 0 \\ 0 & \sigma_{g3}^2 \end{bmatrix}$	-1852.47		
2	$\begin{bmatrix} \sigma_{g1}^2 & \sigma_{g13} \\ \sigma_{g13} & \sigma_{g3}^2 \end{bmatrix}$	-1847.32	M2 vs M1	0.00
3	$\begin{bmatrix} \sigma_{g2}^2 & \rho\sigma_g^2 \\ \rho\sigma_g^2 & \sigma_g^2 \end{bmatrix}$	-1847.86	M2 vs M3	0.15

Model 3 was chosen as the best to explain the data. It has the same environmental parameters, plus Local fixed effect and correlation between the direct genetic effects in Locals 3 and 533. Table D.18 showed the REML estimates of the variance parameters and their respective standard error from fitted Model 3.

Table D.18. REML estimates of variance parameters for fitted Model 3 to the joint analysis of the Local 3 (1) and 533 (3) of the experiment carried out in Goiás.

Variance parameters		Estimates	Standard errors
Correlation between genetic effects	$(\tilde{\rho}_g)$	0.39	0.12
direct genetic	$(\tilde{\sigma}_g^2)$	299.81	48.73
Local 3:Col	$(\tilde{\sigma}_{c1}^2)$	35.86	20.39
Local 3:Spline(Row)	—	13.50	15.41
Local 533:Spline(Row)	—	19.59	22.82
Local 3:Residual	$(\tilde{\sigma}_1^2)$	292.25	50.32
Spatial:column direction	$(\tilde{\rho}_{c1})$	-0.29	0.11
Spatial:row direction	$(\tilde{\rho}_{r11})$	-0.57	0.10
Spatial:row direction	$(\tilde{\rho}_{r12})$	0.18	0.09
Local 533:Residual	$(\tilde{\sigma}_3^2)$	284.01	51.23
Spatial:row direction	$(\tilde{\rho}_{r31})$	-0.44	0.09

The Wald F tests for fixed terms in Model 3 are presented in Table D.19. Note that there is not a significant difference between the Locals. The predicted values with their standard error of the 15 best test lines of the Model 3 for each local are in Table D.20. It can be observed that in the joint analysis the precision in the selection of the test lines is similar than the individual analysis. For some test lines, the predicted values are a bit higher and standard error are smaller, when compared with individual analysis (Tables D.6 and D.16).

Table D.19. Wald F tests for the fixed terms in Model 3. Df means degrees of freedom and DenDF is the approximate denominator degrees of freedom.

Fixed terms	Df	denDF	F test	p-value
(Intercept)	1.00	23.80	10680.00	0.00
Local	1.00	1.10	16.51	0.14
Control	1.00	0.00	1.45	0.14
Control:Check	3.00	47.80	12.67	0.00
Local:Control	1.00	0.00	0.01	0.00
Local 3:harvest	1.00	0.00	8.75	0.00
Local:lin(Row)	2.00	—	6.15	—
Local:Control:Check	3.00	76.60	4.54	0.01

Table D.20. The 15 best test lines with predicted values and respective standard error from Model 3 for each Local.

Local 3			Local 533		
Test lines	Pred.value	Stand.error	Test lines	Pred.value	Stand.error
GS51	192.03	10.96	GS194	186.42	11.18
GS204	168.00	11.09	GS189	178.99	11.18
GS87	163.80	10.73	GS211	169.92	11.24
GS132	161.88	11.15	GS40	167.16	11.39
GS2	160.26	11.17	GS210	166.21	11.45
GS99	159.52	11.19	GS213	165.57	11.32
GS210	159.43	11.23	GS173	165.04	11.25
GS151	158.11	11.61	GS64	162.94	11.06
GS53	157.65	10.50	GS13	158.93	11.23
GS79	156.90	11.00	GS132	158.61	11.37
GS162	155.42	10.99	GS198	158.16	11.24
GS198	154.37	11.15	GS176	156.29	11.41
GS141	153.66	11.23	GS51	155.84	11.38
GS102	153.07	11.39	GS87	155.49	11.36
GS48	153.00	11.13	GS157	154.31	11.87

D.4 Conclusions

The individual analyses resulted in different models for Locals. The main difference was the selected models: for Local 3 and 533, there were residual competition effects in the row direction, while for Local 521 there was no significant residual correlation, but competition at the genetic level. The same test lines were assessed in different sites although the ratios between direct genetic and residual components were different: they were 0.84, 3.44 and 1.35 for Local 3, 521 and 533, respectively. It was expected that the models with genetic competition effect would have lower residual and high genetic components. However, this was the only experiment for which there was a significant competition effect and it is extraordinary that the direct genetic ratio is so much larger for this Local when compared to other Locals, given that the same test lines were used as all Locals. Consequently a simulation study will be conducted for this case.

The selected model for Local 3 included fixed effect for harvest and random effects for a row spline; column effects and direct genetic effects; residual correlation was modeled with AR(1) and Band(2) in column and row direction, respectively. For Local 521, it was included term for a column linear trend, a row spline, random column effects and unstructured genetics effects (direct and neighbour variances and their covariance). For Local 533, the model included a row spline term and direct genetic effects with residual correlation only in row direction modeled with AR(1) structure.

In the joint analysis, for Locals 3 and 533, the selected model indicates a common genetic component for Locals ($\sigma_g^2 = 299.81$) and a genetic correlation term ($\rho_g = 0.39$). The number of common test lines considering the 15 best lines is a little bigger; 5 of the 15 best predicted values for the test lines were common for the Locals. That is, the joint analysis did not result in greater accuracy of selection, as is to be expected given the low genetic correlation between Locals.

E RIBEIRÃO - GROUP OF EXPERIMENTS

There were three sites, nominated: Local 20, 72 and 140. Each site had different rectangular array and number of filler plot. For Local 20, there were 19 columns by 28 rows; Local 72, 21 columns by 25 rows, and, Local 140, 14 columns by 38 rows. All assessed Locals had the same 429 test lines (R) with 4 commercial varieties nominated “PAD1”, “PAD2”, “PAD3” and “TI” or “INTER”. The experiments was planted in 2014 with the yield in TCH obtained in 2015, in a unique cut. The results were organized in subsections where E.1 presents the descriptive analysis of the sites, E.2 shows the individual analyses for each site and E.3 gives the joint analysis. The concluding remarks are in subsection E.4.

E.1 Descriptive analysis

In the descriptive analysis presented in Table E.1, it is possible to observe that the commercial variety “PAD2” is not so stable, the average yield was close to the other groups, but it has the highest variance in all sites. For Local 72 the interspersed group presents also greater TCH variance, ending up with the highest TCH value for that Local. Considering the mean values, it seems there is no great difference among sites, however Local 140 is the one with had smaller yield average. Figure E.1 shows that some plots can be considered outliers. Then, it can expected that some environmental effects have influenced the productivity and the residuals from the selected models should be examined. On the other hand, the test lines outliers are expected, mainly because our interest is selection of the best sugarcane lines.

Table E.1. Descriptive analysis of the groups of clones carried out in Ribeirão.

	Group of clones	Minimum	Maximum	Mean	Variance
Local 20	Interspersed (TI or INTER)	100.00	158.30	128.63	155.32
	PAD1	90.50	127.40	112.38	172.07
	PAD2	66.70	133.30	99.76	527.39
	PAD3	107.10	145.20	124.28	170.20
	Test lines (R)	66.70	173.80	111.96	351.49
Local 72	Interspersed (TI or INTER)	71.40	210.70	121.28	577.45
	PAD1	103.60	152.40	127.60	218.82
	PAD2	81.00	175.00	110.95	657.55
	PAD3	95.20	147.60	131.29	187.89
	Test lines (R)	58.30	246.40	125.72	508.83
Local 140	Interspersed (TI or INTER)	78.60	138.10	109.62	129.95
	PAD1	72.60	121.40	102.86	316.88
	PAD2	63.10	136.90	85.48	466.35
	PAD3	95.20	135.70	108.56	183.43
	Test lines (R)	57.10	167.90	106.67	431.66

Figures E.2, E.3 and E.4 show the heat maps relating to TCH in each Local. The names “PAD1”, “PAD2”, “PAD3” and “TI” are the check plots; the numbers refer to the test lines and ‘ ‘filler’ ’ indicate plots that were ignored in the experiments. Figure E.3 reports that line 193 has a higher yield than the rest and it is not clear that there is some trend in this area.

Figure E.2. Heat map relating to TCH for Local 20 of Ribeirão.

Figure E.3. Heat map relating to TCH for Local 72 of Ribeirão.

Figure E.4. Heat map relating to TCH for Local 140 of Ribeirão.

E.2 Individual analysis

For Local 20, the algorithms for the models with neighbour genetic effect did not converge, there were singularity problems with the Average Information Matrix in fitting the R-TIM model. Table E.2 presents the sequence of models fitted to find one that describes the global and local effects for Local 20. Ultimately, Model 14 was chosen and had: linear trend in row and column factor; random row and direct genetic effects; and residual with first-order banded correlation in the row direction. The row and

column faces of the empirical variogram, presented in Figure E.5, shows the residuals for the Models 2 and 14.

Table E.2. Summary of the fitted models for Local 20 with restricted log-likelihood (log-lik.) and the p -value for the REML ratio test. All the models include a random direct genetic effects for the test lines (G) and fixed checks varieties effects. The other effects are represented with the symbols: H (harvest); Co (cone); S (sowing); C (column); R (row); N (neighbour genetic); spl(.) (spline in some direction indicated in parentheses) and lin(.) (linear trend in some direction indicated in parentheses). The structures can be: US(.) - unstructured; AR(1) - autoregressive; Band(.) - banded correlation, the number indicates the order of correlation; and Id - identity.

Model	Effects					Log-lik.	Test	p -value
	Fixed	Global/extraneous	Random	Genetic	Local (column $\times$ row)			
1		C + R		US(G:N)	AR(1) $\times$ AR(1)	not converged - singularity		
2		C + R		G	AR(1) $\times$ Band(3)	-1719.86		
3		C + R		G	AR(1) $\times$ Band(2)	-1719.92	M2 vs M3	0.72
4		C + R		G	AR(1) $\times$ Band(1)	-1719.95	M3 vs M4	0.79
5		C + R		G	AR(1) $\times$ Id	-1732.84	M4 vs M5	0.00
6		C + R		G	Id $\times$ Band(1)	-1720.04	M4 vs M6	0.66
7		C + R		G	Id $\times$ AR(1)	-1721.36		
8		C + R		G	Id $\times$ AR(2)	-1719.80	M8 vs M7	0.08
9	H	C + R		G	Id $\times$ Band(1)	-1719.09	not significant	
10	Co	C + R		G	Id $\times$ Band(1)	-1718.94	not significant	
11	S	C + R		G	Id $\times$ Band(1)	-1719.19	not significant	
12	lin(C)+lin(R)	spl(C)+C+spl(R)+R		G	Id $\times$ Band(1)	-1718.31		
13	lin(C)+lin(R)	C+R		G	Id $\times$ Band(1)	-1718.31	M12 vs M13	1.00
14	lin(C)+lin(R)	R		G	Id $\times$ Band(1)	-1718.85	M13 vs M14	0.15
15	lin(C)+lin(R)			G	Id $\times$ Band(1)	-1722.00	M14 vs M15	0.00
16	lin(C)+lin(R)	R		US(G:N)	Id $\times$ Band(1)	not converged - singularity

Figure E.5. Plots of the column and row faces of the empirical variogram for the residuals for the Local 20 of the Ribeirão experiment for models 2 (panels (a) and (c)), and 14 (panels (b) and (d)). The panels (a) and (b) are for the column direction and the others for row.

The REML estimates of variance parameters from fitted Model 14 are presented in Table E.3. Note that the genetic component is smaller than the residual (this ratio is 0.85) and the row correlation is negative, it indicates competition at the residual level.

Table E.3. REML estimates of variance parameters for fitted Model 14 for the experiment in Local 20 from Ribeirão.

Variance parameters		Ratios (γ)	Estimates	Standard errors
Row	(σ_r^2)	0.07	13.05	7.85
Direct genetic	(σ_g^2)	0.85	151.18	32.02
Residual	(σ^2)	1.00	178.00	26.34
Spatial	$(\tilde{\rho}_r)$	—	-0.35	0.06

Table E.4 gives the Wald F tests for the fixed terms in Model 14 and the coefficients of the fixed effects with their respective standard errors are in Table F.5. With the test, there is no evidence of significant difference between the group of test lines and checks (Control), but there is within the checks (Control:Check).

Table E.4. Wald F test for fixed terms for Model 14. Df means degrees of freedom and DenDF is the approximate denominator degrees of freedom.

Fixed terms	Df	denDF	F test	p-value
(Intercept)	1.00	505.00	12940.00	0.00
Control	1.00	505.00	0.38	0.54
lin(COL)	1.00	505.00	7.53	0.01
lin(ROW)	1.00	505.00	6.58	0.01
Control:Check	3.00	505.00	21.76	0.00

Table E.5. Estimates of the fixed effects for Model 14 and their respective standard errors.

Effects	Estimates	Standard errors
TI	18.74	4.46
PAD1	0.00	—
PAD2	-10.14	6.56
PAD3	14.25	6.50
lin(ROW)	-0.31	0.11
lin(COL)	0.30	0.10
overall Checks	0.00	—
overall Test lines	2.16	12.99
overall mean (μ)	111.39	13.11

The 30 best test lines with predicted values and respective standard errors, using Model 14, are presented in Table E.6. The correlation between the EBLUPs from selected model 14 and the traditional model which has fixed checks effects and random direct genetic, row and column effects, is high, 0.953, but as it can be possible to observe in Figure E.8 that the group of select test lines is different in each model.

Table E.6. The 30 best test lines with predicted values and respective standard errors for Model 14.

Test lines	Pred. values	Stand. errors	Test lines	Pred. values	Stand. errors
350	138.85	8.72	8	127.58	8.84
49	137.65	8.77	314	127.54	8.84
358	133.67	8.72	43	127.14	8.87
61	133.49	8.85	97	126.84	8.84
282	133.36	8.85	151	126.62	9.02
375	131.99	8.85	69	126.41	8.70
215	131.89	8.84	291	126.28	8.84
323	130.92	8.85	381	125.60	8.85
220	130.10	8.84	166	125.60	8.84
364	129.89	9.02	356	125.55	8.86
259	129.15	8.70	380	125.43	8.84
384	128.91	8.70	293	125.32	8.86
411	128.90	8.72	130	124.88	8.85
39	128.17	8.87	412	124.87	8.85
429	128.03	8.73	117	124.77	8.86

Figure E.6. Dispersion plot of the genetic EBLUPs values of the Model 14 and the traditional non-spatial analysis model for this design (without spatial dependence) for Ribeirão Local 20. The cut-offs for the 30 best test lines (7 % upper) in each model are indicated by the dotted line.

For Local 72, the initial model fitted included the variance parameters shown in Table E.7: random row, column and genetic effects; residual effects modelled using AR(1) in both the row and

column directions. It is clear that the estimated value of the residual component is much larger than those for the other components and, in particular, that there is very little genetic variation. This was the case irrespective of whether or not observations considered to be outliers were removed. Hence, selection of test lines is not viable for this Local and it will not be included in the joint analysis.

Table E.7. Estimates of the variance parameters and standard error for the model with first-order autoregressive structure in row and column direction, and direct genetic, row and column random effects. The log-likelihood of the model is -1851.92.

Variance parameters		Estimates	Standard errors
Row	σ_r^2	19.92	11.47
Column	σ_c^2	9.81	7.97
Direct genetic	σ_g^2	14.27	79.92
Residual	σ^2	467.41	72.13
Spatial:column	ρ_c^2	-0.11	0.05
Spatial:row	ρ_r^2	-0.13	0.05

For Local 140, as in Local 20, the same problem with the Average Information Matrix happened with the models R-TIM and third-order banded correlation. Table E.8 shows a sequence of fitted models. Using REML ratio test and graphics of residuals diagnostic, Model 14 was considered the best fitted with column linear trend; row spline; direct genetic effect and residual AR(1) and Band(1) in column and row direction, respectively. See in Figure E.7, the empirical semi-variogram for this model.

Table E.8. Summary of the models fitted for Local 140 with restricted log-likelihood (log-lik.) and the p -value for the REML ratio test. All the models include a random direct genetic effects for the test lines (G) and fixed checks varieties effects. The other effects are represented with the symbols: H (harvest); Co (cone); S (sowing); C (column); R (row); N (neighbour genetic); spl(.) (spline in some direction indicated in parentheses) and lin(.) (linear trend in some direction indicated in parentheses). The structures can be: US(.) - unstructured; AR(1) - first-order autoregressive; Band(.) - banded correlation, the number indicates the order of correlation; and Id - identity.

Model	Effects				Log-lik.	Test	p -value
	Fixed	Global/extraneous	Random	Genetic	Local (column $\times$ row)		
1		R + C		US(G:N)	AR(1) $\times$ AR(1)	not converged - singularity	
2		R + C		G	AR(1) $\times$ Band(3)	not converged - singularity	
3		R + C		G	AR(1) $\times$ Band(2)	-1761.54	
4		R + C		G	AR(1) $\times$ Band(1)	-1761.55	M3 vs M4 0.87
5		R + C		G	AR(1) $\times$ Id	-1771.50	M4 vs M5 0.00
6		R + C		G	Id $\times$ Band(1)	-1764.12	M4 vs M6 0.02
7		R + C		G	AR(1) $\times$ AR(1)	-1761.87	
8	H	R + C		G	AR(1) $\times$ Band(1)	-1759.59	not significant
9	S	R + C		G	AR(1) $\times$ Band(1)	-1760.44	not significant
10	Co	R + C		G	AR(1) $\times$ Band(1)	-1760.29	not significant
11	lin(C)+lin(R)	spl(C)+C+spl(R)+R		G	AR(1) $\times$ Band(1)	-1754.10	
12	lin(C)+lin(R)	C+spl(R)+R		G	AR(1) $\times$ Band(1)	-1754.10	M11 vs M12 1.00
13	lin(C)+lin(R)	spl(R)+R		G	AR(1) $\times$ Band(1)	-1754.11	M12 vs M13 0.44
14	lin(C)+lin(R)	spl(R)		G	AR(1) $\times$ Band(1)	-1754.29	M13 vs M14 0.27
15	lin(C)+lin(R)			G	AR(1) $\times$ Band(1)	-1769.27	M14 vs M15 0.00
16	lin(C)+lin(R)	spl(R)		US(G:N)	AR(1) $\times$ Band(1)	not converged - singularity

Figure E.7. Plot of the empirical semi-variogram for the residuals for Local 140 of Ribeirão experiment referring to Model 14.

REML estimates of the variance parameters with γ and standard errors from fitted Model 14 are presented in the Table E.9. The direct genetic component also is smaller than the residual, with ratio equal 0.88. There is competition at the residual level, because the correlation in row direction is negative, -0.40.

Table E.9. REML estimates of variance parameters for fitted Model 14 to the experiment in the Local 140 of the Ribeirão.

Variance parameters		Ratios (γ)	Estimates	Standard errors
spl(Row)		0.11	23.34	23.02
Direct genetic	$(\tilde{\sigma}_g^2)$	0.88	180.48	32.90
Residual	$(\tilde{\sigma}^2)$	1.00	205.52	28.66
Spatial	$(\tilde{\rho}_c)$	—	0.23	0.08
Spatial	$(\tilde{\rho}_r)$	—	-0.40	0.06

The Wald F test for fixed terms of the Model 14 are presented in Table E.10. There is no evidence of significant difference between the group of checks and test lines (Control), but there is within the checks (Control:Checks). The estimates of these fixed effects are in Table E.11 with their respective standard error.

Table E.10. Wald F test of fixed terms from Model 14. Df means degrees of freedom and DenDF is the approximate denominator degrees of freedom.

Fixed terms	Df	denDF	F test	p-value
(Intercept)	1.00	506.00	18960.00	0.00
Control	1.00	506.00	0.08	0.77
lin(COL)	1.00	506.00	12.86	0.00
lin(ROW)	1.00	506.00	2.99	0.08
Control:Check	3.00	506.00	12.59	0.00

Table E.11. Estimates of the fixed effects for Model 14 and their respective standard errors.

Effects	Estimates	Standard errors
TI	12.33	4.50
PAD1	0.00	—
PAD2	-9.96	6.78
PAD3	7.56	6.44
lin(ROW)	-0.09	0.06
lin(COL)	0.55	0.15
overall Checks	0.00	—
overall Test lines	9.21	14.09
overall mean (μ)	95.16	14.17

The 30 best test lines with their respective predicted values and standard errors are presented in Table E.12. The correlation between the EBLUPs from Model 14 and the traditional is high, 0.932, but also the group of the best test lines is different for each model.

Table E.12. The 30 best test lines with predicted values (pred. value) and respective standard errors (stand. error) for Model 14.

Test lines	Pred. values	Stand. errors	Test lines	Pred. values	Stand. errors
152	137.26	9.32	49	131.26	9.40
43	136.13	9.40	266	131.07	9.25
22	135.80	9.05	315	130.87	9.19
223	134.82	9.11	298	130.50	9.39
255	134.44	9.03	426	129.76	9.31
8	134.38	9.35	337	129.38	9.39
188	133.81	9.21	4	129.21	9.33
39	133.58	9.12	219	129.09	9.39
416	132.90	9.34	325	129.01	9.40
348	132.76	9.21	279	128.68	9.40
245	132.40	9.45	274	128.37	9.18
310	132.15	9.11	24	128.30	9.11
314	131.95	9.67	186	127.82	8.94
410	131.82	9.43	15	127.75	9.29
352	131.55	9.42	36	127.18	9.39

Figure E.8. Dispersion plot of the genetic EBLUPs values of the Model 14 and the traditional non-spatial analysis model for this design (without spatial dependence) for Ribeirão Local 140. The cut-offs for the 30 best test lines (7 % upper) in each model are indicated by the dotted line.

E.3 Joint analysis

Based on the separate analysis for each Local, it was observed that the genetic EBLUPs are very different. Figure E.9 presents the relationship between the values for Locals 20 and 140 in the form of a scatterplot; the correlation is 0.401. This means that there is a substantial Local effect on the test lines and only limited correlation between the genetic effects is expected.

Figure E.9. Dispersion plot of the genetic EBLUPs from model selected for Ribeirão Locals 20 and 140. The cut-offs for the 30 best test lines in each Local are indicated by the dotted line.

We started the joint analysis with the models chosen in the individual analysis, adding only fixed effects for Local. Then, three models for genetic components and covariance (see Table E.13) were fitted and the REML ratio tests used to select the model to use for the data.

Table E.13. Summary of the models fitted to joint analysis for the Ribeirão experiments carried out for the Local 20 (1) and 140 (3), with REML log (log-lik.) and the p -value of the REML ratio test. All the models have the same fixed; local and global effects modeled in the individual analysis plus the local fixed effects.

Model	Genetic Effects	Constraint	Log-lik.	Test	p-value
1	$\begin{bmatrix} \sigma_{g1}^2 & 0 \\ 0 & \sigma_{g3}^2 \end{bmatrix}$		-3473.15		
2	$\begin{bmatrix} \sigma_g^2 & 0 \\ 0 & \sigma_g^2 \end{bmatrix}$		-3473.33	M2 vs M1	0.26
3	$\begin{bmatrix} \sigma_g^2 & \sigma_{g13} \\ \sigma_{g12} & \sigma_g^2 \end{bmatrix}$	$\sigma_{g13} = \rho_g \sigma_g^2$	-3434.87	M2 vs M3	0.00

Using the REML ratio test, the Model 3 was chosen as the best to explain the effects that happened in these experiments. Table E.14 shows the REML estimates of the variance parameters and their respective standard errors from fitted Model 3.

Table E.14. REML estimates of variance parameters for fitted Model 3 for the joint analysis of Local 20 (1) and 140 (3) of the experiment carried out in Ribeirão.

Variance parameters		Estimates	Standard errors
Correlation between genetic effects	$(\tilde{\rho}_g)$	0.82	0.12
Direct genetic	$(\tilde{\sigma}_g^2)$	162.09	24.43
Local 20:Row	$(\tilde{\sigma}_{r1}^2)$	16.21	8.21
Local 140:Spline(Row)		9.48	10.36
Local 20:Residual	$(\tilde{\sigma}_1^2)$	166.23	21.43
Spatial:row direction	$(\tilde{\rho}_{r1})$	-0.31	0.06
Local 140:Residual	$(\tilde{\sigma}_3^2)$	213.31	25.23
Spatial:column direction	$(\tilde{\rho}_{c3})$	0.21	0.06
Spatial:row direction	$(\tilde{\rho}_{r3})$	-0.32	0.05

The Wald F test for fixed terms of the Model 3 are presented in Table E.15. Note that there is a significant difference between the Locals, and the 30 best test lines predicted values with their standard errors from Model 3 are in Table E.16 for each Local. It can be observed that in the joint analysis the precision in the selection of the test lines is better, given that the standard errors are smaller than compared with individual analysis (Tables E.6 and E.12).

Table E.15. Wald F test for fixed terms of the Model 3. Df means degrees of freedom and DenDF is the approximate denominator degrees of freedom.

Fixed terms	Df	denDF	F test	p-value
(Intercept)	1.00	1011.00	21230.00	0.00
Local	1.00	1011.00	25.94	0.00
Control	1.00	1011.00	0.04	0.84
Control:Check	3.00	1011.00	32.38	0.00
Local:Control	1.00	1011.00	2.09	0.15
Local:lin(COL)	2.00	1011.00	9.61	0.00
Local:lin(ROW)	2.00	1011.00	6.22	0.00
Local:Control:Check	3.00	1011.00	0.75	0.52

Table E.16. Predicted values and standard errors of the 30 best test lines of the Model 3 for each local.

Local 20			Local 140		
Test lines	Predicted values	Standard errors	Test lines	Predicted values	Standard errors
49	140.56	7.99	49	138.18	8.20
350	138.28	7.99	152	136.82	8.21
282	135.17	8.07	8	136.55	8.17
8	134.82	8.03	43	136.28	8.24
39	134.06	7.99	314	135.48	8.36
314	134.01	8.08	39	135.40	8.15
152	133.43	8.04	350	134.65	8.18
43	133.21	8.05	282	133.18	8.36
61	131.23	7.99	245	132.68	8.26
245	130.79	7.98	266	131.91	8.18
215	130.69	8.05	352	130.27	8.28
384	130.22	7.97	384	130.11	8.16
291	129.85	8.02	325	130.10	8.21
220	129.80	8.04	348	130.09	8.20
411	129.54	7.96	416	129.95	8.16
259	129.38	7.96	291	129.57	8.14
151	129.27	8.16	410	129.43	8.23
325	129.05	8.00	426	129.41	8.17
266	128.39	8.00	223	129.18	8.13
130	127.81	8.04	151	128.98	8.26
97	127.68	8.05	315	128.93	8.22
364	127.54	8.15	310	128.70	8.17
410	127.16	8.05	22	128.69	8.07
426	127.16	8.01	411	128.46	8.15
382	126.93	8.06	217	128.39	8.23
429	126.81	8.02	4	128.30	8.24
352	126.58	8.10	215	128.04	8.23
358	126.41	8.05	382	127.89	8.30
375	125.85	8.06	219	127.81	8.21
380	125.84	8.12	220	127.78	8.22

E.4 Conclusions

The individual analyzes reflect the peculiarities of each local. For this region, convergence problems were experienced when fitting models with competition at the genetic level. For Local 72, it was impossible to select test lines, because all the models indicated a high residual component compared with the genetic component which was not significant. Both Locals 20 and 140 were modeled with trend effect in row and column, and similar correlation in row direction using first-order banded structure (-0.35 and -0.40), indicating competition at the residual level. The difference between the Locals was that Local 20 had a significant random row effect, while Local 140, there is a significant row spline term and AR(1) in column direction. The direct genetic component for both is similar when compared to the ratio between this parameter and the residual, around 0.86, and the correlation between the predicted values from selected models is 0.401. Hence, better accuracy in the selection of the test lines is achieved when it is based on the joint analysis. Lower standard errors values are found. Model 3 was chosen to jointly model Locals 20 and 140, indicates a common genetic component (162.09) for both locals and a high correlation (0.82) between the genetic effects.

F PIRACICABA - GROUP OF EXPERIMENTS

There were three sites, nominated: Local 54, 58 and 76. In the first two sites the experiments were carried out in a rectangular array of 14 columns by 37 rows while in the last there were 15 column by 35 rows. All assessed the same 422 test lines (R) with 4 commercial varieties nominated “PAD1”, “PAD5”, “PAD6” and “Interc”. Planted in 2014, the yield in TCH was obtained in 2015, in a unique cut. The results were organized in subsections where F.1 presents the descriptive analysis of the Locals. F.2 shows the individual analysis for each Local and F.3 the joint analysis. The concluding remarks are in subsection F.4.

F.1 Descriptive analysis

Table F.1 shows a summary of the TCH of the groups of clones by each Local. It is possible to observe that the greatest yield for each group is found in Local 76, but this also had the greatest variance among sugarcane groups, except for “PAD5”. Local 54 has the smallest average value for all groups and the biggest variance value referring to “PAD5”.

Table F.1. Descriptive analysis of the groups of clones carried out in Piracicaba.

	Groups of clones	Minimum	Maximum	Mean	Variance
Local 54	Interspersed	80.40	145.80	110.78	266.25
	PAD1	86.40	136.90	112.42	280.11
	PAD5	59.50	144.30	99.03	788.48
	PAD6	78.80	139.90	110.45	347.36
	Test Lines (R)	44.60	194.90	104.08	475.51
Local 58	Interspersed	92.90	192.90	156.01	304.22
	PAD1	97.60	135.70	117.60	225.21
	PAD5	79.80	140.50	107.29	265.15
	PAD6	107.10	176.20	145.75	537.25
	Test Lines (R)	76.20	209.50	124.91	469.97
Local 76	Interspersed	110.70	225.00	171.61	568.03
	PAD1	108.30	189.30	162.83	595.27
	PAD5	83.30	154.80	105.42	459.49
	PAD6	134.50	214.30	166.53	608.44
	Test Lines (R)	69.00	222.60	148.81	655.01

The boxplot in Figure F.1 illustrates the outliers values for each group of lines given each Local.

Figure F.1. Boxplot of TCH for each group of lines for Locals of Piracicaba.

For all the Locals there are test lines considered outliers. However, as these did not have repetition and external information, the values were accepted. Also some check plots were considered outliers, in Local 58 and 76. The yield of the interspersed plot in Local 58 was found in Local 54. Then, this value must be real. The same is thought for “PAD5” plots Local 58 and “PAD1” Local 76. The “PAD5” plot in Local 76 did not occur in other Locals, hence it will be observed if it represents a possible problem in the analysis, verifying the residuals of the models.

See in Figures F.2, F.3 and F.4, the heat maps relating to TCH in each Local. The names “PAD1”, “PAD5”, “PAD6” and “INTERC” are the check plots, the test lines are labeled alpha-numerically (PI plus number) and “x” represents the empty plots.

Figure F.2. Heat map relating to TCH for Local 54 of Piracicaba.

Figure F.3. Heat map relating to TCH for Local 58 of Piracicaba.

Figure F.4. Heat map relating to TCH for Local 76 of Piracicaba.

F.2 Individual analysis

For Local 54, it is shown, in Table F.2, a summary of the sequence of fitted models. This began with a model that included random effect for direct, neighbour and covariance (G:N) genetic, rows and columns; fixed effect for the check varieties; and spatial dependence, fitted with first-order autoregressive structure in the row and column direction. However, model with competition at the genetic level were not significant. Hence, the local and global effects were assessed without genetic competition.

Table F.2. Summary of the fitted models to Local 54 of Piracicaba with restricted log-likelihood (log-lik.) and the p -value of the REML ratio test. All models include a random direct genetic effects for the test lines (G) and fixed checks varieties effects. The other effects are represented with the symbols: H (harvest); Co (cone); S (sowing); C (column); R (row); N (neighbour genetic); spl(.) (spline in some direction indicated in parentheses) and lin(.) (linear trend in some direction indicated in parentheses). The structures can be: US(.) - unstructured; RR(.) - reduced rank; AR(1) - first-order autoregressive; Band(.) - banded correlation, the number indicates the order of correlation; and Id - identity.

Model	Effects				Log-lik.	Test	p -value
	Fixed	Global/extraneous	Random	Genetic	Local (column $\times$ row)		
1		C + R		US(G:N)	AR(1) $\times$ AR(1)		
2		C + R		RR(G:N)	AR(1) $\times$ AR(1)		
3		C + R		G	AR(1) $\times$ AR(1)	M2 vs M3	0.36
4		C + R		G	AR(1) $\times$ Band(3)		
5		C + R		G	AR(1) $\times$ Band(2)	M4 vs M5	0.81
6		C + R		G	AR(1) $\times$ Band(1)	M5 vs M6	0.97
7		C + R		G	AR(1) $\times$ Id	M6 vs M7	0.00
8		C + R		G	Id $\times$ Band(1)	M6 vs M8	0.25
9	Co	C + R		G	Id $\times$ Band(1)	not significant	
10	S	C + R		G	Id $\times$ Band(1)	not significant	
11	H	C + R		G	Id $\times$ Band(1)	not significant	
12	lin(C)+lin(R)	spl(C)+C+spl(R)+R		G	Id $\times$ Band(1)		
13	lin(C)+lin(R)	spl(C)+C+R		G	Id $\times$ Band(1)	M12 vs M13	1.00
14	lin(C)+lin(R)	spl(C)+R		G	Id $\times$ Band(1)	M13 vs M14	0.23
15	lin(C)+lin(R)	R		G	Id $\times$ Band(1)	M14 vs M15	0.24
16	lin(C)+lin(R)			G	Id $\times$ Band(1)	M15 vs M16	0.00
17	lin(R)	R		G	Id $\times$ Band(1)	not significant	
18	lin(C)	R		G	Id $\times$ Band(1)	significant	
19	lin(C)	R		US(G:N)	Id $\times$ Band(1)	M19 vs M18	0.57

Note in Table F.2 that after testing for global and local environmental effects, the neighbour genetic and covariance between the genetics effects were tested again. However, they were not significant.

Using the REML ratio test, the chosen model was Model 18, with column linear trend, fixed check varieties effects, random rows and direct genetic effects, and residual competition modeled with a first-order banded structure in the row direction. Figure F.5 has the plots of the row and column faces of the empirical variogram for the Models 8 and 18; it shows the effect of the fitted model on the residuals.

Figure F.5. Plots of the row and column faces of the empirical variogram for the residuals for the Local 54 of Piracicaba experiment for models 8 (panels (a) and (c)) and 18 (panels (b) and (d)). The panels (a) and (b) are column direction and the others are row direction.

In table F.3 can be seen the REML estimates of variance parameters of the Model 18. Note that the direct genetic component is smaller than the residual (the ratio is 0.77) and the row correlation is negative 0.27, this indicates residual competition.

Table F.3. REML estimates of variance parameters for fitted Model 18 for the experiment of Piracicaba, Local 54.

Variance parameters		Ratios (γ)	Estimates	Standard errors
Row	$(\tilde{\sigma}_r^2)$	0.25	59.97	20.99
Direct genetic	$(\tilde{\sigma}_g^2)$	0.77	182.15	47.25
Residual	$(\tilde{\sigma}^2)$	1.00	235.89	38.76
Spatial:Row	$(\tilde{\rho}_r)$	—	-0.27	0.07

The Wald F tests for fixed terms and the estimated fixed effects from selected Model 18 are presented in Table F.4 and Table F.5, respectively. In this model for this data, there are no evidences

of significant difference between the group of test lines and checks (Control) or within the checks (Control:Check).

Table F.4. Wald F test for fixed factors for Model 18. Df means degrees of freedom and DenDF is the approximate denominator degrees of freedom.

Fixed terms	Df	denDF	F test	<i>p</i> -value
(Intercept)	1	496	4568.00	0.00
Control	1	496	0.28	0.60
lin(Col)	1	496	8.03	0.00
Control:Check	3	496	1.71	0.16

Table F.5. Estimates of the fixed effects for Model 18 and their respective standard errors.

Effects	Estimates	Standard errors
INTERC	0.00	—
PAD1	-0.03	5.45
PAD5	-11.11	5.43
PAD6	3.70	5.51
lin(Col)	0.49	0.18
overall Checks	0.00	—
overall Test lines	-7.88	13.69
overall mean (μ)	108.11	13.76

The 30 best test lines with predicted values and respective standard errors from Model 18 are presented in Table F.6. Note in Figure F.6 the correlation between genetic EBLUPs from selected model and the traditional model (with random row, column and direct genetic effects) which are very strong, 0.966. However the group of test lines selected in each model are different.

Table F.6. The 30 best test lines with predicted values (pred. value) and respective standard errors (stand. error) for Model 18.

Test line	Pred. values	Stand. errors	Test lines	Pred. values	Stand. errors
PI318	136.13	10.23	PI254	119.43	10.19
PI356	129.61	10.14	PI308	119.17	10.15
PI342	124.94	10.19	PI37	119.14	10.13
PI203	123.11	10.19	PI391	119.01	10.22
PI296	123.06	10.14	PI301	118.84	10.22
PI100	122.96	10.26	PI388	118.76	10.21
PI320	122.46	10.15	PI235	118.56	10.21
PI22	121.71	10.21	PI260	118.19	10.13
PI136	121.60	10.19	PI160	118.19	10.13
PI412	121.28	10.13	PI395	118.09	10.20
PI408	121.12	10.22	PI1	117.90	10.21
PI340	120.36	10.22	PI128	117.87	10.16
PI256	119.87	10.24	PI357	117.85	10.20
PI115	119.74	10.20	PI294	117.80	10.15
PI248	119.64	10.22	PI259	117.80	10.23

Figure F.6. Dispersion plot of the genetic EBLUPs values of the Model 18 and the traditional non-spatial analysis model for this design (without spatial dependence) for Piracicaba Local 54. The cut-offs for the 30 best test lines (7 % upper) in each model are indicated by the dotted line.

For Local 58 the same process than Local 54 was conducted, but the random column component is boundary. Then, the models were analyzed without this effect and the sequence of the models is presented in Table F.7.

Table F.7. Summary of the fitted models to Local 58 of Piracicaba with restricted log-likelihood (log-lik.) and the p -value of the REML ratio test. All models have a random direct genetic effects for the test lines (G) and fixed check varieties effects. The other effects are represented with the symbols: H (harvest); Co(cone); S (sowing); C (column); R (row); N (neighbour genetic); spl(.) (spline in some direction indicated in parentheses) and lin(.) (linear trend in some direction indicated in parentheses). The structures can be: US(.) - unstructured; RR(.) - reduced rank. AR(.) - autoregressive; Band(.) - banded correlation, the number indicates the order of correlation; sar(2) - constrained autoregressive 3; and Id - identity.

Model	Effects			Log-lik.	Test	p -value
	Fixed	Global/extraneous	Random			
1		R		AR(1) $\times$ AR(1)		
2		R		AR(1) $\times$ AR(1)	M1 vs M2	0.22
3		R		AR(1) $\times$ Band(3)		
4		R		AR(1) $\times$ Band(2)	M3 vs M4	0.57
5		R		AR(1) $\times$ Band(1)	M4 vs M5	0.02
6		R		AR(1) $\times$ AR(2)		
7		R		AR(1) $\times$ sar(2)		
8		R		AR(2) $\times$ sar(2)	M8 vs M7	0.15
9		R		Band(3) $\times$ sar(2)		
10		R		Band(2) $\times$ sar(2)	M9 vs M10	0.40
11		R		Band(1) $\times$ sar(2)	M10 vs M11	0.04
12	Co	R		Band(2) $\times$ sar(2)	not significant	
13	S	R		Band(2) $\times$ sar(2)	not significant	
14	H	R		Band(2) $\times$ sar(2)	not significant	
15	lin(C)+lin(R)	spl(C)+spl(R)+R		Band(2) $\times$ sar(2)		
16	lin(C)+lin(R)	spl(R)+R		Band(2) $\times$ sar(2)	M15 vs M16	1.00
17	lin(C)+lin(R)	spl(R)		Band(2) $\times$ sar(2)	M16 vs M17	0.02
18	lin(C)+lin(R)	R		Band(2) $\times$ sar(2)	M16 vs M18	0.00
19	lin(R)	spl(R)+R		Band(2) $\times$ sar(2)		
20	lin(R)	spl(R)+R		Band(2) $\times$ sar(2)		
21	lin(R)	spl(R)+R		Band(2) $\times$ sar(2)	M21 vs M19	0.44

Using the REML test and the plot of the empirical semi-variogram for the residual, in Figure F.7, Model 19 is chosen as the best to explain the data. The chosen model has spline row term, random row and genetic effects, fixed check varieties effects, and residual competition effects in row and column directions modeled with Band(2) in column and sar(2) in row direction. The REML estimates of variance parameters are in Table F.8. Note that the direct genetic component is small, only 0.3 times the residual component and there are negative correlations in both the column and row directions, which, given the spacing between plots, is not expected in the column direction. One hypothesis to be tested is if the lodging effect interferes in the correlation effects.

Figure F.7. Plot of the empirical semi-variogram for the residuals of the Local 58 of Piracicaba. This refers to Model 19.

Table F.8. REML estimates of variance parameters for fitted Model 19 for the experiment of Piracicaba, Local 58.

Variance parameters		Ratios (γ)	Estimates	Standard errors
spl(Row)		0.04	13.53	16.43
Row	$(\tilde{\sigma}_r^2)$	0.07	22.71	13.48
Direct genetic	$(\tilde{\sigma}_g^2)$	0.30	95.58	57.13
Residual	$(\tilde{\sigma}^2)$	1.00	320.24	50.77
Spatial:Column	$(\tilde{\rho}_{c1})$	—	-0.14	0.07
Spatial:Column	$(\tilde{\rho}_{c2})$	—	0.14	0.07
Spatial:Row	$(\tilde{\theta}_{r1})$	—	0.39	0.10
Spatial:Row	$(\tilde{\theta}_{r2})$	—	-0.22	0.05

Table F.9 shows the Wald F tests for fixed terms and the estimated fixed effects are presented in Table F.10. Note that in this model for this experiment, there is evidence of significant differences between the group of test lines and the group of checks (Control) and within the checks (Control:Checks). However the denominator residual degrees of freedom are an approximation and for Control appears to be faulty for this data.

Table F.9. Wald F test for fixed terms from Model 19. Df means degrees of freedom and DenDF is the approximate denominator degrees of freedom.

Fixed terms	Df	denDF	F test	p-value
(Intercept)	1.00	109.70	9404.00	0.00
Control	1.00	0.00	3.95	0.00
lin(Row)	1.00	97.50	2.74	0.10
Control:Check	3.00	14.00	28.10	0.00

Table F.10. Estimates of the fixed effects for Model 19 and their respective standard errors.

Effects	Estimates	Standard errors
INTERC	0.00	—
PAD1	-33.28	6.26
PAD5	-50.73	6.30
PAD6	-9.45	6.30
lin(Row)	-0.19	0.11
overall Checks	0.00	—
overall Test lines	-29.51	10.08
overall mean (μ)	158.29	10.29

The 30 best test lines with predicted values and respective standard errors from Model 19 are presented in Table F.11. Comparing the selection did at the Model 19 and the traditional model (with random row, column and direct genetic effects), there are some best test lines different for both models. The correlation between the EBLUPs is also high, 0.966, as shown in Figure F.8.

Table F.11. The 30 best test lines with predicted values (pred. value) and respective standard errors (stand. error) for Model 19.

Test lines	Pred. values	Stand. errors	Test lines	Pred. values	Stand. errors
PI370	147.98	8.81	PI162	140.10	8.81
PI340	143.56	8.85	PI150	139.87	8.82
PI239	142.72	8.86	PI395	139.12	8.81
PI38	142.67	8.76	PI104	139.08	8.81
PI18	142.37	8.83	PI400	139.01	8.82
PI82	142.18	8.86	PI138	138.62	8.76
PI17	141.48	8.77	PI96	138.34	8.79
PI51	141.42	8.81	PI37	137.95	8.80
PI387	140.79	8.83	PI320	137.82	8.81
PI215	140.70	8.79	PI318	137.74	8.83
PI183	140.45	8.84	PI414	137.44	8.82
PI269	140.42	8.78	PI300	137.25	8.81
PI196	140.30	8.84	PI167	137.09	8.78
PI293	140.28	8.76	PI121	137.03	8.83
PI165	140.20	8.85	PI368	136.91	8.84

Figure F.8. Dispersion plot of the genetic EBLUPs values of the Model 19 and the traditional non-spatial analysis model for this design (without spatial dependence) for Piracicaba Local 58. The cut-offs for the 30 best test lines (7 % upper) in each model are indicated by the dotted line.

Analyzing Local 76, the random column component was boundary for the fitted models. Observe in Table F.12 the sequence of models that was tested. While the genetic competition component was significant, AR(1) was used for the local residual effects (models 2 and 3), it was not significant once the more appropriate local effects model of $\text{Id} \times \text{Band}(2)$ was fitted (models 18 and 19). Hence, Model 19 is selected to represent the data.

Table F.12. Summary of the fitted models to Local 76 of Piracicaba with restricted log-likelihood (log-lik.) and the p -value of the REML ratio test. All models include a random direct genetic effects for the test lines (G) and fixed checks varieties effects. The other effects are represented with the symbols: H (harvest); Co (cone); S (sowing); C (column); R (row); N (neighbour genetic); spl(.) (spline in some direction indicated in parentheses) and lin(.) (linear trend in some direction indicated in parentheses). The structures can be: US(.) - unstructured; RR(.) - reduced rank; AR(.) - autoregressive; Band(.) - banded correlation, the number indicates the order of correlation; sar(2) - constrained autoregressive; and Id - identity.

Model	Effects			log-lik.	Test	p -value
	Fixed	Global/extraneous	Genetic Random			
1		R	US(G:N)	AR(1) $\times$ AR(1)		
2		R	RR(G:N)	AR(1) $\times$ AR(1)	M1 vs M2	1.00
3		R	G	AR(1) $\times$ AR(1)	M2 vs M3	0.05
4		R	RR(G:N)	AR(1) $\times$ Band(3)		
5		R	RR(G:N)	AR(1) $\times$ Band(2)	M4 vs M5	0.09
6		R	RR(G:N)	AR(1) $\times$ Band(1)	M5 vs M6	0.00
7		R	RR(G:N)	AR(1) $\times$ sar(2)		
8		R	RR(G:N)	AR(1) $\times$ AR(2)		
9		R	RR(G:N)	Id $\times$ Band(2)	M5 vs M9	0.78
10	Co	R	RR(G:N)	Id $\times$ Band(2)	not significant	
11	S	R	RR(G:N)	Id $\times$ Band(2)	not significant	
12	H	R	RR(G:N)	Id $\times$ Band(2)	not significant	
13	lin(C)+lin(R)	spl(C)+spl(R)+R	RR(G:N)	Id $\times$ Band(2)		
14	lin(C)+lin(R)	R	RR(G:N)	Id $\times$ Band(2)	M13 vs M14	1.00
15	lin(C)+lin(R)		RR(G:N)	Id $\times$ Band(2)	M14 vs M15	0.00
16	lin(C)	R	RR(G:N)	Id $\times$ Band(2)	not significant	
17	lin(R)	R	RR(G:N)	Id $\times$ Band(2)	not significant	
18		R	RR(G:N)	Id $\times$ Band(2)		
19		R	G	Id $\times$ Band(2)	M18 vs M19	0.13

Figure F.9 shows the plots of the faces of the empirical variogram for the residuals from Model 18 and 19. Using reduced rank models, the empirical variogram plot had some problems with the simulated envelopes. However, it is possible to observe the shape of residuals for both models. The REML estimates of variance parameters from Model 19 are in Table F.13.

Figure F.9. Plots of the column and row faces of the empirical variogram for the residuals for Local 54 of Piracicaba experiment for models 18 (panels (a) and (c)) and 19 (panels (b) and (d)). The panels (a) and (b) are column direction and the others are row direction.

Table F.13. REML estimates of variance parameters for fitted Model 19 for the experiment of Piracicaba, Local 76.

Variance parameters		Ratios (γ)	Estimates	Std. errors
Row	$(\tilde{\sigma}_r^2)$	0.46	145.92	44.42
Direct genetic	$(\tilde{\sigma}_g^2)$	0.66	208.99	62.41
Residual	$(\tilde{\sigma}^2)$	1.00	317.95	53.45
Spatial:Row	$(\tilde{\rho}_{r1})$	—	-0.24	0.08
Spatial:Row	$(\tilde{\rho}_{r2})$	—	0.26	0.07

Note that the genetic component is small than the residual (ratio is 0.66) and the row correlations are opposite in magnitude (-0.24 and 0.26) indicating there are competition at the residual level and spatial dependence. Table F.14 presents the Wald F tests for the fixed terms in Model 19 and Table F.15 shows the estimated fixed effects. As it can be seen, there is no evidence of a significant difference between the check and test lines groups. This contrast is denoted as Control. However, there are significant differences between the checks (Control:Check).

Table F.14. Wald F test for fixed terms from Model 19. Df means degrees of freedom and DenDF is the approximate denominator degrees of freedom.

Fixed terms	Df	denDF	F test	p-value
(Intercept)	1.00	496.00	4128.00	0.00
Control	1.00	496.00	1.16	0.28
Control:Check	3.00	496.00	39.68	0.00

Table F.15. Estimates of the fixed effects from Model 19 and their respective standard errors.

Effects	Estimates	Standard errors
INTERC	0.00	—
PAD1	-12.20	6.08
PAD5	-67.14	6.18
PAD6	-3.53	6.14
overall Checks	0.00	—
overall Test lines	-24.50	14.66
overall mean (μ)	173.89	14.76

The 30 best test lines with predicted values and respective standard errors from Model 19 are presented in Table F.16. Comparing the genetic EBLUPs from models 19 and the traditional (with random row, column and direct genetic effects), the correlation is strong, 0.956, but the selected test lines are not the same between the models. See Figure F.10.

Table F.16. The 30 best test lines with predicted values (pred. value) and respective standard errors (stand. error) for Model 19.

Test lines	Pred. values	Stand. errors	Test lines	Pred. values	Stand. errors
PI163	184.06	11.44	PI171	166.24	11.26
PI318	173.91	11.33	PI27	166.20	11.37
PI269	173.74	11.32	PI395	164.63	11.43
PI385	171.58	11.32	PI42	164.36	11.31
PI53	171.43	11.30	PI336	164.17	11.30
PI388	171.23	11.28	PI356	164.07	11.32
PI7	169.36	11.25	PI348	163.90	11.32
PI17	168.86	11.29	PI206	163.88	11.28
PI220	168.82	11.31	PI204	163.71	11.31
PI359	168.00	11.27	PI95	163.63	11.26
PI207	167.67	11.29	PI180	163.59	11.28
PI267	167.05	11.28	PI83	163.22	11.25
PI115	166.43	11.31	PI422	162.98	11.30
PI340	166.38	11.37	PI416	162.97	11.27
PI405	166.34	11.28	PI191	162.77	11.32

Figure F.10. Dispersion plot of the genetic EBLUPs values of the Model 19 and the traditional non-spatial analysis model for this design (without spatial dependence) for Piracicaba Local 76. The cut-offs for the 30 best test lines (7 % upper) in each model are indicated by the dotted line.

F.3 Joint analysis

There is low correlation, around 0.3, between the genetic EBLUPs from each Local obtained using the models selected in the individual analyses. As shown in Figure F.11, there are few test lines considered the best in pairs of Locals; only five or six.

Figure F.11. Dispersion plots for the genetic EBLUPs values from selected model for each Piracicaba Local. Panels refer the relation between Local (a) 54 and 58; (b) 54 and 76; (c) 58 and 76. The cut-offs for the 30 best test lines in each Local are indicated by the dotted line.

Based on the separate analysis for each Local, a joint analysis of all three Locals was formulated

in order to investigate the magnitudes of the genetic variance at each Local and to assess the genetic covariance between Locals. To do this the models shown in Table F.17 were tested and the results of the REML tests are also given in Table F.17.

Table F.17. Summary of the fitted models to joint analysis of the Piracicaba experiments carried out in the Local 54 (1), 58 (2) and 76 (3), with REML log (log-lik.) and the p -value of the REML ratio test. All the models have the same effects modeled in the individual analysis plus the fixed Local effects. Here σ_{gi}^2 is the genetic variance at the i th Local, σ_{gij} is the genetic correlation between the i th and j th Locals and σ_{g*} is the genetic correlation between two locals which has been hypothesis to be equal for two or more pairs of Locals.

Model	Genetic Effects	Log-lik.	Test	p-value
1	$\begin{bmatrix} \sigma_{g1}^2 & 0 & 0 \\ 0 & \sigma_{g2}^2 & 0 \\ 0 & 0 & \sigma_{g3}^2 \end{bmatrix}$	-5353.00		
2	$\begin{bmatrix} \sigma_{g1}^2 & \sigma_{g12} & \sigma_{g13} \\ \sigma_{g12} & \sigma_{g2}^2 & \sigma_{g23} \\ \sigma_{g13} & \sigma_{g23} & \sigma_{g3}^2 \end{bmatrix}$	-5302.79	M2 vs M1	0.00
3	$\begin{bmatrix} \sigma_{g1}^2 & \rho\sigma_{g1}\sigma_{g2} & \rho\sigma_{g1}\sigma_{g3} \\ \rho\sigma_{g1}\sigma_{g2} & \sigma_{g2}^2 & \rho\sigma_{g2}\sigma_{g3} \\ \rho\sigma_{g1}\sigma_{g3} & \rho\sigma_{g2}\sigma_{g3} & \sigma_{g3}^2 \end{bmatrix}$	-5303.10	M2 vs M3	0.73
4	$\begin{bmatrix} \sigma_g^2 & \rho\sigma_g\sigma_{g2} & \rho\sigma_g^2 \\ \rho\sigma_g\sigma_{g2} & \sigma_{g2}^2 & \rho\sigma_g\sigma_{g2} \\ \rho\sigma_g^2 & \rho\sigma_g\sigma_{g2} & \sigma_g^2 \end{bmatrix}$	-5303.11	M3 vs M4	0.44
5	$\sigma_g^2 \begin{bmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{bmatrix}$	-5303.64	M4 vs M5	0.15

Using the REML ratio test, Model 5 is chosen as the best to explain the effects that occurred in these experiments. The model has the same nuisance effects found in individual analysis added fixed local effects, a single direct genetic variance for all Locals and genetic correlation between Locals that was equal for all pairs of Locals. Table F.18 shows the REML estimates of the variance parameters from select model.

Table F.18. REML estimates of the variance parameters for fitted Model 5 in the joint analysis of the Locals in Piracicaba.

Variance parameters		Estimates	Standard errors
Local: Direct genetic	$(\tilde{\sigma}_g^2)$	155.10	32.28
cor(Local):Direct genetic	$(\tilde{\rho}_g)$	0.84	0.17
Local54:Row	$(\tilde{\sigma}_{r1}^2)$	60.14	20.24
Local58:Row	$(\tilde{\sigma}_{r2}^2)$	23.10	12.78
Local76:Row	$(\tilde{\sigma}_{r3}^2)$	130.07	39.60
Local58:spl(Row)		12.53	14.73
Local 54: Residual	$(\tilde{\sigma}_1^2)$	251.17	31.98
Spatial:Row	$(\tilde{\rho}_{1r})$	-0.23	0.06
Local 58:Residual	$(\tilde{\sigma}_2^2)$	275.36	31.96
Spatial:Column	$(\tilde{\rho}_{2c1})$	-0.12	0.07
Spatial:Column	$(\tilde{\rho}_{2c2})$	0.15	0.07
Spatial:Row	$(\tilde{\theta}_{2r1})$	0.39	0.10
Spatial:Row	$(\tilde{\theta}_{2r2})$	-0.21	0.05
Local 76: Residual	$(\tilde{\sigma}_3^2)$	354.68	38.54
Spatial:Row	$(\tilde{\theta}_{3r1})$	-0.20	0.06
Spatial:Row	$(\tilde{\theta}_{3r2})$	0.19	0.05

Comparing the predicted values from model 5, given in Table F.19, it is observed that these results had small standard errors, when compared to the separate analysis of the Locals, and a large

number of clones in common among the Locals within the top 30 test lines.

Table F.19. Predicted values (pred.) and standard errors (error) of the 30 best test lines from Model 5 for each Local.

Local 54			Local 58			Local 76		
Test lines	Pred.	Error	Test lines	Pred.	Error	Test lines	Pred.	Error
PI318	138.75	8.46	PI318	160.99	8.67	PI318	182.22	8.77
PI340	130.75	8.48	PI340	158.44	8.69	PI340	176.23	8.80
PI269	127.20	8.46	PI269	154.56	8.63	PI269	174.54	8.78
PI356	125.43	8.46	PI395	150.99	8.67	PI395	169.73	8.81
PI395	124.20	8.47	PI370	150.64	8.63	PI388	168.53	8.77
PI320	123.31	8.43	PI38	149.54	8.60	PI356	168.48	8.78
PI388	122.31	8.46	PI18	148.55	8.66	PI320	166.85	8.75
PI203	121.59	8.45	PI320	148.30	8.64	PI385	166.74	8.78
PI342	121.17	8.44	PI196	147.09	8.67	PI220	166.57	8.76
PI136	120.97	8.44	PI356	147.08	8.63	PI7	165.90	8.76
PI38	120.68	8.45	PI82	146.62	8.70	PI38	165.53	8.76
PI220	120.15	8.45	PI17	146.33	8.62	PI18	165.26	8.76
PI37	120.01	8.45	PI388	146.08	8.69	PI17	165.23	8.78
PI408	119.70	8.45	PI37	145.97	8.64	PI136	164.70	8.75
PI196	119.41	8.46	PI203	145.90	8.63	PI191	164.61	8.79
PI160	119.26	8.44	PI150	145.67	8.66	PI394	164.04	8.77
PI370	119.16	8.44	PI96	145.43	8.63	PI218	163.89	8.76
PI385	119.05	8.47	PI385	145.35	8.67	PI203	163.87	8.75
PI7	118.87	8.45	PI215	145.30	8.64	PI82	163.84	8.78
PI18	118.86	8.45	PI160	144.91	8.68	PI348	163.79	8.76
PI100	118.71	8.48	PI7	144.87	8.68	PI196	163.59	8.77
PI191	118.59	8.50	PI138	144.86	8.58	PI370	163.58	8.76
PI150	118.47	8.44	PI220	144.78	8.61	PI192	163.31	8.79
PI296	118.30	8.44	PI191	144.69	8.70	PI160	163.28	8.76
PI192	118.22	8.50	PI192	144.57	8.68	PI115	163.11	8.77
PI218	118.19	8.46	PI51	144.56	8.65	PI37	162.98	8.77
PI394	118.14	8.46	PI136	144.46	8.64	PI408	162.90	8.75
PI297	118.11	8.45	PI342	144.19	8.63	PI95	162.87	8.75
PI348	118.03	8.45	PI165	144.10	8.69	PI237	162.87	8.75
PI115	117.68	8.45	PI408	143.92	8.65	PI287	162.72	8.74

F.4 Conclusions

With each individual analysis it can be noted the peculiarity of the Local, in which was modeled random row effects and residual competition in this direction, but spatial effects were not found in Local 54. The models in Local 58 showed a negative correlation in the column direction. A possible explanation is that there is an operational effect such that the sugarcane for one plot ends up being erroneously included in that for a plot in a neighbouring column. The ratios between genetic and residual components were approximately 0.77, 0.30 and 0.66 for Locals 54, 58 and 76 respectively. This results in better selection for Local 54, because it has the largest genetic variance relative to the residual variance. When comparing the 30 best predicted values for the test lines from each Local there were few similarities among the lines selected; there were around three to six common test lines across Locals and the correlation between the EBLUPs from pairs of Locals was around 0.30. However, in the joint analysis better accuracy in the selection of the test lines is obtained, with predictions having lower standard errors. The model chosen for the genetic variation indicates the same genetic component (155.10) for all three Locals and a positive strong correlation (0.84) between the genetic effects from different Locals. In this case, the better selection is in Local 54, because it has the lowest residual variance. The number of common test lines in the best 30 lines is better; around half of the 30 best predicted values for the test lines for each

Local occur in all three Locals.

G ARAÇATUBA - GROUP OF EXPERIMENTS

There were three sites, nominated: Local 101, 130 and 551. However, each site had small areas with different test lines (New) and rectangular arrays. For Local 101, there are three areas with 3×18 , 15×21 and 3×15 columns by rows. In Local 130 and 551 were two areas each and dimensions 26×12 , 5×15 , 9×21 and 11×24 columns by rows. The number of test lines assessed were 325 for Local 101; 321 for Local 130; and 318 for Local 551. There were 4 commercial varieties nominated “PAD1”, “PAD2”, “PAD3” and “Interc”, each having different numbers of replicates for each Local. These experiments were planted in 2014 and the yield in TCH was obtained in 2015, in a single cut.

In subsection G.1 the results of the descriptive analysis for Local 101 will be described and the model testing and the EBLUPs obtained using the selected model for this Local will be presented in subsubsection G.1.1. Subsection G.2 and subsubsection G.2.1 contain the same information for Local 130 and that for Local 551 is in subsection G.3 and subsubsection G.3.1. Subsection G.4 contains the conclusions for these trials.

G.1 Local 101

A summary for the TCH of the groups of clones carried out in Local 101 of Araçatuba is presented in Table G.1. It can be observed that in areas 1 and 3 there are no replications of “PAD” checks. The value of the average for PAD1 found in area 3 is smaller than in other sites. Observations of Figure G.1, that shows the boxplot of TCH of the groups of clones in each area, reveals some possible outliers, but these are test lines and PAD3 and the values are similar to those in the other experiments.

Table G.1. Descriptive analysis of the groups of clones carried out in Local 101 of Araçatuba.

	Groups of clones	Minimum	Maximum	Mean	Variance	Number of plots
Area 1	PAD1	-	-	135.70	-	1
	PAD2	-	-	123.80	-	1
	PAD3	-	-	123.80	-	1
	Inerspersed (INTERC)	117.90	151.20	134.13	222.67	6
	Test line (New)	76.20	152.40	114.09	373.03	42
Area 2	PAD1	119.00	129.80	125.24	24.05	5
	PAD2	89.30	133.30	109.76	387.13	5
	PAD3	110.70	156.00	124.76	318.66	5
	Interspersed (INTERC)	72.60	154.80	111.80	421.28	35
	Test line (New)	66.70	173.80	114.35	360.41	252
Area 3	PAD1	-	-	69.00	-	1
	PAD2	-	-	133.30	-	1
	PAD3	-	-	114.30	-	1
	Interspersed (INTERC)	72.60	129.80	99.30	550.82	5
	Test line (New)	82.10	150.00	115.70	290.98	31

Figure G.1. Boxplot of TCH in each area for Local 101 of Araçatuba.

See in Figure G.2 the heat maps relating to TCH in each area.

Figure G.2. Heat maps relating to TCH in Local 101 of Araçatuba. The panel refers to (a) area 1; (b) area 3 and (c) area 2. The names PAD1, PAD2, PAD3 and INTERC are the check plots and the test lines are labeled alpha-numerically (AR plus number).

G.1.1 Analysis

In area 1, there is no evidence of spatial dependence or competition effect. Table G.2 presents the summary of the fitted models. When comparing the models, it can be observed that the log likelihood for Model 15 is smallest and was significantly different from Model 14; hence it was selected as the best model.

Table G.2. Summary of the fitted models to Local 101 area 1 with restricted log-likelihood (log-lik.) and the p -value of the REML ratio test. All the models include a random direct genetic effects for the test lines (G) and fixed checks varieties effects. The other effects are represented with the symbols: H (harvest); Co (cone); S (sowing); C (column); R (row); N (neighbour genetic); spl(.) (spline in some direction indicated in parentheses) and lin(.) (linear trend in some direction indicated in parentheses). The structures can be: US(.) - unstructured; RR(.) - reduced rank; AR(1) - autoregressive; Band(.) - banded correlation, the number indicates the order of correlation; and Id - identity.

Model	Effects			Log-lik.	Test	p -value
	Global/extraneous Fixed	Random	Genetic (column $\times$ row)			
1		C + R	US(G:N)	-159.54		
2		C	US(G:N)	-159.54	M1 vs M2	1.00
3		C	RR(G:N)	-159.52	M2 vs M3	1.00
4		C	G	-159.65	M3 vs M4	0.30
5		C	G	not converged - singularity		
6		C	G	-159.56		
7		C	G	-159.56	M6 vs M7	1.00
8		C	G	-159.92	M7 vs M8	0.39
9		C	G	-160.05	M8 vs M9	0.61
10	H	C	G	-158.38	not significant	
11	Co	C	G	-158.41	not significant	
12	S	C	G	-158.40	not significant	
13	lin(C)+lin(R)	C+spl(R)	G	-157.24		
14	lin(C)+lin(R)		G	-157.24	M13 vs M14	1.00
15	lin(C)		G	-156.93	significant	
16	lin(R)		G	-161.21	not significant	
17	lin(C)		RR(G:N)	-161.21	M17 vs M15	0.44

In Figure G.3, there are the plots of the row and column faces of the empirical variogram for the residuals referring to Models 4 and 15. The values obtained are inside the expected range, this shows that there is no lack of adjustment on the spatial dependence.

Figure G.3. Plots of the row and column faces of the empirical variogram for the residuals for Local 101 area 1 of Araçatuba experiment for models 4 (panels (a) and (c)) and 15 (panels (b) and (d)). The panels (a) and (b) are column direction and the others are row direction.

Table G.3 presents the Wald F tests for the fixed terms for Model 15 and Table G.4 shows the estimated fixed effects. As it can be seen, there is no evidence of significant difference between check and the mean of the test lines (Control), nor among the checks (Control:Check).

Table G.3. Wald F test for fixed effects for Model 15. Df means degrees of freedom and DenDF is the approximate denominator degrees of freedom.

Fixed terms	Df	denDF	F test	p-value
(Intercept)	1	38.90	1548.00	0.00
Control	1	13.70	1.14	0.30
lin(Col)	1	19.20	7.69	0.01
Control:Check	3	4.50	0.35	0.79

Table G.4. Estimates of the fixed effects and their respective standard errors for Model 15.

Effects	Estimates	Standard errors
INTERC	0.00	—
PAD1	-6.44	11.51
PAD2	-2.33	11.51
PAD3	-10.33	11.11
lin(Col)	-8.00	3.01
overall Test lines	0.00	—
overall checks	19.47	16.99
overall mean (μ)	130.67	6.90

The 15 predicted best test lines with standard errors are presented in Table G.5.

Table G.5. The 15 best test lines with predicted values (pred. values) and respective standard errors (stand. errors) for Model 15.

Test lines	Pred. values	Stand. errors	Test lines	Pred. values	Stand. errors
AR 154	144.70	9.00	AR 158	111.63	9.00
AR 143	142.99	9.00	AR 323	111.26	8.97
AR 65	139.00	8.72	AR 157	110.77	9.00
AR 324	135.86	8.97	AR 83	110.77	9.00
AR 319	132.43	8.97	AR 148	110.13	8.72
AR 91	132.22	8.72	AR 311	109.55	8.97
AR 68	132.01	9.00	AR 315	109.55	8.97
AR 142	126.03	9.00	AR 317	108.70	8.97
AR 82	126.03	9.00	AR 138	107.63	8.72
AR 326	123.10	8.97	AR 72	107.63	8.72
AR 3	122.89	8.72	AR 139	106.78	8.72
AR 313	122.24	8.97	AR 98	105.28	8.97
AR 159	121.82	9.00	AR 149	103.36	8.72
AR 69	121.18	8.72	AR 67	103.14	9.00
AR 153	119.46	8.72	AR 64	102.29	9.00
AR 90	115.26	8.72	AR 160	101.64	8.72
AR 145	114.98	9.00	AR 310	100.22	8.97
AR 163	114.19	9.00	AR 141	98.87	9.00
AR 140	113.55	8.72	AR 66	96.58	8.72
AR 144	112.69	8.72	AR 63	92.95	9.00
AR 308	112.05	8.97	AR 309	82.40	8.97

In area 2, an analysis of the data using different models was tried, but none were satisfactory. The problem was that the residual variance was high and the direct genetic component was close to zero. Thus, two models were compared: Model 1 considered competition effects at the residual level; and Model 2 as the simplest model with only variety effects (checks fixed and test lines random). The results are given in Table G.6.

```
Model1<- asreml(tch ~ Control/Check, random=~ Reg +Col + Row,
               rcov=~ ar1(Col):corb(Row,k=3),data=place2,
               na.method.Y = "include",na.method.X = "include")
```

```
Model2<- asreml(tch ~ Control/Check, random=~ Reg, rcov=~ (Col):(Row),
               data=place2,na.method.Y = "include",na.method.X = "include")
```

Table G.6. REML estimates of variance parameters from models 1 and 2 for the experiment in Local 101 of Araçatuba in area 2. The values in brackets are the standard errors of the estimates.

Variance parameters		Model 1		Model 2	
		Estimates	Standard errors	Estimates	Standard errors
Column	$(\tilde{\sigma}_c^2)$	20.51	(13.15)	—	—
Row	$(\tilde{\sigma}_r^2)$	0.00	(0.00)	—	—
Direct genetic	$(\tilde{\sigma}_g^2)$	0.00	(0.00)	0.00	(0.00)
Residual	$(\tilde{\sigma}^2)$	342.71	(29.05)	362.65	(29.76)
Spatial	$(\tilde{\rho}_c)$	0.05	(0.06)	—	—
Spatial	$(\tilde{\rho}_{r_1})$	-0.03	(0.06)	—	—
Spatial	$(\tilde{\rho}_{r_2})$	0.01	(0.06)	—	—
Spatial	$(\tilde{\rho}_{r_3})$	-0.12	(0.06)	—	—

The same models were fitted for area 3 and the results are presented in Table G.7.

Table G.7. REML estimates of variance parameters from models 1 and 2 for the experiment in Local 101 of Araçatuba in area 3. The values in brackets are the standard errors of the estimates.

Variance parameters		Model 1		Model 2	
		Estimates	Standard errors	Estimates	Standard errors
Column	$(\tilde{\sigma}_c^2)$	0.00	(0.00)	—	—
Row	$(\tilde{\sigma}_r^2)$	106.98	(67.96)	—	—
Direct genetic	$(\tilde{\sigma}_g^2)$	0.00	(0.00)	0.00	(0.00)
Residual	$(\tilde{\sigma}^2)$	226.08	(77.48)	321.55	(77.99)
Spatial	$(\tilde{\rho}_c)$	-0.20	(0.42)	—	—
Spatial	$(\tilde{\rho}_{r_1})$	-0.08	(0.26)	—	—
Spatial	$(\tilde{\rho}_{r_2})$	0.28	(0.17)	—	—
Spatial	$(\tilde{\rho}_{r_2})$	-0.38	(0.15)	—	—

G.2 Local 130

For Local 130, high values of variance were found in the groups of clones in area 1, mainly for “PAD2”, as can be seen in Table G.8. In Figure G.4, the dispersion of the data can be observed for each group. Some plots are considered outliers and comparing the yield between sites seems to confirm this. Hence, it can be expected that the yield of a clone can be affected by the position of the plot to which it is sown. Figure G.5 illustrates the heat maps relating to TCH in each area of the Local 130 of Araçatuba.

Table G.8. Descriptive analysis of the groups of clones carried out in Local 130 of Araçatuba.

	Groups of clones	Minimum	Maximum	Mean	Variance	Number of plots
area 1	PAD1	95.20	128.60	112.86	217.02	5
	PAD2	88.10	166.70	117.86	914.87	5
	PAD3	82.10	129.80	115.00	367.02	5
	Inerspersed (INTERC)	69.00	169.00	113.51	515.85	35
	Test lines (New)	73.80	191.70	120.52	451.16	262
area 2	PAD1	-	-	104.80	-	1
	PAD2	-	-	169.00	-	1
	PAD3	-	-	158.30	-	1
	Inerspersed (INTERC)	98.80	140.50	120.72	223.87	10
	Test lines (New)	85.70	164.30	125.05	350.97	62

Figure G.4. Boxplot of TCH in each area for Local 130 of Araçatuba.

(a)

(b)

Figure G.5. Heat maps relating to TCH in Local 130 of Araçatuba. The panel refers to (a) area 1 and (b) area 2. The names PAD1, PAD2, PAD3 and INTERC are the check plots and the test lines are labeled alpha-numerically (AR plus number). “x” represents the empty plots.

G.2.1 Analysis

As for the Local 101, the direct genetic variance was estimated as zero for all fitted models. Consequently, the following two models were compared and the resulting estimates of the variance parameters are in Table G.9.

```
Model1<- asreml(tch ~ Control/Check, random=~ Reg +
  Col + Row,rcov=~ ar1(Col):corb(Row,k=3),data=place21,
  na.method.Y = "include",na.method.X = "include")
```

```
Model2<- asreml(tch ~ Control/Check, random=~ Reg,
  rcov=~ (Col):(Row),data=place21,
  na.method.Y = "include",na.method.X = "include")
```


Table G.9. REML estimates of variance parameters for models 1 and 2 for the experiment in Local 130 of Araçatuba in area 1. The values in brackets are the standard errors of the estimates.

Variance parameters		Model 1		Model 2	
		Estimates	Standard errors	Estimates	Standard errors
Column	$(\tilde{\sigma}_c^2)$	46.04	(25.08)	—	—
Row	$(\tilde{\sigma}_r^2)$	0.00	(0.00)	—	—
Direct genetic	$(\tilde{\sigma}_g^2)$	0.00	(0.00)	0.00	(0.00)
Residual	$(\tilde{\sigma}^2)$	417.59	(36.95)	460.31	(37.34)
Spatial	$(\tilde{\rho}_c)$	-0.04	(0.06)	—	—
Spatial	$(\tilde{\rho}_{r_1})$	-0.05	(0.08)	—	—
Spatial	$(\tilde{\rho}_{r_2})$	0.08	(0.07)	—	—
Spatial	$(\tilde{\rho}_{r_3})$	-0.02	(0.08)	—	—

For area 2, the models capture a direct genetic effect, however models with competition at the genetic level did not converge and all the models had a high standard error for the variance components. Table G.10 shows a sequence of fitted models. Using REML ratio test and diagnostic graphs of the residuals, Model 8 was considered to be the best fit; it included only fixed varieties and random direct genetic effects. See the empirical variogram for this model in Figure G.6.

Table G.10. Summary of the models fitted to area 2 of the Local 130 with restricted log-likelihood (log-lik.) and the p -value of the REML ratio test. All the models include a random direct genetic effects for the test lines (G) and fixed checks varieties effects. The other effects are represented with the symbols: H (harvest); Co (cone); S (sowing); C (column); R (row); N (neighbour genetic); spl(.) (spline in some direction indicated in parentheses) and lin(.) (linear trend in some direction indicated in parentheses). The structures can be: US(.) - unstructured; AR(1) - first-order autoregressive; Band(.) - banded correlation, the number indicates the order of correlation; and Id - identity.

Model	Effects				Log-lik.	Test	p -value
	Global/extraneous		Genetic	Local (column $\times$ row)			
	Fixed	Random					
1		R + C	US(G:N)	AR(1) $\times$ AR(1)	not converged - singularity		
2			US(G:N)	AR(1) $\times$ AR(1)	not converged - singularity		
3		R + C	G	AR(1) $\times$ Band(3)	-240.63	M3 vs M4	1.00
4			G	AR(1) $\times$ Band(3)	-240.63	M4 vs M5	0.83
5			G	AR(1) $\times$ Band(2)	-240.65	M5 vs M6	0.36
6			G	AR(1) $\times$ Band(1)	-241.07	M6 vs M7	0.50
7			G	AR(1) $\times$ Id	-241.30	M7 vs M8	0.88
8			G	Id $\times$ Id	-241.31	not significant	
9	H		G	Id $\times$ Id	-238.92	not significant	
10	S		G	Id $\times$ Id	-239.78	not significant	
11	Co		G	Id $\times$ Id	-239.84	not significant	
12	lin(C)+lin(R)	spl(C)+spl(R)	G	Id $\times$ Id	-240.33		
13	lin(C)+lin(R)	spl(R)	G	Id $\times$ Id	-240.33	M12 vs M13	1.00
14	lin(C)+lin(R)		G	Id $\times$ Id	-240.53	M13 vs M14	0.27
15	lin(C)		G	Id $\times$ Id	-239.94	not significant	
16	lin(R)		G	Id $\times$ Id	-241.94	not significant	
17			US(G:N)	Id $\times$ Id	not converged - singularity

Figure G.6. Plots of the column and row faces of the empirical variogram for the residuals for Local 130 area 2 of Araçatuba experiment for Model 8. The panels (a) is column direction and (b) is row direction.

The REML estimates of variance parameters, with ratios (γ) and standard errors, from fitted Model 8 are presented in the Table G.11. As said before, the standard errors are very high when compared with their respective components, mainly for direct genetic term. Note that the parameter estimate was 127.10, while the standard error was 123.19.

Table G.11. REML estimates of variance parameters for fitted Model 8 to the experiment in the Local 130 of the Araçatuba.

Variance parameters	Ratios (γ)	Estimates	Standard errors
Direct genetic ($\tilde{\sigma}_g^2$)	0.57	127.10	123.19
Residual ($\tilde{\sigma}^2$)	1.00	223.87	105.53

Table G.12 presents the Wald F tests for the fixed terms of the Model 8 and, in the Table G.13, there are the estimated fixed effects. See that there is no significant difference between check and the mean of the test lines, nor among the checks.

Table G.12. Wald F tests for the fixed effects for Model 8. Df means degrees of freedom and DenDF is the approximate denominator degrees of freedom.

Fixed terms	Df	denDF	F test	p-value
(Intercept)	1	53.90	2873.00	0.00
Control	1	3.30	0.01	0.94
Control:Check	3	9.00	5.39	0.02

Table G.13. Estimated fixed parameters and their respective standard errors for Model 8.

Effects	Estimates	Standard errors
INTERC	0.00	—
Check - PAD1	-15.92	15.69
Check - PAD2	48.28	15.69
Check - PAD3	37.58	15.69
overall Test lines	0.00	—
overall checks	-4.33	12.45
overall mean (μ)	125.05	2.38

Table G.14 presents the predicted values and standard errors of test lines for Model 8.

Table G.14. Predicted values and standard errors of the test lines for Model 8.

Test lines	Pred. value	Stand. error	Test lines	Pred. value	Stand. error
AR 178	139.26	9.13	AR 230	125.47	9.13
AR 215	137.09	9.13	AR 136	125.03	9.13
AR 150	135.82	9.13	AR 228	125.03	9.13
AR 163	135.39	9.13	AR 131	124.16	9.13
AR 125	134.52	9.13	AR 221	124.16	9.13
AR 161	134.09	9.13	AR 70	123.73	9.13
AR 158	133.22	9.13	AR 148	123.29	9.13
AR 135	132.78	9.13	AR 164	123.29	9.13
AR 220	132.78	9.13	AR 129	122.86	9.13
AR 162	132.35	9.13	AR 139	121.59	9.13
AR 127	131.91	9.13	AR 146	121.59	9.13
AR 138	131.91	9.13	AR 216	121.59	9.13
AR 140	131.91	9.13	AR 169	120.72	9.13
AR 165	131.91	9.13	AR 157	120.29	9.13
AR 145	130.65	9.13	AR 229	120.29	9.13
AR 156	130.21	9.13	AR 142	119.42	9.13
AR 133	129.78	9.13	AR 153	119.42	9.13
AR 154	129.78	9.13	AR 128	118.98	9.13
AR 143	128.92	9.13	AR 149	118.98	9.13
AR 97	128.92	9.13	AR 175	118.98	9.13
AR 130	128.47	9.13	AR 167	118.55	9.13
AR 155	128.47	9.13	AR 134	118.15	9.13
AR 124	128.04	9.13	AR 159	118.15	9.13
AR 132	127.20	9.13	AR 227	118.15	9.13
AR 137	127.20	9.13	AR 141	116.85	9.13
AR 80	126.34	9.13	AR 108	115.54	9.13
AR 147	125.90	9.13	AR 184	115.11	9.13
AR 223	125.90	9.13	AR 160	112.97	9.13
AR 224	125.90	9.13	AR 144	111.23	9.13
AR 126	125.47	9.13	AR 211	111.23	9.13
AR 151	125.47	9.13	AR 225	110.80	9.13

G.3 Local 551

In the descriptive analysis of the groups of clones for Local 551, one high value for the commercial variety “PAD3” in area 2 is observed, but this seems to be incorrect, because a high yield for this clone was never obtained. The variance for this group is bigger than the other groups, although “PAD2” also has a large variance in area 2 (Table G.15). It is possible to observe, in Figure G.7, that only some plots with test lines are considered outliers, however the dispersion of commercial variety in area 2 is not very symmetrical.

Table G.15. Descriptive analysis of the groups of clones carried out in Local 551 of Araçatuba.

	Groups of clones	Minimum	Maximum	Mean	Variance	Number of plots
area 1	PAD1	110.70	129.80	117.47	114.44	3
	PAD2	114.30	144.00	130.93	230.06	3
	PAD3	86.90	131.00	105.97	512.90	3
	Inerspersion (INTERC)	85.70	163.10	127.63	546.07	19
	Test lines (New)	58.30	201.20	123.71	517.37	161
area 2	PAD1	79.80	115.50	102.40	246.18	4
	PAD2	98.80	163.10	121.72	830.72	4
	PAD3	102.40	246.40	159.82	3800.04	4
	Inerspersion (INTERC)	81.00	147.60	115.03	334.84	29
	Test lines (New)	57.10	210.70	114.33	574.52	223

Figure G.7. Boxplot of TCH in each area for Local 551 of Araçatuba.

Figure G.8 has the heat maps relating to TCH in Local 551 for each area.

Figure G.8. Heat maps relating to TCH in the Local 551 of Araçatuba. The panel (a) refers to area 1 and (b) area 2. The names PAD1, PAD2, PAD3 and INTERC are the check plots and the test lines are labeled alpha-numerically (AR plus number). “x” represents the empty plots.

G.3.1 Analysis

Again, in area 1, there is a problem in estimating genetic effects. Three models fitted for area 1 are shown in Table G.16 and the estimates of the variance parameters from these models are shown in Table G.17. Even though the estimated genetic components are not zero, the standard error is considerably larger than the estimate and so can be taken to be zero.

Table G.16. Summary of the models fitted to Local 551 area 1 of the Araçatuba experiment with restricted log (log-lik.). All the models include a random direct genetic effect for the test lines, fixed effects for checks and there is one model with linear trend in column direction (lin(Col)). AR(1) is the first-order autoregressive structure and Id is the identity structure.

Model	Global/extraneous fixed	Effects		Log-lik.	Test	p-value
		Genetic Random	Local (column $\times$ row)			
1		G	AR(1) $\times$ AR(1)	-482.37		
2	lin(Col)	G	AR(1) $\times$ AR(1)	-480.35		
3	lin(Col)	G	Id $\times$ Id	-480.83		

Table G.17. REML estimates of variance parameters for models 1, 2 and 3 for the experiment in Local 551 of Araçatuba in area 1. The values in brackets are the standard errors (std.error) of the estimates.

Variance parameters		Model 1		Model 2		Model 3	
		Estimate	Std.error	Estimate	Std.error	Estimate	Std.error
Direct genetic	$(\tilde{\sigma}_g^2)$	51.38	(151.62)	98.72	(139.82)	95.40	(141.74)
Residual	$(\tilde{\sigma}^2)$	468.41	(134.29)	416.62	(120.90)	419.65	(122.60)
Spatial	$(\tilde{\rho}_c)$	-0.06	(0.11)	-0.07	(0.12)	—	—
Spatial	$(\tilde{\rho}_r)$	0.09	(0.11)	0.08	(0.12)	—	—

Figure G.9 has the plots of the row and column face of the empirical variogram for the residuals referring to Model 3, this shows that there is not a strong spatial dependence effect for this site.

Figure G.9. Plots of the column and row faces of the empirical variogram for the residuals for Local 551 area 1 of Araçatuba experiment for Model 3. The panels (a) is column direction and (b) is row direction.

For area 2, the direct genetic component is close to zero for all models tested. Table G.18 presents some models and in Table G.19 there are the estimated variance components and their respective standard errors.

Table G.18. Summary of the models fitted to Local 551 area 2 of the Araçatuba experiment with restricted log-likelihood (log-lik.). All the models include a random direct genetic effects for the test lines, fixed effects for checks and there is one model with linear trend in column direction (lin(Col)). AR(1) is the first-order autoregressive structure; Band(3) is the third-order banded correlation and Id is the identity structure.

Model	Global/extraneous		Effects	Local (column $\times$ row)	Log-lik.	Test	p-value
	Fixed	Random	Genetic				
1		Col + Row	G	AR(1) $\times$ Band(3)	-892.55		
2		Col	G	Id $\times$ Id	-892.88		
3	lin(Col)	Col	G	Id $\times$ Id	-887.77		

Table G.19. REML estimates of variance parameters from models 1 and 3 for the experiment in Local 551 of Araçatuba in area 2. The values in brackets are the standard errors (std.error) of the estimates.

Variance parameters		Model 1		Model 3	
		Estimate	Std.error	Estimate	Std.error
Column	$(\tilde{\sigma}_c^2)$	131.09	(69.39)	37.68	(27.56)
Row	$(\tilde{\sigma}_r^2)$	0.00	(0.00)	—	—
Direct genetic	$(\tilde{\sigma}_g^2)$	0.00	(0.00)	0.00	(0.00)
Residual	$(\tilde{\sigma}^2)$	472.40	(44.05)	470.94	(43.42)
Spatial	$(\tilde{\rho}_c)$	-0.04	(0.07)	—	—
Spatial	$(\tilde{\rho}_{r_1})$	0.00	(0.07)	—	—
Spatial	$(\tilde{\rho}_{r_2})$	0.03	(0.07)	—	—
Spatial	$(\tilde{\rho}_{r_3})$	-0.01	(0.07)	—	—

This data was also analysed without the “PAD3’’ outlier (first column 1 and fourth row, with a yield of 246.40), but the genetic component was smaller than its estimated standard error. Hence there is no adequate model for the selection of test lines.

G.4 Conclusions

The experiments carried out in Araçatuba present high environmental variation (σ^2) and small (close to zero) direct genetic components (σ_g^2) for Local 101 (areas 2 and 3), Local 130 (area 1) and Local 551 (both areas). Hence the models tested for these experiments did not result in definitive direct genetic effect.

At Local 101 area 1, it was possible to find nonzero EBLUPs for the test lines of this experiment with random direct genetic component equal to 262.37 and a residual component of 105.71. Local 130 area 2 also produced nonzero EBLUPs for the test lines, but the estimated standard error for direct genetic component was approximately the same as that of the variance parameter.

Two observations are made. Firstly, in every Local there was at least one experiment with less than 10 rows, which makes it difficult to estimate spatial dependence, and/or less than 100 plots which limits the ability to estimate variances. Secondly, because even the larger experiments performed poorly, it could be that the sites in Araçatuba are inherently more variable or there were greater difficulties encountered in carrying out the operations for the experiment. To improve experiments carried out in this region will require more uniform areas and/or more careful performance of the operations during the experiment.