

**UNIVERSIDADE FEDERAL DE SANTA MARIA
CAMPUS FREDERICO WESTPHALEN
PROGRAMA DE PÓS-GRADUAÇÃO EM AGRONOMIA:
AGRICULTURA E AMBIENTE**

Tiago Olivoto

**VIÉS ASSOCIADO AO ARRANJO DE DADOS E TAMANHO
AMOSTRAL E SUAS IMPLICAÇÕES NA ACURÁCIA DA SELEÇÃO
INDIRETA NO MELHORAMENTO DE PLANTAS**

Frederico Westphalen, RS

2017

Tiago Olivoto

**VIÉS ASSOCIADO AO ARRANJO DE DADOS E TAMANHO AMOSTRAL E SUAS
IMPLICAÇÕES NA ACURÁCIA DA SELEÇÃO INDIRETA NO MELHORAMENTO
DE PLANTAS**

Dissertação apresentada ao Curso de Pós-Graduação em Agronomia: Agricultura e Ambiente, da Universidade Federal de Santa Maria (UFSM, RS), como requisito parcial para obtenção do título de **Mestre em Agronomia**.

Orientador: Prof. Dr. Velci Queiróz de Souza

Frederico Westphalen, RS

2017

Olivoto, Tiago

VIÉS ASSOCIADO AO ARRANJO DE DADOS E TAMANHO AMOSTRAL
E SUAS IMPLICAÇÕES NA ACURÁCIA DA SELEÇÃO INDIRETA NO
MELHORAMENTO DE PLANTAS / Tiago Olivoto.- 2017.

125 p.; 30 cm

Orientador: Velci Queiróz de Souza

Coorientadores: Denise Schmidt, Braulio Otomar Caron,
Maicon Nardino

Dissertação (mestrado) - Universidade Federal de Santa
Maria, Campus de Frederico Westphalen, Programa de Pós-
Graduação em Agronomia - Agricultura e Ambiente, RS, 2017

1. Coeficiente de correlação 2. Multicolinearidade 3.
Simulações 4. Zea mays L. 5. Intervalo de confiança I.
Queiróz de Souza, Velci II. Schmidt, Denise III. Otomar
Caron, Braulio IV. Nardino, Maicon V. Título.

© 2017

Todos os direitos autorais reservados a Tiago Olivoto. A reprodução de partes ou do todo deste trabalho só poderá ser feita mediante a citação da fonte.

E-mail: tiagoolivoto@gmail.com

Tiago Olivoto

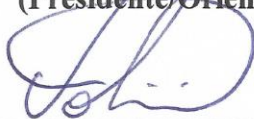
**VIÉS ASSOCIADO AO ARRANJO DE DADOS E TAMANHO AMOSTRAL E SUAS
IMPLICAÇÕES NA ACURÁCIA DA SELEÇÃO INDIRETA NO MELHORAMENTO
DE PLANTAS**

Dissertação apresentada ao Curso de Pós-Graduação em Agronomia: Agricultura e Ambiente, da Universidade Federal de Santa Maria (UFSM, RS), como requisito parcial para obtenção do título de **Mestre em Agronomia**.

Aprovado em 20 de fevereiro de 2017



**Velci Queiróz de Souza, Dr. (UNIPAMPA)
(Presidente/Orientador)**



Volmir Sergio Marchioro, Dr. (UFSM)



Marcos Vinícius Marques Pinheiro, Dr. (UFSM)

Frederico Westphalen, RS

2017

DEDICATÓRIA

Aos meus pais Luiz Antônio Olivoto e Sidnei Salette Carniel Olivoto, meu irmão Luiz Gustavo Olivoto, pelos ensinamentos de vida, exemplo de caráter e amor incondicional, dedico-lhes este trabalho.

AGRADECIMENTOS

Agradeço primeiramente à Deus, pela oportunidade da existência, pela saúde e paz e por ter iluminado os caminhos que me trouxeram até aqui.

Aos meus pais Luiz Antônio Olivoto e Sidnei Salette Carniel Olivoto, meu irmão Luiz Gustavo Olivoto, pelo amor incondicional, pelos exemplos de humildade e ética, pelos ensinamentos e conselhos e pela compreensão nos momentos de ausência.

Ao meu orientador Dr. Velci Queiróz de Souza, pela oportunidade de integrar seu grupo de pesquisa, pela orientação e pela confiança depositada em mim. Obrigado pela amizade verdadeira, pelos conselhos, ensinamentos e por ser além de um professor, um educador formador de pessoas.

A todos os colegas do Laboratório de Melhoramento Genético e Produção de Plantas, pelo companheirismo e amizade, e pela ajuda na realização das avaliações desta e demais pesquisas.

Aos amigos e membros da banca examinadora, Dr. Marcos Vinícius Marques Pinheiro e Dr. Volmir Sergio Marchioro, pela valiosa colaboração na redação do manuscrito final.

À Comissão de Aperfeiçoamento de Pessoal do Nível Superior, pela concessão da bolsa de Mestrado.

À Universidade Federal de Santa Maria, Campus Frederico Westphalen pela estrutura física e humana disponibilizada.

Ao programa de Pós-Graduação em Agronomia: Agricultura e Ambiente, pela possibilidade de cursar o Mestrado.

À Amanda Baseggio e Jaksson Ferreira Klin, pela valiosa colaboração na condução dos experimentos de campo e coleta dos dados que originou esta Dissertação de Mestrado.

A todos que direta ou indiretamente contribuíram para meu crescimento pessoal e profissional, possibilitando que este sonho se tornasse realidade.

Meu sincero MUITO OBRIGADO!

O sucesso nasce do querer, da determinação e persistência em se chegar a um objetivo. Mesmo não atingindo o alvo, quem busca e vence obstáculos, no mínimo fará coisas admiráveis.

(José Martiniano de Alencar)

RESUMO

VIÉS ASSOCIADO AO ARRANJO DE DADOS E TAMANHO AMOSTRAL E SUAS IMPLICAÇÕES NA ACURÁCIA DA SELEÇÃO INDIRETA NO MELHORAMENTO DE PLANTAS

AUTOR: Tiago Olivoto

ORIENTADOR: Velci Queiróz de Souza

Alguns métodos de arranjo de dados utilizados atualmente podem superestimar os coeficientes de correlação de Pearson (r) entre variáveis explicativas, aumentando a multicolinearidade em análises que utilizam regressão múltipla. Neste sentido, os objetivos da presente pesquisa foram revelar o impacto de diferentes cenários de arranjos de dados na multicolinearidade de matrizes, na eficiência dos métodos utilizados para ajustá-la, nas estimativas dos coeficientes e acurácia da análise de trilha, bem como fazer uso de simulações para revelar o comportamento estatístico do r e o tamanho amostral ótimo para estimativas de r entre caracteres do milho. Para isto, foram utilizados dados de um experimento conduzido em delineamento de blocos completos casualizados em esquema fatorial 15×3 (15 híbridos simples de milho e três locais), dispostos em quatro repetições. As variáveis analisadas em cinco plantas de cada parcela foram: altura de planta, altura de inserção da espiga, diâmetro e comprimento da espiga, número de fileiras de grãos por espiga, número de grãos por fileira, diâmetro e comprimento do sabugo, relação diâmetro do sabugo/diâmetro da espiga, número de grãos por espiga, massa de grãos por espiga e massa de mil grãos. Em um primeiro momento, três métodos de análise de trilha (tradicional, com inclusão de k e com exclusão de variáveis) tendo como variável dependente a massa de grãos por espiga, foram testados em dois cenários: 1) com a matriz de correlação linear ($X'X$) entre as variáveis estimada com todas as observações amostradas, $n = 900$ e 2) com a matriz $X'X$ estimada com o valor médio das cinco plantas amostradas em cada parcela, $n = 180$. Posteriormente, visando avaliar o comportamento estatístico do r , além dos dois cenários descritos, o valor médio dos tratamentos em cada local, $n = 45$, também foi considerado. Em cada cenário foram simulados 60 tamanhos amostrais utilizando simulações bootstrap com reposição. Intervalos de confiança para combinações de diferentes magnitudes foram estimados em cada cenário e tamanho amostral. Cento e oitenta matrizes de correlação (três cenários \times 60 tamanhos amostrais) foram estimadas e a multicolinearidade avaliada. O número de grãos por espiga e a massa de mil grãos apresentam os efeitos diretos mais expressivos sob a massa de grãos por espiga ($r = 0,892$ e $r = 0,733$, respectivamente). A utilização de valores oriundos de médias reduz a variância individual de um conjunto de n -variáveis, superestima a magnitude do r entre os pares de combinação, aumenta a multicolinearidade da matriz e reduz a eficiência dos métodos utilizados para ajustá-la, bem como a acurácia das estimativas dos coeficientes de trilha. O número de plantas necessário para estimativa de coeficientes de correlação com intervalo de confiança bootstrap de 95% é maior quando todas as observações da amostra são utilizadas e aumenta no sentido de pares de combinação com menor magnitude. Utilizando todas as observações amostradas, 210 plantas são suficientes para estimativa do r entre caracteres de híbridos simples de milho, no intervalo de confiança “bootstrap” de $95\% < 0,30$. Um método simples para reduzir a multicolinearidade das matrizes e melhorar a acurácia da análise de trilha é proposto.

Palavras-chave: *Zea mays* L. Coeficiente de correlação. Multicolinearidade. Simulações.

ABSTRACT

BIAS ASSOCIATED WITH DATA ARRANGEMENT AND SAMPLE SIZE AND ITS IMPLICATIONS ON THE ACCURACY OF INDIRECT SELECTION IN PLANT BREEDING

AUTHOR: Tiago Olivoto
ADVISOR: Velci Queiróz de Souza

Some data arrangement methods currently used may overestimate Pearson correlation coefficient (r) among explanatory traits, increasing multicollinearity in analysis that uses multiple regression. In this sense, the aims of the present research were to reveal the impact of different data arrangement scenarios on the multicollinearity of matrices, on the efficiency of the used methods to adjust it, on the estimates of coefficients and accuracy of the path analysis, as well as to use simulations to reveal the statistical behavior of the r and the optimal sample size for estimating r between maize traits. For this, data from an experiment conducted in a randomized complete design in a 15×3 factorial scheme (15 maize hybrids \times three growing sites), arranged in four replicates were used. The traits analyzed in five plants of each plot were: plant height, ear insertion height, diameter and length of ear, number of rows per ear, number of kernels per row, diameter and length of cob, cob diameter/ear diameter ratio, number of kernels per ear, kernel mass per ear and thousand-kernel weight. At first, three path analysis methods (traditional, with k inclusion and with the exclusion of traits) having as a dependent trait the kernel mass per ear were tested in two scenarios: 1) with the linear correlation matrix ($X'X$) between the traits estimated with all sampled observations, $n = 900$ and 2) with the $X'X$ matrix estimated with the average value of the five sampled plants in each plot, $n = 180$. Subsequently, aiming to evaluate the statistical behavior of r , in addition to the two described scenarios, the average value of treatments at each site, $n = 45$, was also considered. In each scenario, 60 sample sizes were simulated by using bootstrap simulations with replacement. Confidence intervals for combinations of different magnitudes were estimated in each scenario and sample size. One hundred and eighty correlation matrices (three scenarios \times 60 sample sizes) were estimated and the multicollinearity evaluated. The number of kernels per ear and the thousand-kernel weight presented the most expressive direct effects to kernel mass per ear ($r = 0.892$ and $r = 0.733$, respectively). The use of average values reduces the individual variance of a set of n -traits, overestimates the magnitude of the r between the trait pairs, increases the multicollinearity of the matrix, and reduces the effectiveness of the used methods to adjust it as well as the accuracy of the path coefficient estimates. The number of plants required to estimate correlation coefficients with a 95% bootstrap confidence interval is greater when all sampled observations are used and increases in the sense of combination pairs with lower magnitude. By using all sampled observations, 210 plants are sufficient to estimate r between traits of simple maize hybrids in the 95% bootstrap confidence interval < 0.30 . A simple method that reduces the multicollinearity of matrices and improves the accuracy of path analysis is proposed.

Key words: *Zea mays* L. Correlation coefficient. Multicollinearity. Simulations.

LISTA DE ILUSTRAÇÕES

ARTIGO II

Figure 1 - β values for plant height (PH), ear height (EH), ear length (EL), ear diameter (ED), number of rows per ear (NRE), number of kernels per row (NKR), cob diameter (CD), cob length (CL), total number of kernels per plant (TNK), cob diameter/ear diameter ratio (CD/ED), and thousand-kernel weight (TKW), obtained with 21 k values, where β estimated with $k = 0$, matches to the estimations of least squares. Estimations performed for ASO (a) and AVP (b) scenarios.....86

ARTIGO III

Figure 1 - Descriptive analysis of 1000 bootstrap estimates of Pearson's correlation coefficient. Symbols represent the maximum values, percentile 97.5%, average, percentile 2.5% and minimum, obtained for the pair of traits plant height \times ear height estimated in ASO (a) in AVP (b) and AVT (c) scenarios; number of kernels row \times ear diameter estimated in ASO (d), AVP (e) and AVT (f) scenarios and cob diameter / ear diameter ratio \times ear length, estimated in ASO (g), AVP (h) and AVT (i) scenarios.....105

Figure 2 - Amplitude of the correlation coefficient for the confidence interval of 95%. (a) ASO scenario. (b) AVP scenario and (c) AVT scenario. Lines in greyscale represent the pair cob diameter / ear diameter ratio \times ear length (CD/ED \times EL), number of kernels rows \times ear diameter (NKR \times ED) and plant height \times ear height (PH \times EH).....106

Figure 3 - Distribution of average values of correlation coefficient in ASO \times AVT scenarios combination. Columns represent the observed values. Black and gray lines represent the normal distribution and Kernel density estimation, respectively. ASO and AVT scenarios represent the correlation coefficients estimated by all sampled observations, and by the average values of treatments, respectively. In the lower plot, the average (rhombus), the median (vertical line), the distance between the 25th and 75th percentiles (length of the box) and the maximum and minimum values (outer spread) of the estimated correlation coefficient are presented for each scenario.....107

Figure 4 - Distribution of average values of correlation coefficient in ASO \times AVP scenarios combination. Columns represent the observed values. Black and gray lines represent the normal distribution and Kernel density estimation, respectively. ASO and AVP scenarios represent the correlation coefficients estimated by all sampled observations, and by the average values of plots, respectively. In the lower plot, the average (rhombus), the median (vertical line), the distance between the 25th and 75th percentiles (length of the box) and the maximum and minimum values (outer spread) of the estimated correlation coefficient are presented for each scenario.....108

Figure 5 - Distribution of average values of correlation coefficient in AVP × AVT scenarios combination. Black and gray lines represent the normal distribution and Kernel density estimation, respectively. AVP and AVT scenarios represent the correlation coefficients estimated by average values of plots and treatments, respectively. In the lower plot, the average (rhombus), the median (vertical line), the distance between the 25th and 75th percentiles (length of the box) and the maximum and minimum values (outer spread) of the estimated correlation coefficient are presented for each scenario.....109

Figure 6 - Descriptive analysis of correlation coefficients of 55 trait pairs estimated in 60 sample sizes by 1000 bootstrap simulations. Scenarios represent the original data coming from all sampled observations (ASO), coming from average values of each plot (AVP) and coming from average values of treatments (AVT). The rhombus within the box represents the average in the scenario. The horizontal line within the box represents the median value. The length of the box is the distance between the 25th and 75th percentiles. Outer spread represents the maximum and minimum values.....110

Figure 7- Condition number of correlation's matrices among explanatory traits estimated with 60 different sample sizes. For each sample size, the traits' values were estimated by average of 1000 bootstrap simulations of the original data coming from all sampled observations (ASO), coming from average values each plot (AVP) and coming from average values of treatments (AVT).....111

LISTA DE TABELAS

ARTIGO I

Table 1 - Multiple Coefficient of Determination (R^2) and the noise observed in 25 studies involving path analysis.....	46
---	----

ARTIGO II

Table 1 - Descriptive analysis for the 11 explanatory traits estimated in the two data arrangement scenarios.....	79
---	----

Table 2 - Correlation and covariance matrices among 11 explanatory traits obtained with all sampled observations, $n = 900$ (upper diagonal) and with the average values of each plot, $n = 180$ (below diagonal).....	80
--	----

Table 3 - Multicollinearity diagnosis for Pearson product-moment correlation matrices among 11 explanatory traits estimated in the two data arrangement scenarios.....	81
--	----

Table 4 - Eigenvalues and components of the eigenvectors of Pearson product-moment correlation matrix among the 11 explanatory traits estimated with all sampled observations, $n = 900$	82
--	----

Table 5 - Eigenvalues and component of the eigenvectors of Pearson product-moment correlation matrix among the 11 explanatory traits estimated with the average values of each plot, $n = 180$	83
--	----

Table 6 - Multicollinearity diagnosis of Pearson product-moment correlation matrices among the 11 explanatory traits estimated in two data arrangement scenarios and three path analysis methodologies.....	84
---	----

Table 7 - Direct effects for the 11 explanatory traits on kernel weight per ear with the regression estimators estimated in two data arrangement scenarios and three path analysis methodologies.....	85
---	----

ARTIGO III

Table 1 – t -statistics for the average correlation coefficient (r) of 55 traits combination estimated in 60 different numbers of plants. Average values represent 1000 bootstrap simulations of the original data coming from all sampled observations (ASO), the average of each plot (AVP) and the average of treatments (AVT). Coefficients in bold indicate the combinations in which r was lower with the use of averages.....	103
--	-----

SUMÁRIO

1	INTRODUÇÃO	15
1.1	PROBLEMÁTICA.....	15
1.2	HIPÓTESES.....	16
1.3	OBJETIVOS	17
1.3.1	Objetivo geral	17
1.3.2	Objetivos específicos	17
1.4	JUSTIFICATIVA.....	17
2	ARTIGO I - PEARSON CORRELATION COEFFICIENTS AND ACCURACY OF PATH ANALYSIS USED IN MAIZE BREEDING: A CRITICAL REVIEW	19
2.1	ABSTRACT.....	21
2.2	INTRODUCTION	21
2.2.1	Maize crop	21
2.2.2	Maize breeding	22
2.2.3	Biometric models used in maize hybrids	25
2.2.4	Path analysis conception	26
2.2.5	Estimation of linear correlation	27
2.2.6	Path analysis estimation	28
2.2.7	Difficulties observed in path analysis	29
2.2.8	Matrices multicollinearity	30
2.2.8.1	<i>What is it?</i>	30
2.2.8.2	<i>Methods for adjusting multicollinearity</i>	31
2.2.8.3	<i>Can multicollinearity be reduced?</i>	33
2.2.9	A theoretical explanation	34
2.2.10	Path analysis accuracy in ecological experiments	35
2.2.11	Future perspectives	36
2.3	FINAL CONSIDERATIONS	36
2.4	REFERENCES.....	37
3	ARTIGO II - MULTICOLLINEARITY IN PATH ANALYSIS: A SIMPLE METHOD TO REDUCE ITS EFFECTS	47
3.1	ABSTRACT.....	48
3.2	INTRODUCTION	49
3.3	MATERIALS AND METHODS.....	53
3.3.1	Material and experimental design	53
3.3.2	Assessed traits	54
3.3.3	Data analysis	55
3.3.4	Correlation and covariance matrices	55
3.3.5	Multicollinearity diagnosis	56
3.3.6	Methods for adjusting multicollinearity	58
3.3.6.1	<i>Determining which traits should be excluded from the model</i>	58
3.3.6.2	<i>Including the k constant into correlation matrices</i>	58
3.3.7	Direct and indirect effects	59
3.4	RESULTS	62
3.4.1	Descriptive statistics	62
3.4.2	Correlation and covariance matrices	62
3.4.3	Multicollinearity Diagnosis	63
3.4.4	Multicollinearity-generating traits	64
3.4.5	Multicollinearity determination after adjustment	64

3.4.6	Direct effects and accuracy	65
3.5	DISCUSSION.....	67
3.5.1	Correlation coefficients estimated with data average are overestimated.....	67
3.5.2	Preserving individual variances, multicollinearity is reduced	67
3.5.3	Data arrangement change the efficiency of methods to adjusting multicollinearity	69
3.5.4	Data based on averages reduce direct effects and increase the noise in path analysis	70
3.6	CONCLUSIONS	72
3.7	ACKNOWLEDGMENTS	72
3.8	REFERENCES	73
4	ARTIGO III - OPTIMAL SAMPLE SIZE AND DATA ARRANGEMENT METHOD IN ESTIMATING CORRELATION MATRICES WITH LESSER COLLINEARITY: A STATISTICAL FOCUS IN MAIZE BREEDING	87
4.1	ABSTRACT	89
4.2	INTRODUCTION	90
4.3	MATERIALS AND METHODS	91
4.3.1	Site description and experimental design.....	91
4.3.2	Accessed traits.....	92
4.3.3	Statistical procedures	93
4.3.3.1	<i>Bootstrap simulations</i>	93
4.3.3.2	<i>Descriptive analysis of correlation coefficients</i>	94
4.3.3.3	<i>t-test to compare the correlation coefficient among the scenarios</i>	94
4.3.3.4	<i>Diagnosis of multicollinearity in the scenarios</i>	95
4.4	RESULTS	95
4.4.1	Statistical properties of the correlation coefficient.....	95
4.4.2	Comparison of correlation pairs between the scenarios	96
4.4.3	Multicollinearity	97
4.5	DISCUSSION.....	98
4.6	CONCLUSION	100
4.7	ACKNOWLEDGMENT	100
4.8	REFERENCES	101
5	DISCUSSÃO GERAL	112
6	CONCLUSÃO GERAL	115
	REFERÊNCIAS	116
	APÊNDICE A – ANÁLISE DESCRITIVA DAS VARIÁVEIS ANALISADAS EM CADA CENÁRIO DE ARRANJO DE DADOS.	118
	APÊNDICE B – EFEITOS INDIRETOS ESTIMADOS EM DIFERENTES CENÁRIOS E MÉTODOS DE ANÁLISE DE TRILHA.	119

1 INTRODUÇÃO

Modelos biométricos desempenham um importante papel no melhoramento genético vegetal atual, aumentando a eficiência e reduzindo o tempo de seleção por meio da seleção indireta. Para realizar a seleção indireta o melhorista precisa compreender o sentido e o grau de associação entre os caracteres avaliados de uma determinada espécie. Para isto, o coeficiente de correlação produto-momento de Pearson (PEARSON, 1920), vem sendo amplamente utilizado.

Embora a correlação de Pearson revele o sentido e o grau de associação linear entre um par de caracteres, esta ferramenta não revela associações de causa e efeito. Assim, Sewall Wright em seu trabalho publicado com o título ‘Correlation and causation’ (WRIGHT, 1921), propôs um método conhecido como análise de trilha ou ‘Path analysis’ permitindo esta compreensão. O método é baseado no particionamento do coeficiente de correlação linear em efeitos diretos e indiretos de um grupo de variáveis consideradas preditoras ou explicativas, na resposta de uma variável dependente ou principal.

Na cultura do milho, bem como em diversas culturas de importância mundial, trabalhos utilizando análise de trilha têm obtido sucesso no sentido de revelar as inter-relações entre caracteres, sejam eles produtivos, de qualidade de grão ou de efeitos da interação do genótipo × ambiente/manejo de cultivo (ADESOJI; ABUBAKAR; LABE, 2015; JADHAV; KASHID; KULKARNI, 2014; MA et al., 2015; NARDINO et al., 2016).

1.1 PROBLEMÁTICA

Embora o modelo estatístico seja consolidado, alguns problemas de natureza inevitável são observados nas estimativas dos coeficientes de trilha. O principal entrave encontrado, principalmente por se tratar de um modelo de regressão múltipla, é a presença de multicolinearidade entre variáveis preditoras, ou seja, a alta correlação entre as variáveis preditoras incluídas no modelo. No decorrer de quase um século de utilização desta análise, diversos pesquisadores têm trabalhado com o intuito de ajustar a multicolinearidade em matrizes de variáveis explicativas, seja excluindo variáveis não aditivas ou modificando o modelo estatístico utilizado (ALIN, 2010; AUCOTT; GARTHWAITE; CURRALL, 2015; FARRAR; GLAUBER, 1967; GUNST; MASON, 1977; HUANG; JOU; CHO, 2015; KIERS; SMILDE, 2006; MANSFIELD; HELMS, 1982; YU; JIANG; LAND, 2015). Resultados satisfatórios vêm sendo observados, no entanto, a grande maioria dos estudos realizados têm o

foco principal no ajuste da multicolinearidade, ou seja, o que e como fazer para minimizar os efeitos danosos da multicolinearidade após a estimativa da matriz de correlação. São escassos na literatura, entretanto, trabalhos que abordam métodos para reduzi-la.

Como discutido, a multicolinearidade está diretamente associada a magnitude das correlações entre variáveis preditoras. Neste sentido, para estimativa da real correlação entre duas variáveis aleatórias (ex. X e Y), a covariância e o desvio padrão destas variáveis devem representar a população em estudo. Em experimentos agrônomicos, é comum mensurar os caracteres elencados de acordo com as hipóteses e objetivos do programa de melhoramento em diversos indivíduos (plantas) em cada parcela de cada tratamento, visando representar a população (tratamento) em estudo. Tais plantas rotineiramente compõem a média desta parcela, qual será posteriormente utilizada para estimativas de análise de variância e análises complementares. No entanto, são encontrados diversos estudos que fizeram uso destas médias para estimar os coeficientes de correlação e posteriormente os coeficientes de trilha (ADESOJI; ABUBAKAR; LABE, 2015; FARIA et al., 2015; KHAMENAE et al., 2012; KUMAR et al., 2015; NATARAJ; SHAHI; AGARWAL, 2014; NATARAJ; SHAHI; VANDANA, 2015; RIGON et al., 2012; TOEBE; CARGNELUTTI, 2013; TORRES et al., 2015). Partindo-se do pressuposto que a média mascara as variâncias individuais (das observações coletadas), correlações estimadas a partir destas médias não representarão a real variância e desvio padrão das variáveis mensuradas (X , Y , ..., Z) na população original.

Diante do exposto, na presente pesquisa serão abordados problemas conceituais e metodológicos evidenciados atualmente na estimativa dos coeficientes de correlação entre variáveis explicativas e qual o impacto da utilização de diferentes cenários de arranjo de dados na acurácia da análise de trilha, direcionada para a área do melhoramento genético do milho. Em adição a isto, tamanhos amostrais são estudados em diferentes cenários de arranjo de dados, visando avaliar o comportamento estatístico do coeficiente de correlação e o condicionamento das matrizes de correlação.

1.2 HIPÓTESES

Levando-se em consideração a premissa para estimativa do coeficiente de correlação abordada anteriormente e a observação de diversos trabalhos metodologicamente tendenciosos, as seguintes hipóteses foram formuladas.

- A utilização de valores médios reduz a variância do conjunto de n -variáveis, superestimando os coeficientes de correlação dos pares de combinação;

- Os métodos para ajuste de multicolinearidade conhecidos atualmente são mais eficazes quando aplicados em matrizes de correlação estimadas com os valores de todas as observações amostradas;

- A acurácia da análise de trilha é maior quando os coeficientes de trilha são estimados com matrizes de correlação estimadas com todas as observações da amostra;

- O aumento no tamanho amostral proporciona menor intervalo de confiança do coeficiente de correlação;

- Matrizes de correlação estimadas com todas as observações amostradas apresentam melhor condicionamento e menores problemas de multicolinearidade;

1.3 OBJETIVOS

1.3.1 Objetivo geral

As hipóteses formuladas fundamentaram o seguinte objetivo geral: fazer uso de modelos estatísticos e biométricos para analisar o comportamento do coeficiente de correlação, a multicolinearidade das matrizes e a acurácia da análise de trilha em diferentes cenários de arranjos de dados.

1.3.2 Objetivos específicos

Propor um método de arranjo de dados que reduza os níveis de multicolinearidade, melhore a eficiência dos métodos conhecidos para ajustá-la, bem como aumente a acurácia das estimativas dos coeficientes de trilha.

Revelar quais os impactos de diferentes cenários de arranjo de dados e tamanhos amostrais no comportamento estatístico dos coeficientes de correlação e no condicionamento das matrizes.

1.4 JUSTIFICATIVA

A realização da presente pesquisa é justificada pela ampla utilização e importância mundial do coeficiente de correlação de Pearson e da análise de trilha, apoiada pela inexistência, até onde se sabe, de estudos que avaliaram o impacto do arranjo de dados e tamanhos amostrais

nas estimativas dos coeficientes de correlação e condicionamento de matrizes de variáveis consideradas explicativas.

Espera-se que as hipóteses sejam comprovadas e, se assim evidenciado, os resultados pioneiros poderão contribuir para uma melhor acurácia desta análise. A aplicabilidade de tais resultados será ampla, principalmente na área de melhoramento genético vegetal, contudo, outras áreas da ciência que fazem uso desta análise, também poderão ser beneficiadas.

**2 ARTIGO I - PEARSON CORRELATION COEFFICIENTS AND ACCURACY OF
PATH ANALYSIS USED IN MAIZE BREEDING: A CRITICAL REVIEW**

Submetido para o periódico: International Journal of Current Research

Situação: Publicado

URL: <http://www.journalcra.com/article/pearson-correlation-coefficients-and-accuracy-path-analysis-used-maize-breeding-critical-rev>

**PEARSON CORRELATION COEFFICIENTS AND ACCURACY OF PATH
ANALYSIS USED IN MAIZE BREEDING: A CRITICAL REVIEW**

¹, *Tiago Olivoto, ²Maicon Nardino, ³Ivan Ricardo Carvalho, ⁴Diego Nicolau Follmann, ⁵Vinicius Szareski, ³Mauricio Ferrari, ³Alan Junior Pelegrin, ⁶Velci Queiróz de Souza.

¹ Department of Agronomic and Environmental Sciences, Federal University of Santa Maria Frederico Westphalen, Rio Grande do Sul, Brazil.

² Department of Mathematics and Statistics, Federal University of Pelotas, Capão do Leão, Rio Grande do Sul, Brazil.

³ Plant Genomics and Breeding Center, Federal University of Pelotas, Capão do Leão, Rio Grande do Sul, Brazil.

⁴ Agronomy Department, Federal University of Santa Maria, Santa Maria, Rio Grande do Sul, Brazil.

⁵ Dept. of Crop Science, Federal University of Pelotas, Capão do Leão, Rio Grande do Sul, Brazil.

⁶ Federal University of Pampa, Dom Pedrito, Rio Grande do Sul, Brazil.

*Corresponding author: Tiago Olivoto

Department of Agronomic and Environmental Sciences, Federal University of Santa Maria Frederico Westphalen, Rio Grande do Sul, Brazil.

Email address <tiagoolivoto@gmail.com>

2.1 ABSTRACT

Maize (*Zea mays* L.) has been the subject of several studies involving correlation coefficient estimates and path analysis. This critical review discusses some systematic errors that have been observed in estimating of correlation coefficients and its possible impacts on accuracy of path analysis. In a first moment, an approach about the maize crop, origin, characteristics and biometric models commonly used in genetic breeding of this crop is presented. Some obstacles found in estimates of path coefficients and the methods used to adjust them are discussed. We also present evidences and a theoretical explanation that some data arrangement methods currently used may be overestimating the correlation coefficients in scientific studies. Data from a literature search revealing the accuracy of path analysis of some research are presented and discussed. In a last moment, we present a future perspective about how the correct estimate of the correlation coefficients may improve the accuracy of path analysis, underscoring the need for research directed to this subject.

Key-words: *Zea mays*. Average data. Correlation matrices. Systematic errors.

2.2 INTRODUCTION

2.2.1 Maize crop

Maize (*Zea mays* L.), belonging to the grass family Poaceae is the most produced cereal in the world, surpassing the mark of 1 billion tons produced in 2016 growing season. This crop has great economic, social and recently environmental importance due their grain serve as alternative raw material for ethanol production (Hertel et al., 2010). The world's leading producers of this cereal is the United States, China and Brazil.

It is known that the currently known maize is the result of a long evolutionary process, being the most accepted hypothesis that it evolved from the teosinte, with the center of origin in Central America, specifically in Mexico (Gaut and Doebley, 1997).

During this evolutionary process, genetic events were decisive for the change in plant architecture and crop's inflorescence characteristics. Two quantitative traits loci (QTLs) were identified as the main responsible for the morphological differences between these species. The first (TB1) located on arm of chromosome 1L has effects on the gender of the inflorescence and the number and length of internodes on the lateral branches; the second, located on arm of chromosome 3L, affects the same characteristics. A study evaluating the segregation of these loci revealed that they present epistatic interrelationships turning together, substantially, the plant architecture and inflorescence (Doebley et al., 1995).

The number of chromosomes present in the modern maize is 10, but it has long been suspected that this number was the result of a historical tetraploid event. Several observations point to this possibility, including the fact that the culture has duplicated chromosome segments (Gaut, 2001). Some of these segments were sequenced and the standard divergence between 14 pairs of duplicated genes was examined. The results indicated that the time in this sequences' duplication vary in two distinct groups, corresponding to about 20.5, and 11.4 million years ago. (Gaut and Doebley, 1997). This observation indicates the possibility of an allotetraploid genomic event where his two diploid progenitors diverged about 20.5 million years ago, and that the allotetraploid event probably occurred approximately 11.4 million years ago.

2.2.2 Maize breeding

It is attributed to Darwin the first works with plant pollination, however, were East and Shull the pioneers in the study of the influence of successive self-pollination and exploitation of heterosis in maize. During the era of hybrid maize (1908 to present), the crop yield has increased almost six times (Lee and Tollenaar, 2007).

In early 1908, George Harrison Shull, published a paper with the title “The composition of the field of maize”, marking the beginning of the exploitation of heterosis in plant breeding,

certainly one of the greatest genetic triumphs of our time. In his work, Shull showed that inbred lines of maize, subjected to several cycles of self-pollination showed significant reduction in vigor and grain yield; however, the hybrids resulting from two inbred lines had these features recovered, often featuring performance and superior vigor of varieties from which the inbred lines were derived (Shull, 1908).

At the same time, Edward Murray East made similar experiments and also recognized the deleterious effects of inbreeding in maize plants; however, did not realize the value of crossing inbred lines, up to study Shull's paper. East was not convinced of the usefulness of the idea, because, really, inbred lines produced a very small amount of seeds, burdening any increase in production provided by hybrids. Both were at odds, but have remained true to their findings (Crow, 1998).

The limitation in seeds' production was surpassed later (1918) from an idea of Donald Forsha Jones, who while still a graduate student, defended the idea of using four genetic bases or double-cross hybrids. The principle involved crossing two inbred lines and later, crossing of this hybrid with other, resulting from two other inbred lines. These hybrids were somewhat more variable compared with simple hybrids, however, much less than the open-pollinated varieties existing at that time. As seeds were coming from a simple hybrid, the largest quantity of available seed improved the program viability (Jones, 1918).

Increases in maize productivity was, no doubt, largely due to the discovery of heterotic effect; however, the evolution of agricultural practices, such as increased use of fertilizers, changes in plant's arrangement, cultivation practices and agricultural mechanization, were useful tools and that combined with the use of higher-genetically plants enabled the achievement of high yields currently observed. But, it would be possible to separate the contribution of these effects? Studies evaluating the productivity of maize in a period of 70 yrs.

showed an average increase of 65-75 kg ha⁻¹ yr⁻¹, and that genetic breeding was responsible for about 50% of this increase (Duvick, 1977, 2005).

A maize ideotype had been proposed by Mock and Pearce (1975). The ideotype that should produce optimally when grown in an environment without limitations of edaphoclimatic factors, high plant density and reduced spacing between rows, it is characterized by a) rigid vertically-oriented leaves above ear (leaves below the ear should be horizontally-oriented); b) maximum photosynthetic efficiency; c) efficient conversion of assimilates in grains; d) short interval between pollination and the emergence of style-stigmas; e) prolificacy; f) small size of cobs; g) insensitivity to photoperiod; h) cold tolerance in the germination (for cultivated genotypes in areas where early sowing takes place in cold or wet soil); i) as long as possible grain filling; and j) slow leaves senescence.

In this regard, studies aiming at a higher-plant architecture (Tian et al., 2011), better floral sync (Buckler et al., 2009), improved photosynthetic efficiency (Fracheboud et al., 1999) and absorption of nutrients (Gallais and Hirel, 2004) has been successful. The combination of all the favorable characteristics in a single hybrid, however, is a daunting task for breeders mainly due, in most part, the traits be expressed by different genic actions (Sa et al., 2014).

Success in maize breeding, as well as in others economically important crops also was due to wider use of statistic-experimental models in the selection of superior hybrid, introduced by Fisher, involving replication, randomization and local control. The author states the importance of a thorough selection in a plant breeding program. In the case of simple maize hybrids this process occurs in three steps. 1) choice of individuals in a population to start the process; 2) artificial self-pollination of these individuals aiming to inbreeding and selection of pure lines and 3) artificial crosses. If plants are randomly selected in each step, the hybrids will be a random sample of the original population. Thus, the criteria-based selection in the three steps should be considered. At first, the selection resembles the mass selection, practiced in

breeding of open-pollinated varieties. In the second, the selection is neutralized quickly by rapid fixation, due to homozygosity increase in 50% each generation; so, Fisher emphasized that the selection in the last step, should have greater emphasis. In fact, the selection at this step is important as it is being practiced in the studied subject (Fisher, 1925).

2.2.3 Biometric models used in maize hybrids

Several statistical models have been used to evaluate the performance of maize hybrids. Models that allow the partition of genotype-vs-environment interaction into environmental and genetic components are useful to evaluate the adaptability and stability of hybrids, especially in the assessment of value for cultivation and use. Mixed models with fixed and random variance components also have proven also efficient to identify promising hybrids in breeding programs (Baretta et al., 2016).

Knowledge of association degree between traits is of fundamental importance in plant breeding programs. This importance increases, especially if some desirable trait present difficulty in assessment or low heritability (Cruz et al., 2014). The Pearson product-moment correlation coefficient (Pearson, 1920), has been widely used for this purpose. Although this correlation reveals the direction and degree of linear association between a pair of traits, it does not reveal interrelationships of cause and effect. Thus, Sewall Wright in his work entitled “Correlation and causation” (Wright, 1921) proposed a method known as “path analysis” allowing this understanding. The method is based on the partitioning of the linear correlation coefficient into direct and indirect effects of a group of explanatory traits on the response of a dependent trait.

Path analysis has been highlighted in breeding area because the selection aiming to improve a desirable trait that has difficulty-measure and low heritability, can be indirectly

carried out by another trait, directly associated with the desirable trait, but that shows high heritability and easy assessment.

In maize, as well as in several world-important crops, studies using path analysis has been successful in the sense of revealing the interrelationships between traits, be them yielding, grain quality or the effects of interaction genotype-vs-environment or management of cultivation (Adesoji et al., 2015, 2015; Jadhav et al., 2014; Ma et al., 2015; Nardino et al., 2016). In summary, the results converge to a common conclusion: the number of kernels per ear and thousand-kernel weight are the traits with greater direct association with grain yield (Adesoji et al., 2015; Khameneh et al., 2012; Mohammadi et al., 2003; Reddy et al., 2012). As the heritability in the broad sense of these traits is high ($h^2 > 0.90$), the indirect selection from these traits aiming at increasing grain yield (trait highly influenced by the environment) can be effective (Ojo et al., 2006).

2.2.4 Path analysis conception

Path analysis is originally based on ideas developed by Sewall Wright (Wright, 1921), however, from its conception to the method's consolidation, some disagreement about the reliability of the mathematical method were observed. In 1922, Henry E. Niles, in his paper entitled "Correlation, causation and Wright's theory of path coefficients", made a criticism of the method proposed by Wright, claiming that the philosophical basis of the path coefficients method was doubtful. Niles, testing Wright's method had observed in some of its results correlations exceeding | 1 |, saying "these results are ridiculous" and that Wright would have to provide much more convincing evidence than he was presenting (Niles, 1922).

In the following year (1923), Sewall Wright in his paper entitled "The theory of path coefficients: a reply to Niles' criticism", consolidates his method concluding that Niles seemed to be based on incorrect mathematical concepts, result of a failure to recognize that path

coefficient it is not a symmetric function of two traits, but it necessarily has direction. Wright concludes his work by stating that the path analysis does not provide a formula to infer causal relationships from knowledge of the correlations; it is, however, within certain limitations, a method of evaluating the logical consequences of a causal hypothesis relationship in a system of correlated traits. It adds yet that the criticism offered by Niles nothing invalidates the theory or application of path coefficient (Wright, 1923). Currently, the statistical method of path coefficient is consolidated and worldwide used in several areas of science.

To estimate path coefficients, normal equations models are used to partition the linear coefficients into direct and indirect effects of a set of explanatory traits on a dependent trait. Thus, their estimates need a previously-estimated linear correlation matrix among traits under study.

2.2.5 Estimation of linear correlation

One of the most used measures in breeding to estimate the direction and degree of linear association between two random traits is the Pearson product-moment correlation coefficient. To estimate the degree of association between two hypothetical traits X and Y , let's consider the following assumption. The traits should form the following dataset. $(X_1, Y_1), (X_2, Y_2) \dots (X_n, Y_n)$. Thus, correlation coefficient estimates between X and Y is obtained by the following

equation:
$$r = \frac{\sum_{i=1}^n [(X_i - \bar{X})(Y_i - \bar{Y})]}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}}$$

where $\sum_{i=1}^n [(X_i - \bar{X})(Y_i - \bar{Y})]$ is the covariance XY ; $\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}$

is the product of standard deviation of X and Y , respectively, being $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ and

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i .$$

Although the merit of this analysis had been attributed to Karl Pearson, the method was originally designed by Francis Galton, who defined the term correlation as to the following: “two variables are said to be co-related when the variation of the one is accompanied on the average by more or less variation of the other, and in the same direction” (Galton, 1888). So, your estimate considers the covariance between two traits, represented here by XY divided by the product of respective standard deviation of X and Y .

Considering the premise of this analysis, the traits which will be correlated, will have, mandatorily, be assessed in the same subject, in order to represent the actual covariance and standard deviation of the set of observations.

2.2.6 Path analysis estimation

After obtaining linear correlation estimates (r), partitioning of linear correlations into direct and indirect effects of an explanatory dataset with p -traits can be performed by derivation of the set of normal equations ($X'X\beta = X'Y$) to estimate parameters of multiple regression using OLS (*Ordinary Last Squares*). Thus, β estimate is given by $\beta = X'X^{-1} X'Y$, where β is the partial regression coefficient ($\beta_1, \beta_2, \beta_3, \dots, \beta_p$) to $p + 1$ rows; $X'X^{-1}$ is the inverse of linear correlation matrix among explanatory traits; and $X'Y$ is the correlation matrix between each explanatory trait with the dependent trait.

After estimating the regression coefficients (β_p), the direct and indirect effects of a set of p -explanatory trait towards the dependent trait can be estimated. Consider the following example, where a set of explanatory traits (a, b, c and d) are used to explain the relationship of cause and effect on the response of dependent variable (y). After partial regression estimations ($\beta_1, \beta_2, \beta_3$ and β_4), direct and indirect effects of ‘ a ’ on ‘ y ’ are given by: $r_{a:y} = \beta_1 + \beta_2 r_{a:b} + \beta_3 r_{a:c} + \beta_4 r_{a:d}$ where $r_{a:y}$ is the linear correlation between ‘ a ’ and ‘ y ’; β_1 is the direct effect of ‘ a ’ on ‘ y ’; $\beta_2 r_{a:b}$ is the indirect effect of ‘ a ’ on ‘ y ’ via ‘ b ’; $\beta_3 r_{a:c}$ is the indirect effect of ‘ a ’ on ‘ y ’ via ‘ c ’; and

$\beta_{a:d}$ is the indirect effect of ‘*a*’ on ‘*y*’ via ‘*d*’. Similar equations are used to estimate direct and indirect effects of *b*, *c*, and *d*. The coefficient of determination of the model, i.e., how much of the variance in the dependent trait is explained by the interrelationship on explanatory traits, is given by $R^2 = \beta_1 r_{a:y} + \beta_2 r_{b:y} + \beta_3 r_{c:y} + \beta_4 r_{d:y}$. Residual effect is estimated by $\text{Noise} = \sqrt{1 - R^2}$.

This technique has facilitated the understanding of the interrelationship among traits and their effects on dependent trait in several areas of science, as in plant breeding and crop management (Abdala et al., 2016; Dewey and Lu, 1959; Farooq et al., 2015; Mohammadi et al., 2016; Nardino et al., 2016; Olivoto et al., 2015; Souza et al., 2015), animal breeding (Norris et al., 2015; Önder and Abaci, 2015), environmental and social sciences (Hong et al., 2016; Xu et al., 2014), humanities (Hagger et al., 2016) and several related areas.

Indeed, path analysis has been a useful tool particularly in plant breeding, however, care must be taken prior the estimation of this analysis. Below we discuss some obstacles encountered in the estimates of the path coefficients.

2.2.7 Difficulties observed in path analysis

Although this analysis shows associations of cause and effect, its estimate is based on multiple regression principles; thus, it can be biased by complex nature of the data, where the response of the dependent trait is linked to many explanatory traits that are often correlated with each other (Graham, 2003). Correlated traits are difficult to analyze because its effects on dependent trait may be due to any synergistic relationship between traits or spurious correlations. Thus, where two explanatory traits are highly associated, it is difficult to estimate the relationship of each individual explanatory trait, since these, as a whole contribute to the explanation of the linear relationship. This particularity is known as multicollinearity (Blalock, 1963).

2.2.8 Matrices multicollinearity

2.2.8.1 What is it?

In multiple linear regression, data is fitted to a multiple linear model that predicts the values of a response trait (Y) from the weighted sum of several explanatory traits (X_i) and the random error (ε) $Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \dots + \beta_iX_i + \varepsilon$; where β are regression coefficients.

The main goal is to fit a model using the smallest number of traits that explain the most variance of the response variable. If all explanatory traits are independent, each of the regression coefficients (β_i) represent the total contribution of a given predictor in response trait; if however, two or more explanatory traits are associated, partial regression coefficients need to be estimated to isolate the contribution of a single explanatory trait. The distinction between single contributions is the crucial point in multiple regression analysis being also the largest inferential problem found due to the presence of multicollinearity (Graham, 2003; Gunst and Mason, 1977).

When this phenomenon occurs in moderate or severe levels, the variances associated with path estimators can reach too high values, making unreliable estimates. Montgomery et al (2012), proposed a classification for multicollinearity based on the condition number (CN), i.e. the ratio between the largest and smallest eigenvalue of explanatory traits matrix ($CN = \lambda_{Max}/\lambda_{Min}$). Thus, the degree of multicollinearity is considered weak, moderate and severe when $CN \leq 10$ between 10 and 100 and ≥ 1000 , respectively.

Other indicator used to identify the presence of multicollinearity is called variance inflation factor (VIF), and demonstrated the extent of the effects of other independent traits on the variance of the selected independent trait [$VIF = 1 / (1 - R_i^2)$]. For each of β coefficients in a multiple regression model there is one VIF. When the VIF for a given predictor is 1, it means that there is no correlation between this predictor and the remainder of explanatory traits. This fact is hardly observed. Can be taking as a rule, that the existence of VIFs greater than 10, are

serious multicollinearity signals, being necessary to take some action to adjust it (Mansfield and Helms, 1982; O'Brien, 2007).

Path coefficients at odds with biological expectation were observed when the analysis was performed in the presence of severe multicollinearity (Toebe and Cargnelutti, 2013). In addition, a study by Petraitis et al. (1996) revealed that from 24 path analysis published in ecological studies, 15 had problems with multicollinearity, resulting in 13 cases with biased path coefficients. This information is worrying because in the case of plant breeding, path coefficients wrongly-estimated and interpreted, may result in an inefficient selection, brought into play the financial, human and time spent in the conduct of a plant breeding program.

2.2.8.2 Methods for adjusting multicollinearity

Although the problems related to multicollinearity presents itself as a difficulty in estimating path coefficients, some steps can be taken to mitigate its undesirable effects when it is detected by the aforementioned methods.

It is now known that the exclusion of the traits responsible for inflating the variance of a regression coefficient is an effective technique that reduces the multicollinearity in matrices of explanatory traits (Jadhav et al., 2014). The identification of these traits, however, can become a difficult task. As previously discussed, the purpose of multiple regression (path analysis) is to identify a set of explanatory traits with high explanatory power, but which do not exhibit highly correlated. In this sense, there are several variable selection methods to choose a subset of predictors with minimal multicollinearity, such as hierarchical models, stepwise procedures and criteria-based models (George and McCulloch, 1993; Mitchell and Beauchamp, 1988; Nishii, 1984; Wold et al., 1984).

In a focused approach to plant breeding, Cruz et al. (2014) discuss a method to identify the traits responsible for the multicollinearity in a set of explanatory traits. This method is based

on analysis of eigenvalues and eigenvectors of a symmetric positive definite matrix of explanatory traits and identifies the traits responsible for this problem, as that with the highest weight (component of the eigenvector) associated to the eigenvalues of lesser magnitude.

The exclusion of traits responsible for multicollinearity allowed estimating path coefficients, without its harmful effect, in research with several crops such as rice (Shrivastava and Sharma, 1976), canola (Coimbra et al., 1999), soybean (Bizeti et al., 2004) and maize (Toebe and Cargnelutti, 2013). It should be noted that the choice of traits for exclusion must be careful, because traits with high explanatory power removed from the model, can reduce the coefficient of determination (R^2), and increase the noise's model (Cruz et al., 2014). When the exclusion of multicollinearity-generating traits is not a procedure considered by researcher, e.g., due to a small number of explanatory traits, or the importance of knowing their effects, a second option is to perform the path analysis with all the explanatory traits, but with the addition of a small value in diagonal elements of $X'X$, known as ridge regression (Hoerl and Kennard, 1970). This method aims to reduce the variance associated with the OLS (*Ordinary Last Squares*) estimators. Thus, β estimates in ridge regression are obtained similarly to the conventional method, however solving the partially-modified normal equations system $(X'X+k)\beta = X'Y$ generating $\beta = (X'X+k)^{-1} X'Y$, for $0 < k < 1$. Where, β is the partial regression coefficient ($\beta_1, \beta_2, \beta_3, \dots, \beta_p$) to $p + 1$ rows; $(X'X+k)^{-1}$ is the inverse of linear correlation among explanatory traits with k constant included in diagonal elements; and $X'Y$ is the correlation matrix between each explanatory traits with the dependent trait.

Using numerical examples to illustrate the effectiveness of this method, Marquardt (1970), concluded that the ridge regression method is efficient in estimating path analysis coefficients from non-orthogonal data. In plant breeding, this technique also has been proven effective in improving the conditioning of explanatory traits matrices in studies with several

economically-important crops (Bizeti et al., 2004; Coimbra et al., 1999; Luz et al., 2011; Nardino et al., 2016; Nogueira et al., 2012; Olivoto et al., 2015; Souza et al., 2015).

2.2.8.3 *Can multicollinearity be reduced?*

Although the techniques for adjusting multicollinearity have been effective and widely known, such techniques are used after the diagnosis of the correlation matrix among explanatory traits, that is, its use is only possible after the estimation of linear correlation matrix.

As previously discussed, multicollinearity is directly associated with the high magnitude of correlation between explanatory traits in the model. In this sense, to estimate the actual correlation between two random traits (X and Y), the covariance and standard deviation should represent the population under study. In agronomic experiments, it is common assessing several samples (plants) in each plot of each treatment, to represent the population (treatment). Such plants routinely make up an average of this specific plot, which will be used later for ANOVA and supplementary analysis, such as multiple comparison analysis.

In a bibliographic research project were found, however, several studies that has been using these averages to estimate the correlation coefficients and then the path coefficients (Adesoji et al., 2015; Faria et al., 2015; Khameneh et al., 2012; Kumar and Babu, 2015; Nataraj et al., 2015, 2014; Rigon et al., 2012; Toebe and Cargnelutti, 2013; Torres et al., 2015). Starting from the assumption that the average can mask the individual variances (of assessed plants), correlations estimated from these averages do not represent the actual variance and standard deviation of the traits (X , Y , ..., Z) of the original population.

In addition to the statistical concept methodologically biased, the inference of magnitude and direction of interrelationships between traits when the correlation is estimated with average data is misleading, because this inference is performed in a different population of the original (e.g. when all plants are used for this estimate). As a large number of agronomic

studies makes populational inferences based on sampling (plants), using average value of these plants to estimate correlations and make an inference to the original population is a misconception that, without a doubt, should be considered.

2.2.9 A theoretical explanation

We take as an example an experiment to evaluate the direction and degree of association between traits of maize hybrids, conducted in a randomized complete block design with five treatments (simple hybrid) and four replications. In each replication (plot) is common to assess traits in several plants, aiming to represent the population of this specific plot. In experiments with maize hybrids are usually sampled 3 to 5 plants per plot, mainly because they present low phenotypic variation. So, in this hypothetical experiment, we assume that in five plants of each plot were evaluated three traits (X , Y , and Z). The researcher would then have the values of these three traits assessed in 100 plants (5 hybrids \times 4 replications \times 5 plants). To estimate the correlation between X and Y , e.g., the following dataset is required: $(X_1, Y_1), (X_2, Y_2), \dots, (X_{100}, Y_{100})$. The correlation coefficient is then given by applying the formulae described in “estimation of linear correlation”.

When the researcher uses the average values of plots in order to estimate the correlation, he is masking the deviations of each trait (X , Y and Z) relative to the overall average of these traits. In this case, the observed deviations among the five plants of each plot will be canceled out by the average of these plants. The new data set used for the same estimation of the correlation between X and Y in this methodology will be then: $(X_1, Y_1), (X_2, Y_2), \dots, (X_{20}, Y_{20})$. The observed variance in the new dataset is representing then the variance of average from five original sampled plants, and not the variance coming from all sampled plants; therefore, this variance is masked and tends to present lower itself, compared to the original variance. This

fact should be considered because the inference of the direction and magnitude of association between traits is being made for a population with variance different than the original.

After an in-depth evaluation of the correlation's formula (see estimation of linear correlation), it is noted that the formula's divisor is estimated by the product of the standard deviations of X and Y . Then, when the correlation is estimated based on average data, generally showing less variation, the product of these deviations will be lesser. Assuming that the covariance between X and Y remain similar, dividing by a smallest divider, will result in an overestimated coefficient of correlation. But, could this mistake found in the correlation estimates be associated with higher multicollinearity problems and with the reduction of accuracy in path analysis? This approach, as far as we known, is still limited in the literature.

2.2.10 Path analysis accuracy in ecological experiments

In a randomized research of 25 studies using path analysis, we observed a certain contradiction regarding information of the coefficient of determination (R^2) and model's noise. For example, only five studies (20%) clearly showed the R^2 and the noise in their results. In four studies (16%), only the R^2 was presented, while in six studies (24%) only the noise was presented. In 10 studies (40%), neither of these parameters were found. This is alarming because it can mask the interpretation of the reader in not to know how much of the variation in the dependent trait was explained by the model.

In studies that showed both adjustment measures, were observed R^2 fluctuating between 0.31 and 0.99 and noises ranging from 0.105 to 0.680. It is also observed that in some cases, the noise approached of the R^2 , a fact that may cast doubt on the reliability of the estimated path coefficients (Table 1).

2.2.11 Future perspectives

Research aiming to demonstrate if and how much the use of average values may overestimate the correlation coefficients, increase multicollinearity in analysis that uses multiple regression and reduces its accuracy are necessary and certainly will be welcome. Thus, by combining the correct estimate of correlation coefficients with the known methods to adjust multicollinearity, the accuracy of path analysis in biological studies could be increased. It is noteworthy that, completely eliminate the multicollinearity in matrices of explanatory traits is an almost impossible task, because the degree of interrelationship coming from the nature of the traits is inevitable. In this context, studies adopting a sequential path analysis model with first-, second-, n-order predictors might be considered to determine the interrelationships among traits with smallest problems of multicollinearity (Mohammadi et al., 2003).

From a breeding viewpoint, the effectiveness of indirect selection based on path coefficients will depend then of: (i) researcher's ability to correctly estimating correlation coefficients; (ii) take the right steps to adjust multicollinearity of their matrices; (iii) include in group of predictors, traits that explain most of the observed variance in the dependent trait; and (iv) carry out the selection based on traits with high heritability and directly associated with the response of the dependent trait.

2.3 FINAL CONSIDERATIONS

Path analysis has been helping researchers from several areas of science to reveal logical relationships of cause and effect. In maize genetic breeding, in particular, this technique has allowed the knowledge of the interrelationships between traits, enabling faster-indirect selection of lines in inbreeding process. The methods currently used for adjusting the multicollinearity of explanatory traits matrices are effective. Observation of studies with correlation coefficients tendentiously-estimated and studies in which have been hidden

important information such as coefficient of determination and model's noise, however, is worrying. In this sense, research aiming to compare the influence of average values on estimates of correlation coefficients and its impact on path analysis accuracy are needed and could help researchers reduce the systematic errors in their experiments.

2.4 REFERENCES

- Abdala, A.J., Bokosi, J.M., Mwangwela, A.M., and Mzengeza, T.R. 2016. Correlation and path co-efficient analysis for grain quality traits in F1 generation of rice (*Oryza sativa* L.). *Journal of Plant Breeding and Crop Science*, 8:109–116.
- Adesoji, A.G., Abubakar, I.U., and Labe, D.A. 2015. Character Association and Path Coefficient Analysis of Maize (*Zea mays* L.) Grown under Incorporated Legumes and Nitrogen. *Journal of Agronomy*, 14:158–163.
- Agrama, H.A.S. 1996. Sequential path analysis of grain yield and its components in maize. *Plant Breeding*, 115:343–346.
- Alvi, M.B., Rafique, M., Tariq, M.S., Hussain, A., Mahmood, T., and Sarwar, M. 2003. Character association and path coefficient analysis of grain yield and yield components maize (*Zea mays* L.). *Pakistan Journal of Biological Sciences*, 6:136–138.
- Bárbaro, I.M., Centurion, M.A.P.D.C., Di Mauro, A.O., Unêda-Trevisoli, S.H., Arriel, N.H.C., and Costa, M.M. 2006. Path analysis and expected response in indirect selection for grain yield in soybean. *Crop Breeding and Applied Biotechnology*, 6:151–159.
- Baretta, D., Nardino, M., Carvalho, I.R., Oliveira, A.C. de, Souza, V.Q. de, and Maia, L.C. da, 2016. Performance of maize genotypes of Rio Grande do Sul using mixed models. *Científica*, 44:403-411.
- Bello, O.B., Abdulmalik, S.Y., Afolabi, M.S., and Ige, S.A. 2010. Correlation and path coefficient analysis of yield and agronomic characters among open pollinated maize

varieties and their F₁ hybrids in a diallel cross. *African Journal of Biotechnology* 9:2633–2639.

Bizeti, H.S., Carvalho, C.G.P. de, Souza, J.R.P. de, and Destro, D. 2004. Path analysis under multicollinearity in soybean. *Brazilian Archives of Biology and Technology* 47:669–676.

Blalock, H.M., 1963. Correlated Independent Variables: The Problem of Multicollinearity. *Social Forces*, 42:233–237.

Buckler, E.S., Holland, J.B., Bradbury, P.J., Acharya, C.B., Brown, P.J., Browne, C., Ersoz, E., Flint-Garcia, S., Garcia, A., Glaubitz, J.C., Goodman, M.M., Harjes, C., Guill, K., Kroon, D.E., Larsson, S., Lepak, N.K., Li, H., Mitchell, S.E., Pressoir, G., Peiffer, J.A., Rosas, M.O., Rocheford, T.R., Romay, M.C., Romero, S., Salvo, S., Villeda, H.S., Silva, H.S. da, Sun, Q., Tian, F., Upadaya, N., Ware, D., Yates, H., Yu, J., Zhang, Z., Kresovich, S., and McMullen, M.D. 2009. The Genetic Architecture of Maize Flowering Time. *Science*, 325:714–718.

Carvalho, C.G.P. de, 2001. Path analysis under multicollinearity in So × So maize hybrids. *Crop Breeding and Applied Biotechnology*, 1:263–269.

Coimbra, J.L.M., Guidolin, A.F., Carvalho, F.I.F., Coimbra, S.M.M., and Marchioro, V.S. 1999. Análise de trilha I: análise do rendimento de grãos e seus componentes. *Ciência Rural*, 29:213–218.

Crow, J.F. 1998. 90 Years Ago: The Beginning of Hybrid Maize. *Genetics* 148:923–928.

Cruz, C.D., Carneiro, P.C.S. and Regazzi, A.J., 2014. Modelos Biométricos Aplicados ao Melhoramento Genético, 3rd ed. UFV, Viçosa, MG.

Dewey, D.R., and Lu, K., 1959. A correlation and path-coefficient analysis of components of crested wheatgrass seed production. *Agronomy Journal* 51:515–518.

- Diz, D.A., Wofford, D.S., and Schank, S.C., 1994. Correlation and path-coefficient analyses of seed-yield components in pearl millet \times elephantgrass hybrids. *Theoretical and Applied Genetics*, 89:112–115.
- Doebley, J., Stec, A., and Gustus, C. 1995. teosinte branched1 and the origin of maize: evidence for epistasis and the evolution of dominance. *Genetics*, 141:333–346.
- Duvick, D.N. 1977. Genetic rates of gain in hybrid maize yields during the past 40 years. *Maydica*, 22:187–196.
- Duvick, D.N. 2005. Genetic progress in yield of United States maize (*Zea mays* L.). *Maydica*, 50:193-202.
- Faria, L.A., Peluzio, J.M., Afféri, F.S., Carvalho, E.V., Dotto, M.A., and Faria, E.A. 2015. Análise de trilha para crescimento e rendimento de genótipos de milho sob diferentes doses nitrogenadas. *Journal of Bioenergy and Food Science*, 2:1–11.
- Farooq, J., Anwar, M., Rizwan, M., Riaz, M., Mahmood, K., and Mahpara, S. 2015. Estimation of correlation and path analysis of various yield and related parameters in cotton (*Gossypium hirsutum* L.). *Cotton Genomics and Genetics*, 6:1–6.
- Fisher, R.A., 1925. Statistical methods for research workers, 1st ed. Oliver and Boyd, London.
- Fracheboud, Y., Haldimann, P., Leipner, J., and Stamp, P. 1999. Chlorophyll fluorescence as a selection tool for cold tolerance of photosynthesis in maize (*Zea mays* L.). *Journal of Experimental Botany*, 50:1533–1540.
- Gallais, A., and Hirel, B. 2004. An approach to the genetics of nitrogen use efficiency in maize. *Journal of Experimental Botany*, 55:295–306.
- Galton, F. 1888. Co-relations and their measurement, chiefly from anthropometric data. *Proceedings of the Royal Society of London*, 45:135–145.
- Gaut, B.S. 2001. Patterns of chromosomal duplication in maize and their implications for comparative maps of the grasses. *Genome Research*, 11:55–66.

- Gaut, B.S., and Doebley, J.F. 1997. DNA sequence evidence for the segmental allotetraploid origin of maize. *Proceedings of the National Academy of Sciences*, 94:6809–6814.
- George, E.I., and McCulloch, R.E. 1993. Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88:881–889.
- Graham, M.H., 2003. Confronting Multicollinearity in Ecological Multiple Regression. *Ecology*, 84:2809–2815.
- Gunst, R.F., and Mason, R.L. 1977. Advantages of examining multicollinearities in regression analysis. *Biometrics*, 33:249–260.
- Hagger, M.S., Chan, D.K.C., Protogerou, C., and Chatzisarantis, N.L.D. 2016. Using meta-analytic path analysis to test theoretical predictions in health behavior: An illustration based on meta-analyses of the theory of planned behavior. *Preventive Medicine* 89:154–161.
- Hertel, T.W., Golub, A.A., Jones, A.D., O’Hare, M., Plevin, R.J., and Kammen, D.M. 2010. Effects of us maize ethanol on global land use and greenhouse gas emissions: estimating market-mediated responses. *BioScience*, 60:223–231.
- Hoerl, A.E., and Kennard, R.W. 1970. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 12:55–67.
- Hong, J., Shen, Q., and Xue, F. 2016. A multi-regional structural path analysis of the energy supply chain in China’s construction industry. *Energy Policy*, 92:56–68.
- Jadhav, N.H., Kashid, D.N., and Kulkarni, S.R. 2014. Subset selection in multiple linear regression in the presence of outlier and multicollinearity. *Statistical Methodology*, 19:44–59.
- Iqbal, S., Mahmood, T., Tahira, M.A., Anwar, M., Sarwar, and M. 2003. Path coefficient analysis in different genotypes of soybean (*Glycine max* (L) Merrill). *Pakistan Journal of Biological Science*, 6:1085–1087.

- Jones, D.F. 1918. The effect of inbreeding and crossbreeding upon development. *Proceedings of the National Academy of Sciences of the United States of America*, 4:246–250.
- Khameneh, M.M., Bahraminejad, S., Sadeghi, F., Honarmand, S.J., and Maniee, M. 2012. Path analysis and multivariate factorial analyses for determining interrelationships between grain yield and related characters in maize hybrids. *African Journal of Agricultural Research*, 7:6437–6446.
- Khan, N., and Naqvi, F.N. 2012. Correlation and path coefficient analysis in wheat genotypes under irrigated and non-irrigated conditions. *Asian Journal of Agricultural Sciences*, 4:346–351.
- Kumar, S.V.V., and Babu, D.R., 2015. Character association and path analysis of grain yield and yield components in Maize (*Zea Mays* L.). *Electronic Journal of Plant Breeding*, 6:550–554.
- Kumar, T.S., Reddy, D.M., Reddy, K.H., and Sudhakar, P. 2011. Targeting of traits through assessment of interrelationship and path analysis between yield and yield components for grain yield improvement in single cross hybrids of maize (*Zea mays* L.). *International Journal of Applied Biology and Pharmaceutical Technology*, 2:123–129.
- Kumar, K.V., Sudarshan, M.R., Dangi, K.S., and Reddy, S.M. 2013. Character association and path coefficient analysis for seed yield in quality protein maize *Zea mays* L. *Journal of Research ANGRAU*, 41:153–157.
- Lee, E.A., and Tollenaar, M. 2007. Physiological basis of successful breeding strategies for maize grain yield. *Crop Science*, 47:202–215.
- Luz, L.N. da, Santos, R.C. dos, Filho, M., and Albuquerque, P. 2011. Correlations and path analysis of peanut traits associated with the peg. *Crop Breeding and Applied Biotechnology*, 11:88–95.

- Ma, Z., Qin, Y., Wang, Y., Zhao, X., Zhang, F., Tang, J., and Fu, Z., 2015. Proteomic analysis of silk viability in maize inbred lines and their corresponding hybrids. *PLoS One*, 10: e0144050.
- Mansfield, E.R., and Helms, B.P., 1982. Detecting multicollinearity. *The American Statistician*, 36:158–160.
- Marquardt, D.W. 1970. generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. *Technometrics*, 12:591–612.
- Mitchell, T.J., and Beauchamp, J.J. 1988. Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83:1023–1032.
- Mock, J.J., and Pearce, R.B., 1975. An ideotype of maize. *Euphytica* 24:613–623.
- Mohammadi, R., Farshadfar, E., and Amri, A. 2016. Path analysis of genotype \times environment interactions in rainfed durum wheat. *Plant Production Science*, 19:43–50.
- Mohammadi, S.A., Prasanna, B.M., and Singh, N.N. 2003. Sequential path model for determining interrelationships among grain yield and related characters in maize. *Crop Science*, 43:1690–1697.
- Montgomery, D.C., Peck, E.A. and Vining, G. 2012. Introduction to linear regression analysis. 5th ed. John Wiley & Sons, Hoboken, NJ.
- Nardino, M., Souza, V.Q. de, Baretta, D., Konflanz, V.A., Carvalho, I.R., Follmann, D.N., and Caron, B.O., 2016. Association of secondary traits with yield in maize F₁'s. *Ciência Rural*, 46:776–782.
- Nataraj, V., Shahi, J.P., and Agarwal, V., 2014. Correlation and path analysis in certain inbred genotypes of maize (*Zea Mays* L.) at Varanasi. *International Journal of Innovative Research and Development*, 3:14–17.
- Nataraj, V., Shahi, J.P., and Vandana, D. 2015. Character association and path analyses in maize (*Zea mays* L.). *Environment and Ecology*, 33:78–81.

- Niles, H.E. 1922. Correlation, causation and Wright's theory of "path coefficients." *Genetics*, 7:258–273.
- Nishii, R. 1984. Asymptotic properties of criteria for selection of variables in multiple regression. *Annals of Statistics*, 12:758–765.
- Nogueira, A.P.O., Sedyama, T., Sousa, L.B. de, Hamawaki, O.T., Cruz, C.D., Pereira, D.G., and Matsuo, É., 2012. Análise de trilha e correlações entre caracteres em soja cultivada em duas épocas de semeadura. *Bioscience Journal*, 28:877-888.
- Norris, D., Brown, D., Moela, A.K., Selolo, T.C., Mabelebele, M., Ngambi, J.W., and Tyasi, T.L. 2015. Path coefficient and path analysis of body weight and biometric traits in indigenous goats. *Indian Journal of Animal Research*, 49:573–578.
- O'brien, R.M., 2007. A caution regarding rules of thumb for variance inflation factors. *Qual Quant*, 41:673–690.
- Ojo, D.K., Omikunle, O.A., Oduwaye, O.A., Ajala, M.O., and Ogunbayo, S.A., 2006. Heritability, character correlation and path coefficient analysis among six inbred-lines of maize (*Zea mays* L.). *World Journal of Agricultural Sciences*, 2:352–358.
- Olivoto, T., Souza, V.Q. de, Carvalho, I.R.C., Nardino, M., and Follmann, D.N. 2015. Análise de trilha para caracteres relacionados ao crescimento de mudas de pepineiro. *Enciclopédia Biosfera*, 11:69–80.
- Önder, H., and Abaci, S.H. 2015. Path analysis for body measurements on body weight of saanen kids. *Kafkas Üniversitesi Veteriner Fakültesi Dergisi*, 21:351–354.
- Pearson, K. 1920. Notes on the history of correlation. *Biometrika*, 13:25–45.
- Petratis, P.S., Dunham, A.E., and Niewiarowski, P.H. 1996. Inferring multiple causality: the limitations of path analysis. *Functional Ecology*, 10:421–431.

- Reddy, V.R., Jabeen, F., Sudarshan, M.R., and Rao, A.S. 2012. Studies on genetic variability, heritability, correlation and path analysis in maize (*Zea mays* L.) over locations. *International Journal of Applied Biology and Pharmaceutical Technology*, 4:196–199.
- Rigon, J.P.G., Capuani, S., Brito Neto, J.F. de, Rosa, G.M. da, Wastowski, A.D., and Rigon, C.A.G. 2012. Dissimilaridade genética e análise de trilha de cultivares de soja avaliada por meio de descritores quantitativos. *Revista Ceres*, 59:233–240.
- Sa, K.J., Park, J.Y., Woo, S.Y., Ramekar, R.V., Jang, C.-S., and Lee, J.K. 2014. Mapping of QTL traits in maize using a RIL population derived from a cross of dent maize × waxy maize. *Genes & Genomics*, 37:1–14.
- Saleem, U.S., Subhani, G.M., Ahmad, N., Rahim, M., and Ali, M.A. 2007. Correlation and path coefficient analysis in maize (*Zea mays* L.). *Journal of Agriculture Research* 45:177–183.
- Shrivastava, M., and Sharma, K. 1976. Analysis of path coefficients in rice. *Zeitschrift fuer Pflanzenzuechtung*, 77:174–177.
- Shull, G.H. 1908. The composition of a field of maize. *Journal of Heredity*, 4:296–301.
- Souza, V.Q. de, Bellé, R., Ferrari, M., de Pelegrin, A.J., Caron, B.O., Nardino, M., Follmann, D.N., and Carvalho, I.R. 2015. Componentes de rendimento em combinações de fungicidas e inseticidas e análise de trilha em soja. *Global Science and Technology*, 8:167-176.
- Tian, F., Bradbury, P.J., Brown, P.J., Hung, H., Sun, Q., Flint-Garcia, S., Rocheford, T.R., McMullen, M.D., Holland, J.B., and Buckler, E.S. 2011. Genome-wide association study of leaf architecture in the maize nested association mapping population. *Nature Genetics*, 43:159–162.
- Toebe, M., and Cargnelutti, A. 2013. Multicollinearity in path analysis of maize (*Zea mays* L.). *Journal of Cereal Science*, 57:453–462

- Torres, F.E., Teodoro, P.E., Ribeiro, L.P., Correa, C.C.G., Hernandez, F.B., Fernandes, R.L., Gomes, A.C., and Lopes, K.V. 2015. Correlations and path analysis on oil content of castor genotypes. *Bioscience Journal*, 31:1363–1369.
- Wold, S., Ruhe, A., Wold, H., and Dunn, W.J. 1984. The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. *SIAM Journal on Scientific and Statistical Computing*, 5:735–743.
- Wright, S. 1921. Correlation and causation. *Journal of Agricultural Research*, 20:557–585.
- Wright, S. 1923. The theory of path coefficients a reply to Niles's criticism. *Genetics*, 8:239–255.
- Xu, L., Lin, T., Xu, Y., Xiao, L., Ye, Z., and Cui, S. 2014. Path analysis of factors influencing household solid waste generation: a case study of Xiamen Island, China. *Journal of Material Cycles and Waste Management*, 18:377–384.

Table 1. Multiple Coefficient of Determination (R^2) and the noise observed in 25 studies involving path analysis.

Species	R^2	Residual	Reference
Castor beans	0.89	np	Torres et al., 2015
Cotton	np†	np	Farooq et al., 2015
<i>Indigenous goats</i>	np	np	Norris et al., 2015
Maize	np	0.345	Adesoji et al., 2015
Maize	np	np	Agrama, H. 1996
Maize	np	np	Alvi et al., 2003
Maize	np	np	Bello et al., 2010
Maize	np	0.560-0.670	Carvalho et al., 2001
Maize	0.555	0.667	Faria et al., 2015
Maize	0.31-0.99	np	Khameneh et al., 2012
Maize	np	0.249	Kumar et al., 2011
Maize	np	0.105	Kumar et al., 2013
Maize	0.851	0.386	Kumar et al., 2015
Maize	np	0.372	Nataraj et al., 2014
Maize	np	np	Nataraj et al., 2015
Maize	0.64	0.53	Rigon et al., 2012
Maize	np	np	Saleem et al., 2007
Maize	0.74‡	0.490‡	Toebe and Cargneluti, 2013
Peanut	np	np	Luz et al., 2011
Pearl millet × Elephantgrass	np	np	Diz et al., 1994
Rice	0.915	np	Abdala et al., 2016
Soybean	0.912	0.295	Bárbaro et al., 2006
Soybean	0.909-0.950	np	Bizeti et al., 2004
Soybean	np	np	Iqbal et al., 2003
Wheat	np	0.470-0.680	Khan and Naqvi, 2012

† np, not presented.

‡ Average from 14 path analysis.

**3 ARTIGO II - MULTICOLLINEARITY IN PATH ANALYSIS: A SIMPLE METHOD
TO REDUCE ITS EFFECTS**

Submetido para o periódico: Agronomy Journal

Situação: Publicado

DOI: 10.2134/agronj2016.04.0196

Multicollinearity in path analysis: a simple method to reduce its effects

Tiago Olivoto*, Velci Q. de Souza, Maicon Nardino, Ivan R. Carvalho, Maurício Ferrari,
Alan J. de Pelegrin, Vinicius J. Szareski and Denise Schmidt.

T. Olivoto and D. Schmidt, Dep. of Agronomy, Federal Univ. of Santa Maria, Frederico Westphalen, RS, Brazil; V.Q. de Souza, Federal Univ. of Pampa, Dom Pedrito, RS, Brazil; M. Nardino, I.R. Carvalho, M. Ferrari, A.J. de Pelegrin and V. J. Szareski, Plant Genomics and Breeding Center, Fed. Univ. of Pelotas, Pelotas, RS, Brazil. *Corresponding author (tiagoolivoto@gmail.com).

3.1 ABSTRACT

Some data arrangement methods often used may mask correlation coefficients among explanatory traits, increasing multicollinearity in multiple regression analysis. This study was performed to determine if the harmful effects of multicollinearity might be reduced in the estimation of the $\mathbf{X}'\mathbf{X}$ correlation matrix among explanatory traits. For this, data on 45 treatments (15 maize [*Zea mays* L.] hybrids sown in three places) were used. Three path analysis methods (traditional, with k inclusion, and traditional with trait exclusion) were tested in two scenarios: with $\mathbf{X}'\mathbf{X}$ matrix estimated with all sampled observations (ASO, $n = 900$) and with the $\mathbf{X}'\mathbf{X}$ matrix estimated with the average values of each plot (AVP, $n = 180$). The condition number (CN) was reduced from 3395 to 2004 when the matrix was estimated with all observations. On average, the factors that inflate the variance of regression coefficients were

increased by 61% in the AVP scenario. The addition of the k coefficient reduced the CN to 85.40 and 51.17 for the ASO and AVP scenarios, respectively. Exclusion of multicollinearity-generating traits was more effective in the ASO than the AVP scenario, resulting in CNs of 29.62 and 63.66, respectively. The largest coefficient of determination (0.977) and the smallest noise (0.150) were obtained in the ASO scenario after the exclusion of the multicollinearity-generating traits. The use of all sampled observations does not mask the individual variances and reduces the magnitude of the correlations among explanatory traits in 90% of cases, improving the accuracy of biological studies involving path analysis.

Abbreviations: ASO, All sampled observations; AVP, Average values of plot; CD, cob diameter; CD/ED, cob diameter/ear diameter ratio; CL, cob length; CN, condition number; ED, ear diameter; EH, ear height; EL, ear length; KWE, kernel weight per ear; MD, matrix determinant; NKR, number of kernels per row; NRE, number of rows per ear; PH, plant height TKW, thousand-kernel weight; TNK, total number of kernels per ear; VIF, variance inflator factor.

3.2 INTRODUCTION

In genetic breeding programs, understanding the sense and degree of association among traits has an important role in the development of selection strategies, which facilitate obtaining superior genotypes. One of the most used techniques to estimate these associations is the Pearson product-moment correlation, which is interpreted as the strength of the linear association between a pair of traits (Pearson, 1920). When more than two traits are considered, this measure by itself does not present the real sense and magnitude of the interrelationships,

making it impossible to determine if the associations are cause or effect (Aliyu et al., 2000). Therefore, path analysis is used when there is a dependent trait (of interest) and interrelations among explanatory traits. This method is based on ideas originally developed in biology (Wright, 1921, 1923, 1934) and economics (Wold, 1954) and enables the partitioning of the linear correlation coefficients into direct and indirect effects of several traits considered as explanatory toward a single dependent trait. In genetic plant breeding, this technique has proven very useful for revealing associations of cause and effect and providing help on indirect selection (Bello et al., 2010; Nardino et al., 2016).

Although path analysis presents the magnitude and sense of interrelations among explanatory traits toward a dependent trait, it is essentially based on the principles of multiple regression. When two or more alleged explanatory traits are highly correlated, it is hard to individually estimate the relations of each explanatory trait because they are associated and because they collectively contribute to explain linear relations. Such particularity is called *multicollinearity* (Blalock, 1963). When this problem is present at moderate or severe levels, the variance associated with estimators of path coefficients reach extremely high values, which makes such estimates untrustworthy, usually inconsistent with the biological expectation (Cruz et al., 2014).

Without the intention of observing such results, Shrivastava and Sharma (1976), performing studies related to the yield components of rice (*Oryza sativa* L.) yields, proved its negative direct effects on grain yields in the presence of multicollinearity. Theoretically, the yield components contribute positively to the grain yield; therefore, their results revealed that multicollinearity causes a bias in path coefficients. Illogical relationships in path coefficients obtained under multicollinearity ($-25.90 \leq \text{direct effect} \leq 21.5$) were also observed in maize

(*Zea mays* L.; Toebe and Cargnelutti, 2013). In addition, a study performed by Petraitis et al. (1996) reported that, from 24 path analysis studies published in ecological studies, 15 cases presented problems of multicollinearity, resulting in 13 cases with incorrectly estimated path coefficients.

The problems related to multicollinearity may be bypassed by excluding the nonadditive traits of the model. This technique depends on the prior diagnosis of the correlation matrix among explanatory traits, adopting procedures that, besides informing the degree of present multicollinearity, also identify the traits that are causing such problems (Mansfield and Helms, 1982; Montgomery et al., 2012). After excluding the multicollinearity-generating traits, path coefficients were estimated without the harmful effects of multicollinearity in several crops, such as rice (Shrivastava and Sharma, 1976), maize (Carvalho et al., 1999a), canola (*Brassica napus* L. ssp. *napus*; Coimbra et al., 2005), and soybean (*Glycine max* L. Merr.; Bizeti et al., 2004). When the exclusion of a trait is not a procedure considered by researchers (e.g., due to the reduced number of explanatory traits), path coefficients may be obtained with partially modified equations by the inclusion of a k constant in the diagonal elements of the $\mathbf{X}'\mathbf{X}$ correlation matrix (Cruz et al., 2012). This technique has been effective in studies with maize (Carvalho et al., 1999a), canola (Coimbra et al., 2005), peanut (*Arachis hypogaea* L.; da Luz et al., 2011), and bell pepper (*Capsicum annuum* L. var. *annuum*; Carvalho et al., 1999b).

The techniques used for adjusting the multicollinearity effects in path analysis are very well known. These methods, however, are applied only after the estimation of the correlation matrix among the explanatory traits (Cruz et al., 2012). In agronomic studies, the tradition is to assess different plants from each plot or, in other words, in each replicate; the traits of several plants are assessed, which routinely composes the average of the trait for this specific plot

(Vaux et al., 2012). In a bibliographic research project, it was observed that the path coefficients of 10 relevant studies were obtained with the average data of the observations of each plot (Silva et al., 2005; Khameneh et al., 2012; Rigon et al., 2012; Toebe and Cargnelutti, 2013; Nataraj et al., 2014, 2015; Adesoji et al., 2015; Faria et al., 2015; Kumar et al., 2015; Torres et al., 2015). After an in-depth study of these studies, the hypothesis was established that it would be possible to reduce the multicollinearity in matrices of explanatory traits by estimating correlation coefficients with data coming from all sampled plants in each plot. It is known that the multicollinearity arises because of high correlations between two or more explanatory traits and that linear correlation is the quotient between covariance XY and the product of the standard deviations of X and Y . Because average values tend to present the smallest standard deviation, the possibility of the correlation coefficient between X and Y getting the largest magnitude is high once the standard deviation is the divisor of the correlation's formula.

In this context, the following hypotheses were formulated: (i) the use of average values, at the plot level, suppresses individual variation and may increase correlations among explanatory traits; (ii) the harmful effects of multicollinearity in $\mathbf{X}'\mathbf{X}$ correlation matrices of explanatory traits are reduced when the values of all sampled observations are considered in estimations; and (iii) the measures to adjusting multicollinearity are more effective in correlation matrices estimated with the values of all sampled observations. These hypotheses motivated our research's aims: to evaluate the multicollinearity effects, and the effectiveness of the current methods for adjusting it, in $\mathbf{X}'\mathbf{X}$ correlation matrices of explanatory traits estimated in two scenarios: using the values of each sampled observation (ASO) and with the average values of each plot's observations (AVP).

3.3 MATERIALS AND METHODS

3.3.1 Material and experimental design

The experimental material used for this study consisted of simple maize hybrids. This specific crop was chosen due to its phenotypic stability and ease of trait assessment, thus reducing the likelihood of the occurrence of systematic and random errors. Fifteen commercial hybrids from five companies, which represent a large part of the Brazilian seed market, were used. Hybrids of each company were as follows: Pioneer P30F53H, P1630H, and P30B39; Biomatrix B2A525 HX, BM915 PRO, and 2B655 PW; Agrocere AG8690, AG8780, and AG9045; Syngenta Velox TL, Status TL, Truck TL, and SX7331; and Biogene BG7318H and BG7648H. The trials were performed in three cities in the Rio Grande do Sul state, Brazil, in the 2015 summer growing season: (i) Santo Expedito do Sul ($27^{\circ}56'S$, $51^{\circ}37'W$ at 728 m asl), with an average daily temperature of $24.5^{\circ}C$ and accumulated rainfall during the crop cycle of 823 mm; (ii) São José do Ouro ($27^{\circ}44'S$, $51^{\circ}32'W$ at 796 m asl), with an average daily temperature of $23.8^{\circ}C$ and accumulated rainfall of 958 mm; and (iii) Viadutos ($27^{\circ}33'S$, $52^{\circ}0'W$ at 628 m asl), with an average daily temperature of $25.2^{\circ}C$ and accumulated rainfall of 746 mm. All locations are within a 70-km radius, have a Haplustox soil, and were chosen due to similarities of soil and climatic characteristics. Thus, abiotic effects on the plants' response were minimized as much as possible.

Prior to the installation of the trials, each site was surveyed for potentially disruptive characteristics. To ensure uniformity inside the block and heterogeneity between the blocks, a randomized complete block design in a 15×3 factorial treatment design (15 simple maize hybrids \times 3 cropping fields) with four replicates was used, totaling 180 plots. Each plot was composed of six, 5-m-long cultivar rows, spaced at a 0.45 m. The hybrid seeds were manually

sown. After emergence and crop establishment, the plant density was adjusted to 70,000 plants ha⁻¹ for all hybrids. Soil management and cultural practices were the same for the three locations, following the phenological stages and needs of the crop. At the harvest stage, to avoid edge effects, only the two central rows were assessed. Data on 12 traits were assessed in five representative plants (observations) from each plot. Each plant and ear were labeled to assure that the traits (of plant and ear) were measured on the same subject; thus, a numerical sample identification system was implemented (place, hybrid, replicate, and plant).

3.3.2 Assessed traits

The plant height (PH) and ear height (EH) were measured (cm) from the surface of the ground to the flag leaf node and the supporting node of the largest ear at the harvest stage, respectively. The labeled ears were then evaluated in the laboratory. For each ear, the following traits were assessed: ear length (EL, cm), ear diameter (ED, cm), number of rows per ear (NRE), number of kernels per row (NKR), cob length (CL, cm), cob diameter (CD, mm), cob diameter/ear diameter ratio (CD/ED), total number of kernels per ear (TNK), kernel weight per ear (KWE, g), and thousand-kernel weight (TKW, g).

The lengths and diameters of ears and cobs were measured with a digital caliper. After counting the number of rows per ear and kernels per row, each ear was manually threshed, and the kernels were cleaned with pressurized air. Later, the kernel weight was measured with an analytical scale (AX 200, Marte Científica), and the total number of kernels was obtained with seed counter equipment (Pfeuffer). Last, the humidity of the kernels was assessed with a universal-humidity bench determiner (Comag), using 60 g of kernels, in a thickness of 0.575 inches. With these data, and with the humidity adjusted to 14% base humidity, it was possible

to determine the thousand-kernel weight of each ear, obtained by the following equation: $TKW = [(KWE/TNK) \times 1000]$. To maintain the actual variance of the sample, all procedures were carefully performed, one ear at a time, keeping sample traceability.

3.3.3 Data analysis

The dataset was tested with the purpose of detecting the presence of outliers. Points considered as discrepant were excluded. To test our hypotheses, we considered two scenarios for estimating $\mathbf{X}'\mathbf{X}$ correlation matrices (consider the columns of \mathbf{X} as the standardized traits): In the first scenario, the data used were obtained from values of each sampled observation (place \times hybrid \times replicate \times plant), totaling a dataset with 900 samples (ASO). In the second, the data used were obtained from the average values of the five plants of each replicate (place \times hybrid \times replicate), totaling a dataset with 180 samples (AVP). In each scenario, the sampled traits were subjected to a descriptive analysis to obtain the values related to average, mode, maximum, minimum, and standard deviation. Later, for each dataset, matrices of correlation (phenotypic) and covariance were estimated.

3.3.4 Correlation and covariance matrices

To estimate the degree of interdependence among the traits in each scenario, the traits (see above), represented here by X and Y , formed a dataset $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$. The covariance between X and Y was estimated by the expression:

$$\text{Cov}_{XY} = \sum_{i=1}^n [(X_i - \bar{X})(Y_i - \bar{Y})]$$

which was used to estimate the nonstandard degree of the interrelation between such traits. To estimate the standardized degree of interrelation between the traits, the values of the Pearson product-moment correlation (r) were estimated by (Puth et al., 2014).

$$r = \frac{\sum_{i=1}^n [(X_i - \bar{X})(Y_i - \bar{Y})]}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}} \quad [1]$$

Where

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

and

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$$

The $\mathbf{X}'\mathbf{X}_{11 \times 11}$ matrices of phenotypic correlation and covariance among the explanatory traits PH, AE, EL, ED, NRE, NKR, CD, CL, TNK, CD/ED, and TKW were formed for two scenarios. The correlation coefficients between each explanatory trait and the dependent trait (KWE) generated a $\mathbf{X}'\mathbf{Y}_{11 \times 1}$ correlation matrix.

3.3.5 Multicollinearity diagnosis

We determined the source and magnitude of the multicollinearity in $\mathbf{X}'\mathbf{X}$ correlation matrices in each scenario by the following methods:

Method 1: Eigenvalues and Eigenvectors

In each scenario, the eigenvalues ($\lambda_1, \lambda_2, \dots, \lambda_{11}$), as well as the associated eigenvectors of each matrix of explanatory traits were estimated by the *eigen()* procedure in R software (R Development Core Team, 2008). The eigenvalues indicate the amount of variance explained by

each factor, and the components of the eigenvector are the weights of the traits to explain the variance determined by the associated eigenvalue.

Method 2: Condition Number

The condition number (CN) was obtained by the ratio between the largest and smallest eigenvalue of the $\mathbf{X}'\mathbf{X}$ correlation matrices, using the expression:

$$\text{CN} = \frac{\lambda_{\text{Max}}}{\lambda_{\text{Min}}} \quad [2]$$

The multicollinearity degree of the matrices was respectively considered as weak, moderated, and severe when $\text{CN} \leq 100$, between 100 and 1000, and ≥ 1000 (Montgomery et al., 2012).

Method 3: $\mathbf{X}'\mathbf{X}$ Correlation Matrix Determinant

The determinant of each correlation matrix (MD) was estimated by the product of its respective eigenvalues, for eigenvalues of $\mathbf{X}'\mathbf{X} > 0$ ($\lambda_j > 0$ for $j = 1, 2, \dots, 11$), as described by:

$$\text{MD}_{\mathbf{X}'\mathbf{X}} = \prod_{j=1}^p \lambda_j \quad [3]$$

An MD closer to zero indicated linear dependency among the explanatory traits, indicating severe multicollinearity problems (Cruz et al., 2014).

Method 4: Variance Inflation Factors

The variance inflation factors (VIFs) were used to measure how much the variance of estimated regression coefficients (β) was inflated in comparison to when the explanatory traits were not linearly associated. We estimated the VIF for the k th element of β by the sum of the quotients of each component of the eigenvector divided by its respective associated eigenvalue:

$$\text{VIF}_{\beta_k} = \left(\frac{\text{EV}_{kC1}}{\lambda_1} + \frac{\text{EV}_{kC2}}{\lambda_2} + \dots + \frac{\text{EV}_{kC11}}{\lambda_{11}} \right) \quad [4]$$

Where VIF_{β_k} is the variance inflation factor the k th element of β for $k = 1, 2, \dots, 11$; EV_{kC_1} is the component of the k th eigenvector for $k = 1, 2, \dots, 11$ and $C = 1, 2, \dots, 11$; and λ is the eigenvalue associated with the respective eigenvector for $\lambda = 1, 2, \dots, 11$.

3.3.6 Methods for adjusting multicollinearity

To confirm the third hypothesis, when detected, two ways for adjusting the harmful effects of multicollinearity were tested in each scenario: (i) eliminating the traits responsible for the multicollinearity in the matrix, and (ii) estimating path coefficients with all traits based on normal equations, partially modified by the addition of the k constant in diagonal elements of the $\mathbf{X}'\mathbf{X}$ correlation matrix among explanatory traits (Hoerl and Kennard, 1976, 1981).

3.3.6.1 Determining which traits should be excluded from the model

For each scenario, the origin of the multicollinearity of the $\mathbf{X}'\mathbf{X}$ correlation matrices was assessed by analyzing the eigenvalues and eigenvectors. We adopted the procedure of identifying the traits causing multicollinearity, in order of importance, as the component of the eigenvector (trait with the largest weight) associated with the eigenvalues with the smallest magnitudes (Kirschvink, 1980; Mansfield and Helms, 1982; Cruz et al., 2014; Hu and Qi, 2014).

3.3.6.2 Including the k constant into correlation matrices

Sequences of β_j values for $j = 1, 2, 3, \dots, 11$, obtained by a set of 21 values of k (0.00, 0.05, 0.10, ..., 1.00), where the values of β_j estimated with the value of $k = 0$ correspond to the

estimation of least squares, were estimated according to Hoerl and Kennard (1976). The procedure was based on:

$$\beta_j = (\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1} \mathbf{X}'\mathbf{Y} \quad [5]$$

where β_j is the partial regression coefficient for $j = 1, 2, 3, \dots, 11$; $(\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}$ is the inverse of the $\mathbf{X}'\mathbf{X}$ correlation matrix among explanatory traits with k included in the diagonal elements; and $\mathbf{X}'\mathbf{Y}$ is the matrix of the correlation coefficients between the explanatory traits and the dependent trait (KWE). Using all 11 coefficients of β , for each scenario a graph was made where the y axis was represented by values of β for each of the 21 k values represented on the x axis (Fig. 1). Thus, it was possible to visually determine the smallest necessary k value in each scenario to stabilize the regression coefficients (Cruz et al., 2014).

After the use of these methods to reduce the multicollinearity of the matrices, a new determination was performed for each scenario, using the methods described for the multicollinearity determination.

3.3.7 Direct and indirect effects

In each scenario, linear correlation coefficients were partitioned into direct and indirect effects by solving normal equation systems to fit multiple regression models using ordinary least squares, as described by Quinn and Keough (2002). The explanatory traits included in the model were PH, AE, EL, ED, NRE, NKR, CD, CL, TNK, CD/ED, and TKW. The dependent trait was kernel weight per ear (KWE). We estimated the direct and indirect effects of each p th explanatory trait on KWE using three path analysis methods: (i) traditional path analysis, (ii) with k inclusion, and (iii) traditional path analysis with trait exclusion.

Traditional path analysis

In this methodology, direct and indirect effects (indirect effects are shown in Supplemental Table S1) were estimated by derivation of the set of normal equations used for estimating the parameters of multiple regression models. To estimate β , we solved the model $\mathbf{X}'\mathbf{X}\beta = \mathbf{X}'\mathbf{Y}$, represented in matrix form as:

$$\begin{bmatrix} 1 & r_{\text{PH:EH}} & \cdots & r_{\text{PH:TKW}} \\ r_{\text{EH:PH}} & 1 & \cdots & r_{\text{EH:TKW}} \\ \vdots & \vdots & \ddots & \vdots \\ r_{\text{TKW:PH}} & r_{\text{TKW:EH}} & \cdots & 1 \end{bmatrix} \times \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{11} \end{bmatrix} = \begin{bmatrix} r_{\text{PH:KWE}} \\ r_{\text{EH:KWE}} \\ \vdots \\ r_{\text{TKW:KWE}} \end{bmatrix} \quad [6]$$

The estimate of β was given by: $\beta = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{Y}$, where β is the vector of sample partial regression coefficients ($\beta_1, \beta_2, \dots, \beta_p$) with $p + 1$ rows; $(\mathbf{X}'\mathbf{X})^{-1}$ is the inverse of the $\mathbf{X}'\mathbf{X}$ correlation matrix among explanatory traits; and $\mathbf{X}'\mathbf{Y}$ is the correlation matrix between each explanatory trait and KWE. Solving this model, it was possible to estimate the direct and indirect effects. Consider, as an example, the direct and indirect effects of PH on KWE, given by: $r_{\text{PH:KWE}} = \beta_1 + \beta_2 r_{\text{PH:AE}} + \dots + \beta_{11} r_{\text{PH:TKW}}$, where $r_{\text{PH:KWE}}$ is the linear correlation between PH and KWE; β_1 is the direct effect of PH on KWE; $\beta_2 r_{\text{PH:AE}}$ is indirect effect of PH on KWE via AE, ..., and $\beta_{11} r_{\text{PH:TKW}}$ is the indirect effect of PH on KWE via TKW. Equivalent equations were used for the other predictors. The coefficient of determination of the model was given by: $R^2 = \beta_1 r_{\text{PH:KWE}} + \beta_2 r_{\text{AE:KWE}} + \beta_3 r_{\text{EL:KWE}} + \beta_4 r_{\text{ED:KWE}} + \beta_5 r_{\text{NRE:KWE}} + \beta_6 r_{\text{NKR:KWE}} + \beta_7 r_{\text{CD:KWE}} + \beta_8 r_{\text{CL:KWE}} + \beta_9 r_{\text{TNK:KWE}} + \beta_{10} r_{\text{CD/ED:KWE}} + \beta_{11} r_{\text{TKW:KWE}}$. The noise of the path analysis model was obtained by: $\text{Noise} = \sqrt{1 - R^2}$.

With k inclusion

In this methodology, for each scenario, the 11 explanatory traits were used to estimate the direct and indirect effects on KWE (indirect effects are shown in Supplemental Table S2), but in contrast to the traditional path analysis method, a constant k was included in the diagonal of the $\mathbf{X}'\mathbf{X}$ correlation matrix to reduce the variance associated with the least squares estimator. Thus, the normal partially modified system of equations was solved $(\mathbf{X}'\mathbf{X} + k)\beta = \mathbf{X}'\mathbf{Y}$. (See discussion on including the k constant into correlation matrices above to understand k chosen).

$$\begin{bmatrix} 1+k & r_{\text{PH:EH}} & \cdots & r_{\text{PH:TKW}} \\ r_{\text{EH:PH}} & 1+k & \cdots & r_{\text{EH:TKW}} \\ \vdots & \vdots & \ddots & \vdots \\ r_{\text{TKW:PH}} & r_{\text{TKW:EH}} & \cdots & 1+k \end{bmatrix} \times \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{11} \end{bmatrix} = \begin{bmatrix} r_{\text{PH:KWE}} \\ r_{\text{EH:KWE}} \\ \vdots \\ r_{\text{TKW:KWE}} \end{bmatrix} \quad [7]$$

The estimates of β were given by $\beta = (\mathbf{X}'\mathbf{X} + k)^{-1} \mathbf{X}'\mathbf{Y}$, where β is the vector of sample partial regression coefficients $(\beta_1, \beta_2, \dots, \beta_p)$ with $p + 1$ rows, and $(\mathbf{X}'\mathbf{X} + k)^{-1}$ is the inverse of the $(\mathbf{X}'\mathbf{X} + k)$ correlation matrix among explanatory traits with k included in diagonal elements. Thus, like traditional path analysis, it was possible to estimate the direct and indirect effects of each explanatory trait on KWE, however without the harmful effects of multicollinearity. The coefficient of determination and the noise were also estimated as described for the traditional path analysis.

With trait exclusion

In this methodology, the direct and indirect effects (indirect effects are given in Supplemental Table S3), the coefficient of determination and the noise were estimated for each scenario, as described for traditional path analysis, after the exclusion of the multicollinearity-generating traits (Cruz et al., 2014).

3.4 RESULTS

The results of the descriptive analysis and the correlation and covariance matrices obtained in the studied scenarios are shown. Next, the results of the multicollinearity diagnosis are presented. Analysis of the eigenvalues and the components of the eigenvectors showed which were the traits that were causing multicollinearity and which were considered for exclusion. A new multicollinearity determination performed after the use of the mentioned measures for adjusting it is also presented. Ultimately, it was demonstrated how the three path analysis methodologies, performed on both scenarios, influenced the direct effects, precision, and noise of the models.

3.4.1 Descriptive statistics

Table 1 shows the descriptive analysis of the explanatory traits obtained for each scenario. It is clear that PH, AE, EL, CD, and CL presented the largest averages in the AVP scenario. The mode of the traits CD, TNK, and CD/ED ratio was largest in the AVP scenario. In this scenario, the amplitude of the data was meaningfully smallest, resulting in the smallest standard deviation for all assessed traits (Table 1).

3.4.2 Correlation and covariance matrices

From 55 tested combinations, the correlation matrix estimated in the ASO scenario ($n = 900$) presented 51 meaningful combinations ($p < 0.01$), with the largest correlation observed between EL and CL ($r = 0.906$). For the correlation matrix estimated for the AVP scenario ($n = 180$), 42 combinations were meaningful, with the largest correlation obtained between AE and PH ($r = 0.925$) (Table 2).

Comparing the same combinations between both scenarios, excluding $TNK \times EL$, $CD/ED \times ED$, $NRE \times TKW$, $NKR \times CD$, $TNK \times CD$, and $CL \times TNK$, approximately 90% of studied combinations (49) presented the largest magnitudes in the AVP scenario.

3.4.3 Multicollinearity Diagnosis

For the ASO scenario, the eigenvalue amplitude ranged from 4.4088 to 2.2×10^{-3} , resulting in a CN of 2004.00 (Eq. [2]), with an MD of 3.024×10^{-6} (Eq. [3]). For the AVP scenario, the eigenvalue amplitude was larger, from 5.0943 to 1.5×10^{-3} , which resulted in a CN of 3395.20 and, consequently, in a MD still smaller: 1.26×10^{-7} . According to Montgomery et al. (2012), severe multicollinearity was observed for both scenarios; however, the largest problems were in the AVP scenario (Table 3).

Three VIFs >10 were observed in both scenarios. For the ASO scenario, the VIFs >10 were 143.2167, 196.3047, and 116.7618 for ED, CD, and the CD/ED ratio, respectively. In the AVP scenario, the variances of those coefficients were inflated to 214.2254, 310.2553, and 147.1021, respectively (see above for understanding VIF estimates). On average, the factors that inflated the variance of the regression coefficients were increased by 61% in the AVP scenario. It is also noticeable that, although there are just three VIFs >10 in the AVP scenario, the VIFs of PH and AE were inflated by 150 and 135%, respectively, presenting magnitudes of 9.5969 and 8.6245, respectively (Table 3). Based on Eq. [4], the largest VIFs of PH and AE in the AVP scenario occurred due to the largest weight of these traits linked to eigenvalues of TNK and the CD/ED ratio close to zero (Table 5).

3.4.4 Multicollinearity-generating traits

The analysis of eigenvalues and eigenvectors in the ASO scenario (Table 4) revealed that the multicollinearity in this scenario was generated principally by the CD, CL, and TNK traits. According to the adopted methodology for exclusion of traits, CD was excluded from the model because it presented the largest weight (0.6556) linked to the smallest associated eigenvalue (0.0022), and CL, which presented the largest weight (0.6772) linked to the second-smallest associated eigenvalue (0.0872). Researchers must carefully choose which traits should be excluded because the exclusion of traits with high explanatory power might reduce the accuracy of the analysis. The TNK and TKW traits (indicated third and fourth, respectively, for exclusion) are important traits in maize genetic breeding and might meaningfully contribute with path coefficients, improving the coefficient of determination of the model. Therefore, the third component of the eigenvector (NKR) with the largest weight (-0.3172) linked to the third-smallest associated eigenvalue (0.1036) was excluded.

Differently from the ASO scenario, the data presented in Table 5 indicated CD, PH, and AE as the traits that should be excluded from the model. The CD presented the largest weight (0.6807) linked to the smallest associated eigenvalue (0.0015). The PH presented the largest weight (0.5984) linked to the second-smallest associated eigenvalue (0.0475). For the third-smallest eigenvalue (0.0744), the trait with the largest weight (0.3330) was also PH, thus the second component of the eigenvector with the largest weight (AE) was considered for exclusion.

3.4.5 Multicollinearity determination after adjustment

The inclusion of the k constant in the diagonals of the matrix reduced the multicollinearity effects for both scenarios compared with the traditional path analysis. In the

ASO scenario, the inclusion of 0.05 in the diagonal of the $\mathbf{X}'\mathbf{X}$ matrix resulted in a CN of 85.404, in a MD of 3.740×10^{-4} , and in the largest VIF of 9.907 linked to CD. For the AVP scenario, the inclusion of 0.10 at the diagonal of the $\mathbf{X}'\mathbf{X}$ matrix was efficient in reducing the harmful effects of multicollinearity for a CN of 51.175, a MD of 5.39×10^{-4} , and no VIF >10 (Table 6).

The exclusion of the traits CD, CL, and NKR in the ASO scenario presented better responses in the sense of reducing the multicollinearity of the matrix. This methodology's CN was 29.620, and the MD was of 1.2×10^{-2} , with the largest VIF of 4.022 linked to ED. For the AVP scenario, however, the exclusion of the traits CD, PH, and AE did not result in a meaningful improvement of the multicollinearity of the matrix, as happened in the ASO scenario. This methodology's CN (63.656) was larger than in ASO scenario, with a meaningfully smaller MD (7.938×10^{-4}). The largest correlation between the remaining traits in the model was $r = 0.924$ between EL and CL, possibly resulting in problems to estimate path coefficients.

3.4.6 Direct effects and accuracy

In the ASO scenario, the distorted estimations of the direct effects become obvious, especially for ED, CD, and the CD/ED ratio, reaching magnitudes > 2 (Table 7). Unexpectedly, the coefficient of determination of 1.020 and the noise 0.000 also revealed that a path analysis performance under multicollinearity is a problem that must be taken into consideration, especially under conditions classified as severe.

The inclusion of the k constant (0.05) presented direct effects more consistent with the biological expectation. The traits TNK and TKW presented the most meaningful direct effects, with $r = 0.636$ and $r = 0.564$, respectively (Table 7). The coefficient of determination of 0.931

and the noise of 0.261 indicated the effectiveness of the k constant to reduce the undesired effects of multicollinearity in path coefficient estimations.

After excluding the traits CD, CL, and NKR, TNK and TKW presented direct effects of $r = 0.892$ and $r = 0.733$, respectively, on KWE (Table 7). For this methodology, the coefficient of determination was 0.977, with noise of 0.150. These findings showed that the exclusion of multicollinearity-generating traits provided the largest coefficient of determination and the smallest noise among the studied methods.

For the AVP scenario, the direct effects traditionally estimated had different magnitudes than in the ASO scenario, especially in relation to ED, CD, and the CD/ED ratio, presenting direct effects of $r = 0.476$, $r = -0.602$, and $r = 0.403$, respectively. Although the coefficient of determination of 0.973 and noise of 0.161 indicate a good precision in the analysis, the findings proved themselves contrary to what was observed when we included the k constant in the correlation matrix. With k inclusion (0.10), ED, CD, and the CD/ED ratio presented direct effects of $r = 0.128$, $r = 0.046$, and $r = -0.040$, respectively. The coefficient of determination was 0.922, with noise of 0.278 (Table 7). It can be observed that the noise of this methodology was larger than for ASO scenario. Thus, we proved that although k inclusion adjusted the multicollinearity to an acceptable degree, the success of this technique depends on the original degree of multicollinearity present in the matrices of the explanatory traits.

After excluding the traits PH, AE, and CD, the direct effects of TNK and TKW were $r = 0.653$ and $r = 0.645$, respectively (Table 7). The coefficient of determination in this path analysis method (0.973) was somewhat smaller than that for the ASO scenario, and the noise analysis (0.165) was largest.

It is obvious that multicollinearity generates bias in the estimation of path coefficients, especially in traits with a large VIF. Therefore, we demonstrate that a previous and reliable

multicollinearity determination of the $\mathbf{X}'\mathbf{X}$ matrix needs to be performed to create reliable path analysis results.

3.5 DISCUSSION

3.5.1 Correlation coefficients estimated with data average are overestimated

Correlation coefficients estimated with the data average are overestimated. The smallest standard deviation observed in the AVP scenario confirmed the first hypothesis that the use of averages masks individual variances. It was proved that when correlation matrices are estimated at the average plot level, correlation coefficients are overestimated, and consequently, multicollinearity in explanatory traits matrices presents the largest problems. The reduction in individual variation (standard deviation) observed in the AVP scenario (Table 1) was the main factor responsible for overvaluing of 90% of the combination pairs (Table 2). This fact can be explained due to the standard deviation being the divisor in the correlation's formula (Eq. [2]). If covariance XY (dividend of the formula) is similar in both scenarios, however, the standard deviation of X and Y traits (divisor of the formula) is smallest, as observed in the AVP scenario, and the magnitude of the correlation coefficients will be greater.

When the correlation between explanatory traits increases, the difficulties in assessing its relative importance in estimating the dependent trait are greatest (Blalock, 1963; Hoerl and Kennard, 1981). Therefore, the determination of the degree of association among explanatory traits and of the degree of multicollinearity in the matrices of the explanatory traits are vital steps that come before the path analysis estimates.

3.5.2 Preserving individual variances, multicollinearity is reduced

It is evident that the AVP scenario presented bigger problems of multicollinearity due to the largest VIF magnitudes, the largest CN (3396.20), the smallest MD (1.22×10^{-7}), and

four eigenvalues near zero (Table 3). Because calculations of multiple regression analysis involve matrix inversion, and this inversion basically involves dividing by MD, when a determinant near zero is observed, values in the inverted matrix become very sensitive to small differences in the data of the original matrix, or in other words, the inverted matrix is unstable (Farrar and Glauber, 1967; Gunst and Mason, 1977; Mansfield and Helms, 1982; Quinn and Keough, 2002). For the ASO scenario, however, just the last two eigenvalues present extremely low estimates, indicating the existence of only two decisive linear relations with the harmful effects of multicollinearity (Table 2). Consequently, we suggest that the correlation should not be estimated based on the sampled averages of the plots because they overvalue the correlation magnitudes, generating bigger problems with multicollinearity in the matrices of the explanatory traits.

The largest VIF, in both scenarios linked to CD, ED, and CD/ED (Table 3), was expected because CD/ED is the result of the ratio between CD and ED. When CD was excluded from the model, the considerable reduction in multicollinearity of the matrices was evident (Table 6). With that, the degree of trustworthiness of the path coefficients depends on the researcher's ability in choosing the explanatory traits with more power for data representations that are not highly correlated and, in the case of problems, such as multicollinearity, to take the right measures to adjust it (Cruz et al., 2012).

Some research prominently showed the magnitude of multicollinearity in their findings (Carvalho et al., 1999a; Toebe and Cargnelutti, 2013). In others, it was only mentioned that it was there and that some measures were taken to adjust it (Bizeti et al., 2004; Coimbra et al., 2005; Nogueira et al., 2012). We observed, however, that several agronomic studies did not clearly reveal whether multicollinearity determination was performed before the estimation of path analysis (Alvi et al., 2003; Saleem et al., 2007; Reddy et al., 2012; Kumar et al., 2013; Nataraj et al., 2014; Adesoji et al., 2015; Pavlov et al., 2015). This is worrying. Because the

problems related to multicollinearity are not generated only by its presence (Cruz et al., 2014), to assess its magnitude and origin is a fundamental step to choose the best method for adjusting it.

3.5.3 Data arrangement change the efficiency of methods to adjusting multicollinearity

In the procedure with k inclusion, when the ridge trace was examined (Fig. 1), the largest value of k necessary (0.10) to stabilize the regression coefficients in the AVP scenario became clear. In this scenario, k inclusion was more efficient to reduce the degree of multicollinearity than in the ASO scenario. However, the smallest coefficients of determination and the largest noise (Table 7) indicated that largest k values generate bias in regression analysis. This observation was also mentioned by Hoerl and Kennard (1970b) and Cruz et al. (2014). In addition, the residual effect was larger in the AVP than the ASO scenario when path analysis was performed with trait exclusion (Table 7).

There are many variable selection methods for choosing a subset of model terms with minimal multicollinearity, such as hierarchical models, stepwise procedures, and criterion-based procedures (George and McCulloch, 1993; Mitchell and Beauchamp, 1988; Nishii, 1984; Wold et al., 1984). However, there is a shortage of research that shows, in theory and in practice, how to discover the traits responsible for multicollinearity in path analysis; our findings in this study make this interpretation possible.

Toebe and Cargnelutti (2013) were successful by excluding the traits of ear height and number of ears, or adding the k constant = 0.10 in the diagonal of the correlation matrices among the explanatory traits obtained from simple maize hybrids trials. However, the largest coefficient of determination and the smallest noise found by the researchers were of 0.950 and 0.230, respectively, which were obtained with five explanatory traits (number of days to 50% tasseling, plant height at harvest, relative ear position, number of plants at harvest, and

prolificacy). It is important to highlight that the correlation matrices of their study were estimated with the average of three observations (plants) in each plot. In the present study, when the values of all sampled observations (ASO scenario) were used and the path coefficients were estimated with eight explanatory traits, except for CD, CL, and NKR, it was possible to reduce the noise of the model by almost 35%. This fact is possibly linked to the larger number of explanatory traits used and to the choice of traits.

3.5.4 Data based on averages reduce direct effects and increase the noise in path analysis

The smallest direct effects of TNK and TKW in the AVP scenario (Table 7) can be attributed to the largest effects of multicollinearity in this scenario because both PH and AE, excluded from the ASO scenario, and NKR and CL, excluded from the AVP scenario, presented meaningful direct effects to change the magnitudes of TNK and TKW in both scenarios (Table 7).

In our research, a logical relation in the path coefficients was noticed. Previous studies also found a direct and positive contribution of TNK on KWE, fluctuating between $r = 0.52$ and $r = 0.78$ (Mohammadi et al., 2003; Bello et al., 2010; Khameneh et al., 2012), and TKW on KWE, fluctuating between $r = 0.48$ and $r = 0.74$ (Mohammadi et al., 2003; Nastasić et al., 2010; Khameneh et al., 2012; Reddy et al., 2012; Adesoji et al., 2015). In addition, the large coefficient of determination and the small noise demonstrate that the model was efficient in explaining the variation in the dependent trait.

It was clear that a path analysis performed with all sampled observations (ASO scenario), excluding the traits NKR, CD, and CL, was the most trustworthy, being linked to the largest coefficient of determination (0.977), the smallest noise (0.150), and the smallest VIFs observed among the three tested methodologies in both scenarios (Table 7). Thus, we thought

that it is unacceptable, after obtaining the data, to use methodologies that mask the correlation coefficient and reduce path analysis accuracy.

To estimate correlations using data from average values is a biased procedure. Furthermore, the inference about the magnitude of relationships between traits is equivocal because this inference is performed on a population with a variance different than the original. A great number of agronomic studies perform populational inferences based on sampled observations (plants); thus, to use average values of these plants to represent the variance of the original population should be considered, without doubt, a mistake.

In the present research, we proved that average values overestimate correlation coefficients. Also, we demonstrated that researchers can reduce systematic errors and avoid the harmful effects of multicollinearity in explanatory traits matrices that are not linked to a trait's nature by adopting a simple methodology: estimating correlation matrices with data from all sampled observations. We believe that this data arrangement method can be easily applied in future plant breeding research projects, as well as in other areas of science, without a substantial increase in time, labor, or financial resources. Thus, aiming at the correct estimate of the correlation coefficients in future research, we encourage researchers to not estimate correlation matrices based on average data but to do it with all sampled observations to prevent masking the real existing variance of each trait. Success in making estimates of path coefficients aimed at indirect selection in breeding programs, however, will depend on researchers' abilities in correctly estimating the correlation coefficients and in correctly using the methods described here for adjusting the multicollinearity of the correlation matrices. In addition to these factors, success in indirect selection also depends on performing it based on traits with high heritability that are directly associated with the dependent trait. If the plant breeder considers the correlations calculated from average values, these correlations will be overestimated, a fact that

can mask the estimates of path coefficients and consequently the choice of the trait for indirect selection.

Considering that in this research the data used to estimate the $\mathbf{X}'\mathbf{X}$ correlation matrices in both scenarios was the same, with the only difference being procedural in nature, we believe that we have proved, for the first time, that accuracy in path analyses will be greater if all identification procedures (sample tracking) are performed and correlation matrices are estimated with all sampled observations after the exclusion of traits that generate multicollinearity.

3.6 CONCLUSIONS

The use of average data suppresses the individual variation and overestimates the magnitude of correlation among traits; thus, the correlation matrices among explanatory traits estimated with average data have the largest multicollinearity. The best strategy to mitigate this problem is to perform the estimates of correlation coefficients with data coming from all sampled observations, excluding the traits responsible for inflating the variance of regression coefficients. By using these methodologies, the fit statistics of path analysis will be more accurate.

3.7 ACKNOWLEDGMENTS

We thank the Coordination for the Improvement of Higher Education Personnel (CAPES) for granting the master's scholarship for the first author, the colleagues Amanda Baseggio and Jaksson A. F. Klin for their valuable collaboration in conducting the field trials and the Mr. John Stolzle by English grammar review. We also would like to thank AJ's editorial review board members and the three anonymous reviewers for the valuable contribution given in this manuscript.

Supplemental material with indirect effects is available. Supplemental table S1: path analysis traditionally estimated; Supplemental table S2; path analysis with k inclusion and Supplemental table S3: path analysis excluding multicollinearity-generating traits.

3.8 REFERENCES

- Adesoji, A.G., I.U. Abubakar, and D.A. Labe. 2015. Character association and path coefficient analysis of maize (*Zea mays* L.) grown under incorporated legumes and nitrogen. *J. Agron.* 14:158–163. doi:10.3923/ja.2015.158.163
- Aliyu, L., M.K. Ahmed, and M.D. Magaji. 2000. Correlation and multiple regression analysis between morphological characters and components of yield in pepper (*Capsicum annuum* L.). *Crop Res.* 19:318–323.
- Alvi, M.B., M. Rafique, M.S. Tariq, A. Hussain, T. Mahmood, and M. Sarwar. 2003. Character association and path coefficient analysis of grain yield and yield components maize (*Zea mays* L.). *Pak. J. Biol. Sci.* 6:136–138. doi:10.3923/pjbs.2003.136.138
- Bello, O.B., S.Y. Abdulmalik, M.S. Afolabi, and S.A. Ige. 2010. Correlation and path coefficient analysis of yield and agronomic characters among open pollinated maize varieties and their F₁ hybrids in a diallel cross. *Afr. J. Biotechnol.* 9:2633–2639. doi:10.4314/ajb.v9i18
- Bizeti, H.S., C.G.P. de Carvalho, J.R.P. de Souza, and D. Destro. 2004. Path analysis under multicollinearity in soybean. *Braz. Arch. Biol. Technol.* 47:669–676. doi:10.1590/S1516-89132004000500001
- Blalock, H.M. 1963. Correlated independent variables: the problem of multicollinearity. *Soc. Forces* 42:233–237. doi:10.1093/sf/42.2.233

- Carvalho, S.P. de, C.D. Cruz, and C.G.P. de Carvalho. 1999a. Estimating gain by use of a classic selection index under multicollinearity in wheat (*Triticum aestivum*). *Genet. Mol. Biol.* 22: 109–113. doi:10.1590/S0100-204X1999000400011
- Carvalho, C.G.P. de, V.R. Oliveira, C.D. Cruz, and V.W.D. Casali. 1999b. Análise de trilha sob multicolinearidade em pimentão (In Portuguese, with English abstract) *Pesq. Agropec. Bras.* 34:603–613. doi:10.1590/S0100-204X1999000400011
- Coimbra, J.L.M., G. Benin, E.A. Vieira, A.C. de Oliveira, F.I.F. Carvalho, A.F. Guidolin, and A.P. Soares. 2005. Conseqüências da multicolinearidade sobre a análise de trilha em canola. (In Portuguese, with English abstract) *Cienc. Rural* 35:347–352. doi:10.1590/S0103-84782005000200015
- Cruz, C.D., P.C.S. Carneiro, and A.J. Regazzi. 2014. *Modelos Biométricos Aplicados ao Melhoramento Genético*. 3rd ed. UFV, Viçosa, MG.
- Cruz, C.D., A.J. Regazzi, and P.C.S. Carneiro. 2012. *Modelos Biométricos Aplicados ao Melhoramento Genético*. 4th ed. UFV, Viçosa, MG.
- Faria, L.A., J.M. Peluzio, F.S. Afférri, E.V. de Carvalho, M.A. Dotto, and E.A. Faria. 2015. Análise de trilha para crescimento e rendimento de genótipos de milho sob diferentes doses nitrogenadas. (In Portuguese, with English abstract) *J. Bioenerg. Food Sci.* 2:1-11. doi:10.18067/jbfs.v2i1.13
- Farrar, D.E., and R.R. Glauber. 1967. Multicollinearity in regression analysis: the problem revisited. *Rev. Econ. Stat.* 49:92–107.
- Gunst, R.F., and R.L. Mason. 1977. Advantages of examining multicollinearities in regression analysis. *Biometrics* 33:249–260. doi:10.2307/2529320
- Hoerl, A.E., and R.W. Kennard. 1970b. Ridge Regression: Biased estimation for nonorthogonal problems. *Technometrics* 12:55-67. doi:10.2307/1267351

- Hoerl, A.E., and R.W. Kennard. 1976. Ridge regression iterative estimation of the biasing parameter. *Commun. Stat. Theor. Methods* 5:77–88. doi:10.1080/03610927608827333
- Hoerl, A.E., and R.W. Kennard. 1981. Ridge regression - 1980: advances, algorithms, and applications. *Am. J. Math. Manag. Sci.* 1:5–83. doi:10.1080/01966324.1981.10737061
- Hu, S., and L. Qi. 2014. The eigenvectors associated with the zero eigenvalues of the Laplacian and signless Laplacian tensors of a uniform hypergraph. *Discrete Appl. Math.* 169:140–151. doi:10.1016/j.dam.2013.12.024
- Khameneh, M.M., S. Bahraminejad, F. Sadeghi, S.J. Honarmand, and M. Maniee. 2012. Path analysis and multivariate factorial analyses for determining interrelationships between grain yield and related characters in maize hybrids. *Afr. J. Agric. Res.* 7:6437–6446. doi:10.5897/AJAR11.1581
- Kirschvink, J.L. 1980. The least-squares line and plane and the analysis of palaeomagnetic data. *Geophys. J. Int.* 62:699–718. doi:10.1111/j.1365-246X.1980.tb02601.x
- Kumar, V., S.K. Singh, P.K. Bhati, A. Sharma, S.K. Sharma, and V. Mahajan. 2015. Correlation, Path and Genetic Diversity Analysis in Maize (*Zea mays* L.). *Environ. Ecol.* 33:971–975.
- Kumar, K.V., M.R. Sudarshan, K.S. Dangi, and S.M. Reddy. 2013. Character association and path coefficient analysis for seed yield in quality protein maize *Zea mays* L. *J. Res. ANKRAU* 41:153–157.
- Luz, L.N. da, R.C. dos Santos, P.A.M. Filho. 2011. Correlations and path analysis of peanut traits associated with the peg. *Crop Breed. Appl. Biotechnol.* 11:88–95. doi:10.1590/S1984-70332011000100013
- Mansfield, E.R., and B.P. Helms. 1982. Detecting multicollinearity. *Am. Stat.* 36:158–160. doi:10.1080/00031305.1982.10482818

- George, E.I., and R.E. McCulloch. 1993. Variable selection via Gibbs sampling. *J. Amer. Statist. Assoc.* 88:881–889. doi:10.1080/01621459.1993.10476353
- Mitchell, T.J., and J.J. Beauchamp. 1988. Bayesian variable selection in linear regression. *J. Amer. Statist. Assoc.* 83:1023-1032. doi: 10.1080/01621459.1988.10478694
- Mohammadi, S.A., B.M. Prasanna, and N.N. Singh. 2003. Sequential path model for determining interrelationships among grain yield and related characters in maize. *Crop Sci.* 43:1690-1697. doi:10.2135/cropsci2003.1690
- Montgomery, D.C., E.A. Peck, and G. Vining. 2012. *Introduction to linear regression analysis*. 5th ed. John Wiley & Sons, New Jersey.
- Nardino, M., V.Q. de Souza, D. Baretta, V.A. Konflanz, I.R. Carvalho, D.N. Follmann, and B.O. Caron. 2016. Association of secondary traits with yield in maize F₁'s. *Cienc. Rural* 46:776–782. doi:10.1590/0103-8478cr20150253
- Nastasić, A., D. Jocković, M. Ivanović, M. Stojaković, J. Boćanski, I. Đalović, and Z. Srećkov. 2010. Genetic relationship between yield and yield components of maize. *Genetika* 42:529–534. doi:10.2298/GENSR1003529N.
- Nataraj, V., J.P. Shahi, and V. Agarwal. 2014. Correlation and path analysis in certain inbred genotypes of maize (*Zea Mays* L.) at Varanasi. *Int. J. Innov. Res. Dev.* 3:14-17.
- Nataraj, V., J.P. Shahi, and D. Vandana. 2015. Character association and path analyses in maize (*Zea mays* L.). *Environ. Ecol.* 33:78–81.
- Nishii, R. 1984. Asymptotic properties of criteria for selection of variables in multiple regression. *Ann. Statist.* 12(2): 758-765. doi:10.1214/aos/1176346522
- Nogueira, A.P.O., T. Sediya, L.B. de Sousa, O.T. Hamawaki, C.D. Cruz, D.G. Pereira, and É. Matsuo. 2012. Análise de trilha e correlações entre caracteres em soja cultivada em duas épocas de semeadura. (In Portuguese, with English abstract) *Biosci. J.* 28:877-888.

- Pavlov, J., N. Delić, K. Marković, M. Crevar, Z. Čamdžija, and M. Stevanović. 2015. Path analysis for morphological traits in maize (*Zea mays* L.). *Genetika* 47:295–301. doi:10.2298/GENSR1501295P
- Pearson, K. 1920. Notes on the history of correlation. *Biometrika* 13:25–45. doi:10.2307/2331722
- Petratis, P.S., A.E. Dunham, and P.H. Niewiarowski. 1996. Inferring multiple causality: the limitations of path analysis. *Funct. Ecol.* 10:421–431. doi:10.2307/2389934
- Puth, M.T., M. Neuhäuser, and G.D. Ruxton. 2014. Effective use of Pearson's product–moment correlation coefficient. *Anim. Behav.* 93:183–189. doi:10.1016/j.anbehav.2014.05.003
- Quinn, G.P., and M.J. Keough. 2002. *Experimental Design and Data Analysis for Biologists*. Cambridge University Press, New York.
- R Development Core Team (2008). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Reddy, V.R., F. Jabeen, M.R. Sudarshan, and A.S. Rao. 2012. Studies on genetic variability, heritability, correlation and path analysis in maize (*Zea mays* L.) Over locations. *Int. J. Appl. Biol. Pharm. Technol.* 4:196–199.
- Rigon, J.P.G., S. Capuani, J.F. de Brito Neto, G.M. da Rosa, A.D. Wastowski, and C.A.G. Rigon. 2012. Dissimilaridade genética e análise de trilha de cultivares de soja avaliada por meio de descritores quantitativos. (In Portuguese, with English abstract) *Rev. Ceres* 59:233–240. doi:10.1590/S0034-737X2012000200012
- Saleem, A.R., U. Saleem, and G.M. Subhani. 2007. Correlation and path coefficient analysis in maize (*Zea mays* L.). *J. Agric. Res.* 45:177–183.
- Shrivastava, M., and K. Sharma. 1976. Analysis of path coefficients in rice. *Z. Fuer Pflanzenzuechtung* 77:174–177.

- Silva, S.A., F.I.F. de Carvalho, J.L. Nedel, P.J. Cruz, J.A.G. da Silva, V. da R. Caetano, I. Hartwig, and C. da S. Sousa. 2005. Análise de trilha para os componentes de rendimento de grãos em trigo. (In Portuguese, with English abstract) *Bragantia* 64:191–196. doi:10.1590/S0006-87052005000200004
- Toebe, M., and A. Cargnelutti Filho. 2013. Multicollinearity in path analysis of maize (*Zea mays* L.). *J. Cereal Sci.* 57:453–462. doi:10.1016/j.jcs.2013.01.014
- Torres, F.E., P.E. Teodoro, L.P. Ribeiro, C.C.G. Correa, F.B. Hernandez, R.L. Fernandes, A.C. Gomes, and K.V. Lopes. 2015. Correlations and path analysis on oil content of castor genotypes. *Biosci. J.* 31:1363-1369. doi:10.14393/BJ-v31n5a2015-26391
- Vaux, D.L., F. Fidler, and G. Cumming. 2012. Replicates and repeats-what is the difference and is it significant? *EMBO Rep.* 13:291–296. doi:10.1038/embor.2012.36
- Wold, H. 1954. Causality and econometrics. *Econometrica* 22:162–177. doi:10.2307/1907540
- Wold, S., A. Ruhe, H. Wold, and I. Dunn W. 1984. The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. *SIAM J. Sci. and Stat. Comput.* 5(3): 735-743. doi:10.1137/0905052
- Wright, S. 1921. Correlation and causation. *J. Agric. Res.* 20:557–585.
- Wright, S. 1923. The theory of path coefficients a reply to Niles's criticism. *Genetics* 8:239–255.
- Wright, S. 1934. The method of path coefficients. *Ann. Math. Stat.* 5:161–215. doi:10.1214/aoms/1177732676

Table 1. Descriptive analysis for the 11 explanatory traits estimated in the two data arrangement scenarios.

Traits†	Average		Mode		Minimum		Maximum		Standard deviation	
	ASO‡	AVP§	ASO	AVP	ASO	AVP	ASO	AVP	ASO	AVP
PH	2.468	2.470	2.600¶	2.680	1.000	1.670	3.300	3.068	0.379	0.351
AE	1.333	1.333	1.450	1.032 ¶	0.500	0.662	2.390	1.912	0.317	0.298
EL	15.129	15.142	15.400¶	15.080¶	0.800	9.700	20.900	19.100	2.277	1.512
ED	4.939	4.938	4.845¶	4.638¶	3.140	3.826	5.969	5.465	0.391	0.323
NRE	16.064	16.048	16.000	15.600	10.000	12.000	22.000	21.600	2.259	1.737
NKR	32.144	32.101	34.000	30.400	12.000	19.667	50.000	45.000	6.078	4.230
CD	28.943	28.971	28.260	29.460¶	19.360	22.985	41.460	33.560	3.017	2.465
CL	15.941	15.959	17.000	15.300¶	7.800	11.360	21.800	19.800	2.131	1.416
TNK	510.215	508.821	419.000¶	444.200¶	83.000	157.800	1104.000	772.600	123.844	92.219
CD/ED	0.587	0.587	0.572	0.596	0.401	0.495	0.846	0.687	0.048	0.035
TKW	339.039	338.704	368.918¶	213.796¶	122.779	213.796	546.309	439.021	63.228	47.887

† PH, plant height; EH, ear height; EL, ear length; ED, ear diameter; NRE, number of rows per ear; NKR, number of kernels per row; CD, cob diameter; CL, cob length; TNK, total number of kernels per ear; CD/ED, cob diameter/ear diameter ratio; TKW, thousand-kernel weight.

‡ All sampled observations.

§ Average values of plot.

¶ In the case of multiple modes, the smallest is shown.

Table 2. Correlation and covariance matrices among 11 explanatory traits obtained with all sampled observations, $n = 900$ (upper diagonal) and with the average values of each plot, $n = 180$ (below diagonal).

Traits†		PH	AE	EL	ED	NRE	NKR	CD	CL	TNK	CD/ED	TKW
PH	r‡	1	0.838**	0.274**	0.490**	0.248**	0.257**	0.273**	0.251**	0.327**	-0.132**	0.402**
	Cov§		0.101	0.236	0.072	0.212	0.590	0.312	0.202	15.273	-0.002	9.529
AE	r	0.925*	1	0.246**	0.474**	0.177**	0.232**	0.328**	0.202**	0.260**	-0.049 ^{ns}	0.415**
	Cov	0.097		0.177	0.059	0.126	0.445	0.313	0.136	10.131	-0.001	8.228
EL	r	0.414*	0.400**	1	0.457**	0.055**	0.654**	0.328**	0.907**	0.578**	-0.034 ^{ns}	0.359**
	Cov	0.221	0.181		0.407	0.285	9.064	2.257	4.400	162.132	-0.004	49.773
ED	r	0.642*	0.610**	0.516**	1	0.503**	0.355**	0.657**	0.464**	0.566**	-0.161**	0.484**
	Cov	0.729	0.590	2.516		0.444	0.843	0.774	0.386	27.181	-0.003	11.492
NRE	r	0.362*	0.283**	0.081 ^{ns}	0.578**	1	0.098**	0.275**	0.050 ^{ns}	0.527**	-0.143**	-0.209**
	Cov	0.221	0.147	0.214	3.240		1.341	1.878	0.242	147.884	-0.015	-29.794
NKR	r	0.402*	0.388**	0.660**	0.395**	0.165 ^{ns}	1	0.097**	0.654**	0.731**	-0.233**	0.093**
	Cov	0.599	0.491	4.220	5.396	1.212		1.782	8.475	553.101	-0.067	34.783
CD	r	0.378*	0.441**	0.349**	0.730**	0.299**	0.064 ^{ns}	1	0.363**	0.218**	0.634**	0.482**
	Cov	0.328	0.325	1.302	5.805	1.281	0.675		2.334	80.970	0.091	91.227
CL	r	0.370*	0.341**	0.924**	0.524**	0.064 ^{ns}	0.648**	0.389**	1	0.560**	0.010 ^{ns}	0.383**
	Cov	0.185	0.145	1.977	2.391	0.157	3.881	1.357		148.057	.001	50.147
TNK	r	0.489*	0.420**	0.543**	0.595**	0.624**	0.778**	0.180 ^{ns}	0.505**	1	-0.290**	-0.105**
	Cov	15.883	11.620	75.743	177.098	99.955	303.528	40.828	65.898		-1.668	-796.122
CD/ED	r	-0.175*	-0.051 ^{ns} ¶	-0.083 ^{ns}	-0.078 ^{ns}	-0.217**	-0.362**	0.622**	-0.033 ^{ns}	-0.422**	1	0.161**
	Cov	-0.002	-0.001	-0.004	-0.009	-0.013	-0.053	0.053	-0.002	-1.346		0.479
TKW	r	0.539*	0.554**	0.451**	0.623**	-0.082 ^{ns}	0.139 ^{ns}	0.645**	0.470**	0.012 ^{ns}	0.237**	1
	Cov	9.094	7.947	32.629	96.271	-6.794	28.110	76.178	31.833	52.757	0.393	

** meaningful at 0.01 probability level.

† PH, plant height; EH, ear height; EL, ear length; ED, ear diameter; NRE, number of rows per ear; NKR, number of kernels per row; CD, cob diameter; CL, cob length; TNK, total number of kernels per ear; CD/ED, cob diameter/ear diameter ratio; TKW, thousand-kernel weight.

‡ r, correlation coefficient.

§ Cov, covariance.

¶ ns, nonmeaningful at the 0.01 probability level.

Table 3. Multicollinearity diagnosis for Pearson product-moment correlation matrices among 11 explanatory traits estimated in the two data arrangement scenarios.

Order	All sampled observations		Average value of plots	
	Eigenvalues	VIF [†]	Eigenvalues	VIF
1	4.4088	3.8361	5.0943	9.5969
2	2.0327	3.6679	2.2155	8.6245
3	1.6130	6.2261	1.5000	7.7949
4	1.3324	143.2167	1.0560	214.2254
5	0.7012	2.7460	0.5208	4.5611
6	0.3632	3.3364	0.3047	5.3588
7	0.2028	196.3047	0.1053	310.2553
8	0.1530	6.5360	0.0801	8.1903
9	0.1036	6.3061	0.0744	9.1945
10	0.0872	116.7618	0.0475	147.1021
11	0.0022	4.4572	0.0015	5.6768

[†] VIF, variance inflation factor.

Table 4. Eigenvalues and components of the eigenvectors of Pearson product-moment correlation matrix among the 11 explanatory traits estimated with all sampled observations, $n = 900$.

Eigenvalues (EV)	Components of the eigenvectors (Weight of the traits)											
	PH	AE	EL	ED	NRE	NKR	CD	CL	TNK	CD/ED	TKW	
EV ₁	4.4088	0.3072	0.2922	0.3726	0.3913	0.1797	0.3231	0.2761	0.3700	0.3511	-0.0344	0.2325
EV ₂	2.0327	0.1325	0.1982	-0.1040	0.0797	-0.1662	-0.3272	0.4463	-0.0811	-0.3754	0.5076	0.4325
EV ₃	1.6130	0.3977	0.3792	-0.3863	0.1994	0.4400	-0.2351	-0.0494	-0.4149	0.0399	-0.2671	-0.1127
EV ₄	1.3324	-0.3258	-0.3264	-0.0661	0.1609	0.5308	-0.0883	0.4125	-0.0346	0.2243	0.3698	-0.3304
EV ₅	0.7012	-0.3031	-0.3649	-0.0038	0.4823	0.0892	-0.2554	0.0007	0.0262	-0.1527	-0.4877	0.4553
EV ₆	0.3632	-0.1889	-0.0020	-0.4431	0.2382	-0.3464	0.5902	0.2017	-0.3929	0.1922	0.0021	0.1045
EV ₇	0.2028	0.1694	-0.2666	-0.1584	-0.3277	0.5167	0.4131	-0.2070	0.0001	-0.1785	0.1543	0.4812
EV ₈	0.1530	-0.6436	0.6193	0.1545	-0.0556	0.2613	0.1598	-0.0673	-0.0454	-0.2680	-0.0306	0.0464
EV ₉	0.1036	-0.1857	0.0922	0.0990	-0.2484	-0.0247	-0.3172 ‡	-0.1713	-0.2595	0.7010	0.1246	0.4250
EV ₁₀	0.0872	-0.1372	0.1696	-0.6644	-0.0449	-0.0180	-0.1228	-0.0694	0.6772	0.1668	0.0080	0.0471
EV ₁₁	0.0022	-0.0018	-0.0015	-0.0160	-0.5579	0.0248	-0.0058	0.6556	0.0100	0.0276	-0.5049	0.0472

† PH, plant height; EH, ear height; EL, ear length; ED, ear diameter; NRE, number of rows per ear; NKR, number of kernels per row; CD, cob diameter; CL, cob length; TNK, total number of kernels per ear; CD/ED, cob diameter/ear diameter ratio; TKW, thousand-kernel weight.

‡ Component of the eigenvectors in bold indicate multicollinearity-generating traits.

Table 5. Eigenvalues and component of the eigenvectors of Pearson product-moment correlation matrix among the 11 explanatory traits estimated with the average values of each plot, $n = 180$.

Eigenvalues (EV)	Components of the eigenvectors (weight of the traits)											
	PH	AE	EL	ED	NRE	NKR	CD	CL	TNK	CD/ED	TKW	
EV ₁	5.0943	0.3508	0.3399	0.3420	0.3860	0.1994	0.2996	0.2677	0.3343	0.3245	-0.0502	0.2733
EV ₂	2.2155	0.0231	0.0921	-0.0525	0.0967	-0.1813	-0.3326	0.4622	-0.0198	-0.3737	0.5664	0.4011
EV ₃	1.5000	0.2380	0.2121	-0.4406	0.2188	0.5670	-0.2783	0.1193	-0.4637	0.0990	-0.0707	-0.1157
EV ₄	1.0560	-0.4556	-0.4598	0.1082	0.1477	0.4075	0.0286	0.3368	0.1647	0.2457	0.3174	-0.2810
EV ₅	0.5208	-0.1977	-0.3568	0.0595	0.4007	0.0945	-0.3436	-0.0644	0.1067	-0.1983	-0.5488	0.4319
EV ₆	0.3047	-0.2272	-0.1335	-0.4360	0.2453	-0.3111	0.5640	0.1727	-0.3440	0.1932	-0.0432	0.2795
EV ₇	0.1053	-0.1032	0.2138	0.1197	0.4309	-0.5081	-0.2706	0.2666	-0.0728	0.0896	-0.1942	-0.5387
EV ₈	0.0801	0.0171	0.1019	-0.4464	0.1674	0.1179	0.2910	0.0713	0.5108	-0.5577	-0.0885	-0.2802
EV ₉	0.0744	0.4875	-0.4208 ‡	-0.4069	-0.0306	-0.2479	-0.2519	0.0150	0.3608	0.3994	0.0612	-0.0120
EV ₁₀	0.0475	-0.5224	0.4912	-0.3188	-0.1413	0.0016	-0.2526	-0.1151	0.3458	0.3600	-0.0064	0.1950
EV ₁₁	0.0015	0.0086	-0.0079	-0.0040	-0.5627	0.0239	0.0054	0.6807	0.0046	-0.0051	-0.4676	0.0230

† PH, plant height; EH, ear height; EL, ear length; ED, ear diameter; NRE, number of rows per ear; NKR, number of kernels per row; CD, cob diameter; CL, cob length; TNK, total number of kernels per ear; CD/ED, cob diameter/ear diameter ratio; TKW, thousand-kernel weight.

‡ Components of the eigenvectors in bold indicate multicollinearity-generating traits.

Table 6. Multicollinearity diagnosis of Pearson product-moment correlation matrices among the 11 explanatory traits estimated in two data arrangement scenarios and three path analysis methodologies.

Multicollinearity diagnosis	All sampled observations (ASO)			Average value of each plot (AVP)		
	Traditional	With k inclusion	Traditional, excluding traits†	Traditional	With k inclusion	Traditional, excluding traits‡
Condition number	2004.000	85.404	29.621	3395.200	51.175	63.656
Matrix determinant	3.024E-6	3.740E-1	1.2E-2	1.260E-7	5.390E-4	7.938E-4
Number of VIFs > 10§	3	0	0	3	0	0
Largest VIF	195.582	9.907	4.022	320.825	6.445	8.756
Trait	cob diameter	cob diameter	ear diameter	cob diameter	cob diameter	kernel number per ear
Multicollinearity	Severe	Weak	Weak	Severe	Weak	Weak
Largest correlation	0.906**	1.050**	0.837**	0.925**	1.100**	0.924**
Smallest correlation	0.010ns¶	0.055ns	-0.034ns	0.012ns	0.081ns	0.011ns

** Meaningful at 0.01 probability level.

† Traits excluded in this scenario: Cob diameter, cob length and number of kernels per row.

‡ Traits excluded in this scenario: Cob diameter, plant height and ear height.

§ VIF, variance inflation factors.

¶ ns, non-meaningful at 0.01 probability level.

Table 7. Direct effects for the 11 explanatory traits on kernel weight per ear with the regression estimators estimated in two data arrangement scenarios and three path analysis methodologies.

Traits†	All sampled observations (ASO)					Average values of each plot (AVP)		
	r ‡	Direct effects			r	Direct effects		
		Traditional	With k inclusion	Traditional, excluding traits§		Traditional	With k inclusion	Traditional, excluding traits¶
PH	0.515**	0.015	0.039	0.012	0.716**	0.019	0.081	-
AE	0.461**	-0.027	-0.024	-0.040	0.671**	-0.003	0.017	-
EL	0.685**	-0.011	-0.018	-0.056	0.679**	-0.071	0.010	-0.073
ED	0.753**	2.379	0.152	-0.068	0.835**	0.476	0.128	-0.019
NRE	0.278**	-0.089	0.020	0.005	0.422**	0.037	0.058	0.061
NKR	0.657**	0.040	0.093	-	0.694**	0.106	0.187	0.111
CD	0.469**	-2.902	-0.132	-	0.525**	-0.602	0.046	-
CL	0.698**	0.016	0.048	-	0.672**	0.041	0.045	0.041
TNK	0.736**	0.763	0.636	0.892	0.760**	0.651	0.359	0.653
CD/ED	-0.114**	2.241	0.110	0.014	-0.182**	0.403	-0.040	-0.013
TKW	0.575**	0.536	0.564	0.733	0.634**	0.615	0.389	0.645
k value		-	0.050	-		-	0.100	-
R ²		1.02	0.931	0.977		0.973	0.922	0.973
Residual Effect		0	0.261	0.150		0.161	0.278	0.165

** Meaningful at 0.01 probability level.

† PH, plant height; EH, ear height; EL, ear length; ED, ear diameter; NRE, number of rows per ear; NKR, number of kernels per row; CD, cob diameter; CL, cob length; TNK, total number of kernels per ear; CD/ED, cob diameter/ear diameter ratio; TKW, thousand-kernel weight.

‡ r , linear correlation coefficient with kernel weight per ear.

§ Traits excluded in this scenario: cob diameter, cob length and number of kernels per row.

¶ Traits excluded in this scenario: cob diameter, plant height and ear height.

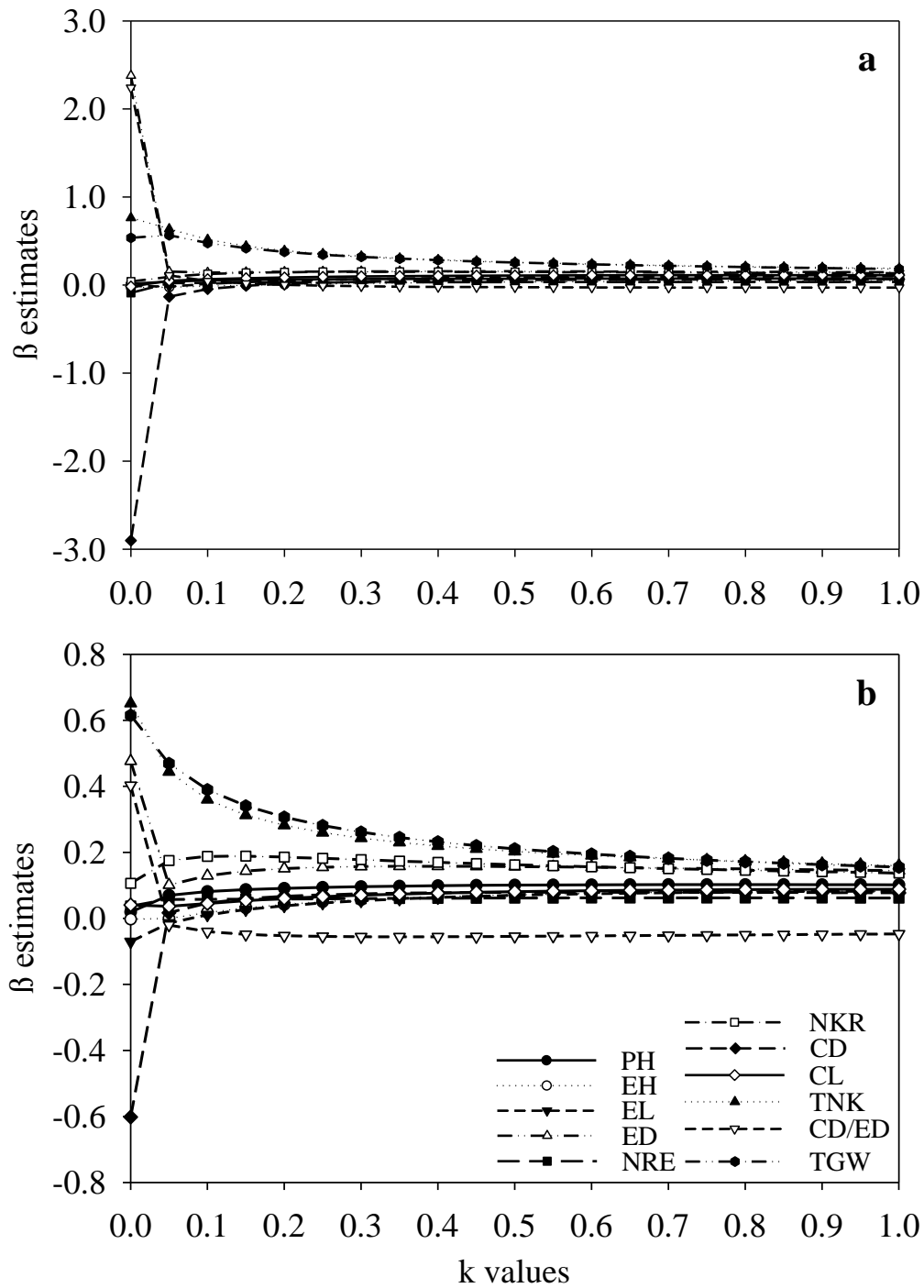


Fig. 1. β values for plant height (PH), ear height (EH), ear length (EL), ear diameter (ED), number of rows per ear (NRE), number of kernels per row (NKR), cob diameter (CD), cob length (CL), total number of kernels per plant (TNK), cob diameter/ear diameter ratio (CD/ED), and thousand-kernel weight (TKW), obtained with 21 k values, where β estimated with $k = 0$, matches to the estimations of least squares. Estimations performed for ASO (a) and AVP (b) scenarios.

**4 ARTIGO III - OPTIMAL SAMPLE SIZE AND DATA ARRANGEMENT METHOD
IN ESTIMATING CORRELATION MATRICES WITH LESSER COLLINEARITY: A
STATISTICAL FOCUS IN MAIZE BREEDING**

Submetido para o periódico: African Journal of Agricultural Research.

Situação: publicado.

DOI: 10.5897/AJAR2016.11799.

**Optimal sample size and data arrangement method in estimating
correlation matrices with lesser collinearity: a statistical focus in maize
breeding**

Tiago Olivoto^{1*}, Maicon Nardino², Ivan Ricardo Carvalho³, Diego Nicolau Follmann⁴,
Mauricio Ferrari³, Alan Junior de Pelegrin³, Vinicius Jardel Szareski⁵, Antônio Costa de
Oliveira³, Bráulio Otomar Caron¹ Velci Queiróz de Souza⁶

¹ Department of Agronomic and Environmental Sciences, Federal University of Santa Maria
Frederico Westphalen, Rio Grande do Sul, Brazil.

² Department of Mathematics and Statistics, Federal University of Pelotas, Capão do Leão,
Rio Grande do Sul, Brazil.

³ Plant Genomics and Breeding Center, Federal University of Pelotas, Capão do Leão, Rio
Grande do Sul, Brazil.

⁴ Agronomy Department, Federal University of Santa Maria, Santa Maria, Rio Grande do Sul,
Brazil.

⁵ Department of Crop Science, Federal University of Pelotas, Capão do Leão, Rio Grande
do Sul, Brazil.

⁶ Federal University of Pampa, Dom Pedrito, Rio Grande do Sul, Brazil.

*Corresponding author. E-mail: tiagoolivoto@gmail.com

4.1 ABSTRACT

Information about data arrangement methodologies and optimal sample size in estimating the Pearson correlation coefficient (r) among maize traits are still limited. Furthermore, some data arrangement methodologies currently used may be increasing multicollinearity in multiple regression analysis. This study aimed to investigate the statistical behavior of the r and the multicollinearity of correlation matrices among maize traits in different data arrangement scenarios and different sample sizes. Data from 45 treatments [15 simple maize hybrids (*Zea mays* L.) conducted in three locations] were used. Eleven traits were accessed and three datasets (scenarios) were formed: 1) Coming from all the sampled observations (plants), $n = 900$; 2) Coming from the average of five plants per plot, $n = 180$; and 3) Coming from the average of treatments, $n = 45$. A thousand estimates of r were held in each scenario to 60 sample sizes by bootstrap simulations with replacement. Confidence intervals (CI) were estimated. One hundred eighty correlation matrices were estimated and the condition number (CN) calculated. Data coming from average values of plots and average values of treatments overestimates the r up to 24 and 34%, resulting in an increase of 24 and 131% in the matrices' CNs, respectively. Trait pairs with high r require a smaller number of plants, being the CI inversely proportional to the magnitude of the r . Two hundred and ten plants are sufficient to estimate the r in the CI of 95% < 0.30 .

Abbreviations: ASO, all sampled observations; AVP, average values of plot; AVT, average values of treatments; CD, cob diameter; CD/ED, cob diameter/ear diameter ratio; CL, cob length; ED, ear diameter; EH, ear height; EL, ear length; NKR, number of kernels per row; NRE, number of rows per ear; PH, plant height; TKW, thousand-kernel weight; TNK, total number of kernels per ear.

Keywords: average values, bootstrap, confidence intervals, sample tracking, *Zea mays* L.

4.2 INTRODUCTION

One of the most used statistical methods to measure the degree of association (linear) between two random traits is the Pearson product-moment correlation coefficient (r) (Pearson, 1920) and has been used in ecological studies to estimate the direction and degree of association among traits (Annicchiarico et al. 1999, Yao and Mehlenbacher 2000, Yang and Su 2016).

As this measure only reveals the linear association between two traits, techniques such as path analysis (Wright 1923) and canonical correlation (Hotelling 1936) were developed in order to explain the interrelationships among traits or group of traits, being worldwide used in plant breeding. These techniques depend on the linear correlation matrix among traits and, due its estimates be based on principles of multiple regression, the low dependence among the traits considered as explanatory is required. When this assumption is not met, it is said that the matrix presents multicollinearity (Blalock 1963).

Although there are techniques to adjust the multicollinearity (Hoerl and Kennard 1970b) these techniques are essentially correctives, applied only after the linear correlation matrix be estimated. Since the estimates of correlation coefficients basically involve the behavior analyses of the variances, i.e., deviations from the average, it is possible that some methods of data arrangement currently used may be masking the actual averages and variances of a trait (x) on a dataset of (n) observations. For example, in a brief survey, we found that the correlation matrices of some agronomic studies using path analysis, were estimated with average values of several plants sampled in each experimental unit (Khameneh et al., 2012, Toebe and Cargnelutti 2013, Adesoji et al., 2015, Kumar and Babu 2015., Nataraj et al., 2015).

In field experiments, it is very common to access values of traits in several plants of each experimental unit. The utilization of average value of these plants in order to estimate the r and perform inferences to the population under study, however, may be questionable. In a theoretical explanation focused on plant breeding, Olivoto et al. (2016) reported that the use of

average values in estimating the r between a traits pair (e.g. $r_{x,y}$) may overestimate its magnitude mainly due the reduction of standard deviation (SD) in the dataset, when compared with estimates performed with values coming from all sampled plants. In addition, the observed SD (e.g., for X and Y) when average values of plots or treatments are used, represents the SD of the average of the originally sampled plants, and not the actual SD coming from all these plants; therefore, this SD is masked and tends to present itself lower. This fact should be taken into consideration, because the inference of the direction and magnitude of association among traits when average values are used, is being made for a different population of the original.

There were no studies in the literature comparing different data arrangement methodologies on estimates of Pearson's correlation coefficients. In addition, the information about the optimal sample size to estimate the r among trait pairs in the maize crop in an acceptable confidence interval is needed. In this context, the aims of the present study were to (i) reveal the statistical behavior of estimated Pearson's correlation coefficients in different data arrangement scenarios and different sample sizes, (ii) reveal the impact of data arrangement scenarios and sample sizes on multicollinearity of matrices, and (iii) propose the optimal sample size to estimate r among trait pairs in the maize crop in an acceptable confidence interval.

4.3 MATERIALS AND METHODS

4.3.1 Site description and experimental design

Field trials were conducted in 2014/2015 growing season in Santo Expedito do Sul (27°56' S, 51°37' W; 728 m above sea level), São José do Ouro (27°44' S, 51°32' W; 796 m above sea level) and Viadutos (27°33' S, 52°00' W; 628 m above sea level), municipalities of northeast region of Rio Grande do Sul State, Brazil. During the experimental period, the air averages temperatures at the sites of the experiments were 24.5, 23.8 and 25.2°C and rainfall of 823, 958 and 746 mm, respectively. All locations are within a 70-km radius, have a Haplustox

soil, and were chosen due to similarities of soil and climatic characteristics, which provided to them low variability of temperature and rainfall. Thus, abiotic effects on the plants' response were minimized as much as possible.

Prior to the installation of the trials, each site was surveyed for potentially disruptive characteristics. To ensure uniformity inside the block and heterogeneity between the blocks, a randomized complete block design in a 15×3 factorial treatment design (15 simple maize hybrids \times three cropping fields) with four replications was used, totaling 180 plots. Each plot contained six 5-m-long cultivar rows, spaced by 0.45 m. Only the two central rows were used to prevent edge effects. In each plot, five representative plants (observations) were selected from which the ear was removed for further evaluation. To ensure that traits (of plant and ear) were assessed in the same individual, a sample tracking system was created, identifying each ear with a label containing a sequence number that characterized the site, the hybrid, the repetition and the evaluated plant.

4.3.2 Accessed traits

Plant height (PH) and the ear insertion height (AE) were measured (cm) from the ground surface to the flag leaf node and the support node of the highest ear at the stem, respectively. Tagged ears were evaluated at a laboratory. The following traits were accessed: ear length (EL) (cm), ear diameter (ED) (cm), number of rows per ear (NRE) (un), number of kernels per row (NKR) (un), cob length (CL) (cm), cob diameter (CD) (mm), cob diameter / ear diameter ratio (CD / ED) (decimal), total number of kernels per ear (TNK) (un) the thousand-kernel weight (TKW) (g). The ratings were performed as follows: the lengths and diameters were measured with a digital caliper. After counting the number of rows per ear and the number of kernels per row, the kernels of each ear were manually-threshed and cleaned with pressurized air. Subsequently, the kernels-weight was measured with an analytical balance and the total number

of kernel each ear was measured with a seed counter equipment. Finally, the grain moisture was measured with a universal moisture meter. With this data, and with the humidity adjusted to 14% base moisture, we determined the thousand-kernel weight each ear by the equation: $TKW = [(KME/TNK) \times 1000]$. Where: TKW = Thousand kernel weight; KME = Kernel mass per ear; TNK = the total number of kernels per ear. All evaluations were carried out carefully in an ear at a time, to maintain traceability of the sample, avoid any systematic errors as well as minimize the random errors.

4.3.3 Statistical procedures

4.3.3.1 Bootstrap simulations

Three data arrangement scenarios were considered: (i) the data used were originated from all sampled observations (ASO), with a total sample size of 900; (ii) in this scenario, the data used were obtained from the average of the five sampled plants of each plot (AVP), with a total sample size of 180; and (iii) finally, the average of the treatments (AVT), with a total sample size of 45 (15 treatments \times 3 locations) was considered.

Aiming to match the sample size in each scenario, 60 sample sizes (plants) were simulated. The size of the initial sample was 15 plants, and the rest were obtained with an increment of 15 plants up to 900 plants. For each one of 55 trait pairs $[n \times (n-1)]/2$, where $n = 11$, in each sample size of each scenario, 1000 simulations of the r were performed by bootstrap resampling with replacement (Efron 1979). Thus, for each pair of traits, 1000 estimates of the r were obtained. Simulations were performed by the Structural Equation Modeling procedure in Statistica 8.0 software (Weiß, 2007).

4.3.3.2 Descriptive analysis of correlation coefficients

In each sample size of each scenario, the 1000 simulated r were subjected to descriptive analysis, where it was determined the maximum, (97.5%), average, (2.5%) and minimum values. Later, the amplitude of the 95% confidence interval was calculated by the difference between the percentile 97.5% and 2.5%. For comparison, three trait pairs that came closest to the following r magnitudes were chosen: $r \approx |0|$, $r \approx |0.5|$ and $r \approx |1.0|$. The statistics mentioned of these three trait pairs has formed scatter diagrams where the x-axis corresponding to the number of plants and the y-axis corresponding to the descriptive statistics.

4.3.3.3 t -test to compare the correlation coefficient among the scenarios

In order to determine whether the inferences could be made with the average of 60 sample sizes, initially the r average of each traits pair at the different sample sizes were compared by t -test at 5% probability error (Steel et al. 1997) in the following scenario combinations: ASO \times AVT, ASO \times AVP and AVP \times AVT. Inferences were made using the average of sample sizes for each pair of trait if the 60 sample size presented the same result on the test.

A test comparing the 3300 values of r (55 trait pairs \times 60 sample size) was also performed. Histograms were developed for each scenario combination (ASO \times AVT, ASO \times AVP and AVP \times AVT) in order to show the behavior of the estimated r distribution. These procedures were performed using *t.test* and *hist* functions in R software (R core Team, 2016). Descriptive statistics such as asymmetry, average, mode, 25th and 75th percentiles, maximum, and minimum applied in each scenario are also presented in boxplot graphics. These procedures were performed using *summary* and *boxplot* functions in R software.

4.3.3.4 Diagnosis of multicollinearity in the scenarios

Data of 11 traits obtained by the average of 1000 bootstrap simulations in each sample size of each scenario were used to estimate correlation matrices. A total of 180 matrices (60 sample size \times three scenarios) were estimated. In each matrix, multicollinearity diagnosis was performed by the condition number (CN) of the matrix. The CN was obtained by the ratio between the largest and the smallest eigenvalue of the matrix. The degree of multicollinearity was considered weak, moderate and severe when $CN \leq 100$, between 100 and 1000 and ≥ 1000 , respectively (Mansfield and Helms 1982). A graph containing the number of plants (x axis) and the CN of each scenario (y axis) was developed. This analysis was performed using the Multicollinearity Diagnostic procedure in Genes software (Cruz 2013).

4.4 RESULTS

4.4.1 Statistical properties of the correlation coefficient

The estimated r presented the largest amplitude when the lowest number of plants was used. For the pair AE \times PH, the magnitude of r oscillates between -0.02 and 0.98 (Fig. 1a), 0.42 to 0.99 (Fig. 1b) and 0.71 to 0.99 (Fig. 1c) in ASO, AVP, and AVT scenarios, respectively. This range was reduced as the number of plants increased, however, it appeared higher in the ASO scenario. The average r between the 60 different numbers of plants evaluated was increased by approximately 11% ($r = 0.92$) and 15% ($r = 0.96$), in AVP and AVT scenarios, respectively (Fig. 1b,c).

For trait the pairs with $r \approx |0.5|$ as NKR \times ED, the amplitude of r was larger, irrespectively of the scenario and the number of assessed plants. With 15 plants, r ranged between -0.33 and 0.89 in the ASO scenario (Fig. 1d), between 0.62 and 0.91 in the AVP scenario (Fig. 1e) and between 0.03 and 0.90 in the AVT scenario (Fig. 1f). The average r was increased by approximately 16% ($r = 0.58$) and 24% ($r = 0.62$), in AVP and AVT scenarios,

respectively. Trait pairs with $r \approx |0|$ as DSDE \times CE presented the highest amplitudes, with similar r distribution in the studied scenarios (Fig. 1g–i).

For the pair PH \times AE, 270 plants were enough to estimate the r in the ASO scenario in the CI $95\% \leq 0.10$ (Fig. 2a). For AVP and AVT scenarios, however, the number of plants needed was only 45 (Fig. 2b) and 30 (Fig. 2c), respectively. Trait pairs with $r \approx |0.5|$ (NKR \times ED), needed 660, 465 and 285 plants, in ASO, AVP, and AVT scenarios, respectively. For CD/ED \times EL combination, CI $95\% \leq 0.10$ was not reached even with 900 plants.

4.4.2 Comparison of correlation pairs between the scenarios

The t -test revealed no differences among the sample sizes in all scenario combinations. Thus, the inferences for each pair of traits were performed with the average of 60 sample sizes. Among the 165 comparisons (55 trait pairs in three scenario combinations), 164 differed. Only one did not differ. In approximately 82% of the cases, average values (AVT and AVT scenarios) overestimated the magnitude of the r (Table 1).

Comparing the estimated r in ASO \times AVT scenarios, of 55 tested pairs, ten (18%) had a higher average when all sampled observations were used (Table 1). Comparing ASO \times AVP scenarios, only seven combinations (13%) had a higher average r in correlation analysis estimated with all observations. Comparing the averages (AVP \times AVT), 12 combinations (22%) were higher when the average of the plots was used (Table 1).

A t -test comparing the average r of 55 trait pairs in ASO \times AVT scenario combination confirmed the difference between these (t -value = -12.89 , $P < 0.001$). The average r with low magnitudes are due to the use of all pairs of correlation, where there are positive and negative values. The estimates in the ASO scenario showed a distribution similar to normal. That is related to the low asymmetry value (0.009), smaller r amplitude (-0.273 to 0.912), and the median value (0.268) that is similar to the average value (0.282), although the tests reject the

hypothesis of normality (Kolmogorov-Smirnov = 0.048, $P = < 0.01$) (Fig. 3). The estimates carried out in the AVT scenario, however, shows a negative asymmetrical distribution of r values (-0.843), with a greater r amplitude (-0.552 to 0.956) and the median value (0.484), higher than the average (0.379). The distribution of r values in this scenario do not follow the normal distribution (Kolmogorov-Smirnov = 0.137 , $P = < 0.01$) (Fig. 3).

The comparison of ASO \times AVP scenarios shows a behavior similar to that discussed above, though with a slightly smaller difference (t -value = -9.60 , $P < 0.0001$). For the AVP scenario, r also presented negative asymmetry (-0.566). The amplitude was also lower (-0.427 to 0.926), with a median value (0.399) higher than the average (0.350) (Fig. 4). The distribution in this scenario was not normal (Kolmogorov-Smirnov = 0.136 , $P < 0.01$).

The t -test comparing the average r between the AVP \times AVT scenarios combinations, revealed difference (t -value = -3.73 , $P < 0.001$). With the measures of central tendency and amplitudes of these scenarios discussed above, both showed non-normal distribution of r , with a clear tendency of most of the observed values being higher than r average (Fig. 5).

The r was increased by approximately 24% and 34% in the AVP and AVT scenarios, respectively. In addition, the r amplitude and standard deviation were higher in these scenarios (Fig. 6).

4.4.3 Multicollinearity

Multicollinearity was considered severe for the three scenarios, regardless of the number of assessed plants (Fig. 7). The use of averages (AVP and AVT scenarios) increased the CN of the correlation matrices. The largest changes occurred when the number of plants was low (< 100). For example, with 45 and 60 plants, the CN increased by 118% and 75% for the AVP scenarios and 250% and 68% for the AVT scenario, respectively. Although in some cases the

CN was higher for the ASO scenario, on mean, CN was increased by 24% and 131% in AVP and AVT scenarios, respectively (Fig. 7).

4.5 DISCUSSION

The reduction of individual variation (standard deviation) observed in the scenarios AVP and AVT was the main factor responsible for overvaluing the r of trait pairs. This fact can be explaining due standard deviation be the divisor on correlation formulae. If covariance XY (dividend of formulae) is similar in both scenarios, however, the standard deviation of X and Y traits (divisor of formulae) is smallest, the magnitude of correlation coefficients will be greater.

The higher number of plants required for estimation of the r at the 95% $CI \leq 0.10$ in trait pairs with less intensity of linear association, shows that the researcher must take into consideration the magnitude of the trait pairs, and the confidence interval will be inversely proportional to the magnitude of its correlations. The magnitude of the CI used here (95% $CI \leq 0.10$) it is not a rule, being it will be up to each researcher adopt the appropriate confidence level for its inferences. If we consider the CI 95% $CI < 0.30$, 210 plants are enough for estimating trait pairs with low magnitude ($r < 0.10$). This number of plants it is perfectly possible of to be evaluated. The experimental design (number of treatments and repetitions) will set so, the number of plants to be sampled in each plot. In experiments with large numbers of experimental units (e.g., factorial designs), the increase in sample size will provide greater confidence in the estimates provided that they are properly followed the sampling procedures and maintained traceability of the samples.

Although for trait pairs with high linear association ($AE \times PH$) AVP and AVT scenarios needed 83% and 89% fewer plants to estimate r , the average r in these scenarios was increased by 11% and 15%, respectively, compared to the ASO scenario ($r = 0.83$). In an analysis that depends on of the linear correlation matrix for their estimates, e.g., canonical correlation, path

analysis and stepwise multiple linear regression procedures, high linear association magnitudes among explanatory traits make it difficult to analyze, threatening the statistic and the inferential interpretation (Graham 2003).

A recent study revealed that multicollinearity begins to seriously distort the estimates of the path coefficients when the explanatory traits show $r > |0.7|$ (Dormann et al. 2013). While there have been observed high correlations in the ASO scenario (e.g., AE \times PH, $r = 0.83$), the higher values for the same pair ($r = 0.92$) and ($r = 0.96$) estimated in APV and AVT scenarios, respectively, demonstrated that these data arrangement methodologies overestimate the magnitude of the r and may result in larger problems in estimates of multiple regression parameters, leading to an erroneous interpretation of predictors in a statistical model. Thus, these methods should be carefully evaluated by the researchers when the goal is to use the correlation matrix in studies involving multiple regression, as for this, the independence or the less degree of dependence among explanatory traits is sought (Prunier et al., 2014; Montgomery et al., 2015).

Average values (AVP and AVT scenarios), visibly elevated the multicollinearity of the matrices, confirming the earlier discussion. Although there are variations in CN in each studied sample size, the multicollinearity was increased on average by 24% and 131% when the AVP and AVT scenarios were considered in the estimation of correlation matrices. Although there are techniques for adjusting the multicollinearity as to delete the traits responsible for inflating the variance of the coefficients (Gunst and Mason 1977) or to perform estimates using equations partially modified by the inclusion of a k constant in the diagonal elements of correlation matrix (Hoerl and Kennard 1970a), these techniques can mask the true biological behavior's response, because the deletion of the traits can reduce the model's explanation power. The inclusion of the k constant is effective in reducing the magnitude of multicollinearity, however, also causes a bias in the regression analysis (Hoerl and Kennard 1970a).

The best strategy to mitigate the problems caused by multicollinearity is to reduce it since it becomes practically impossible to eliminate it. In this research, a simple method for mitigating the multicollinearity in correlation matrices is suggested: estimating the correlation coefficients considering all observations, keeping traceability and individual variance of the sample. This can be accomplished without significant increase of time, labor and financial resources since, a priori, all sampled plants were assessed.

4.6 CONCLUSION

Estimates made with data based on averages (AVP and AVT scenarios) reduce the individual variances, overestimate the correlation coefficients and increase the multicollinearity in correlation matrices. Thus, studies that require explanatory traits in order to predict a dependent trait will present greater misstatements in the estimates of the regression coefficients, if these methods are used. Using values coming from all sampled plants, 210 plants are enough for estimating Pearson product-moment correlation coefficients among maize traits. The current study about data arrangement on Pearson's correlation coefficients presents useful information on the planning of future experiments in plant breeding involving biometric templates that require the correlation matrix for their estimates.

Conflict of Interests

The authors have not declared any conflict of interests.

4.7 ACKNOWLEDGMENT

We thank the Higher Coordination for the Improvement of Higher Education Personnel (CAPES) and the National Council for Scientific and Technological Development (CNPq) for granting to master's scholarship and research productivity's scholarship. We also are grateful

to the colleagues Amanda Baseggio and Jaksson Klin for their valuable collaboration in conducting the field trials.

4.8 REFERENCES

- Adesoji AG, Abubakar IU, Labe DA (2015). Character association and path coefficient analysis of maize (*Zea mays* L). grown under incorporated legumes and nitrogen. *J. Agron.* 14(3):158-163.
- Annicchiarico P, Piano E, Rhodes I, 1999: Heritability of, and genetic correlations among, forage and seed yield traits in Ladino white clover. *Plant Breeding* 118(4):341–346.
- Blalock HM (1963). Correlated independent variables: the problem of multicollinearity *Social Forces* 42(2):233-237.
- Cruz CD (2013). GENES - a software package for analysis in experimental statistics and quantitative genetics. *Acta Sci., Agron.* 35(3):271–276.
- Dormann CF, Elith J, Bacher S, Buchmann C, Carl G, Carré, G, Marquéz JRG, Gruber B, Lafourcade B, Leitão PJ, Münkemüller T, McClean C, Osborne PE, Reineking B, Schröder B, Skidmore AK, Zurell D, Lautenbach S (2013). Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography* 36(1):27-46.
- Efron B (1979). Bootstrap methods: another look at the jackknife. *Ann. Statist.* 7(1):1-26.
- Graham MH (2003). Confronting multicollinearity in ecological multiple regression. *Ecology* 84(11):2809-2815.
- Gunst RF, Mason RL (1977). Advantages of examining multicollinearities in regression analysis. *Biometrics* 33(1):249-260.
- Hoerl AE, Kennard RW (1970a). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* 12(1):55-67.
- Hoerl AE, Kennard RW (1970b). Ridge regression: applications to nonorthogonal problems. *Technometrics* 12(1):69-82.
- Hotelling H (1936). Relations between two sets of variates. *Biometrika* 28(3/4):321-377.
- Khameneh MM, Bahraminejad S, Sadeghi F, Honarmand SJ, Maniee M (2012). Path analysis and multivariate factorial analyses for determining interrelationships between grain yield and related characters in maize hybrids. *Afr. J. Agric. Res.* 7(48):6437-6446.
- Kumar SVV, Babu, DR (2015). Character association and path analysis of grain yield and yield components in maize (*Zea Mays* L). *Electronic J. Plant Breeding* 6(2):550-554.
- Mansfield ER, Helms BP (1982). Detecting multicollinearity. *Am. Stat.* 36(3):158-160.

- Montgomery DC, Peck EA, Vining GG (2012). Introduction to linear regression analysis 5th ed John Wiley & Sons New Jersey.
- Nataraj V, Shahi JP, Vandana D (2015). Character association and path analyses in maize (*Zea mays* L). Environ. Ecol. 33(1):78-81.
- Olivoto T, Nardino M, Carvalho IRC, Follmann DN, Szareski VJ, Ferrari M, Pelegrin AJ, Souza VQ (2016). Pearson correlation coefficient and accuracy of path analysis used in maize breeding: a critical review. Int. J. Curr. Res. 8(9):37787-37795.
- Pearson K (1920). Notes on the history of correlation. Biometrika 13(1):25-45.
- Prunier JG, Colyn M, Legendre X, Nimon KF, Flamand MC (2014). Multicollinearity in spatial genetics: separating the wheat from the chaff using commonality analyses. Mol. Ecol. 24(2):263-283.
- R core Team. 2016: R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna.
- Steel RGD, Torrie JH, Dickey D (1997). Principles and procedures of statistics: a biometrical approach 3rd ed McGraw-Hill New York NY, USA.
- Toebe M, Cargnelutti A (2013). Multicollinearity in path analysis of maize (*Zea mays* L). J. Cereal Sci. 57(3):453-462.
- Wei CH (2007). StatSoft, Inc., Tulsa, OK.: STATISTICA, Version 8. AStA Adv. Stat. Anal 91(3):339-341.
- Wright S (1923). The theory of path coefficients a reply to Niles's criticism Genetics 8(3):239-255.
- Yang H, Su, G (2016). Impact of phenotypic information of previous generations and depth of pedigree on estimates of genetic parameters and breeding values. Livest. Sci. 187: 61-67.
- Yao Q, Mehlenbacher SA (2000). Heritability, variance components and correlation of morphological and phenological traits in hazelnut. Plant Breeding 119(5):369-381.

Table 1. *t*-statistics for the average correlation coefficient (*r*) of 55 trait pairs estimated in 60 different numbers of plants. Average values represent 1000 bootstrap simulations of the original data coming from all sampled observations (ASO), coming from the average of each plot (AVP) and coming from the average of treatments (AVT). Coefficients in bold indicate the pairs in which *r* was lower with the use of averages.

Trait pairs	ASO × AVT			ASO × AVP			AVP × AVT		
	Average <i>r</i>		<i>t</i>	Average <i>r</i>		<i>t</i>	Average <i>r</i>		<i>t</i>
	ASO	AVT		ASO	AVP		AVP	AVT	
AE × PH	0.834	0.955	782.08**	0.834	0.955	-782.08**	0.925	0.955	484.84**
EL × PH	0.249	0.573	944.63**	0.249	0.414	-460.71**	0.414	0.573	547.30**
EL × AE	0.215	0.546	1079.40**	0.215	0.399	-559.79**	0.399	0.546	516.39**
ED × PH	0.478	0.750	910.05**	0.478	0.641	-559.57**	0.641	0.750	389.69**
ED × AE	0.458	0.712	805.67**	0.458	0.610	-558.02**	0.610	0.712	315.06**
ED × EL	0.417	0.513	205.64**	0.417	0.514	-113.86**	0.514	0.513	-1.62ns
NRE × PH	0.234	0.447	458.76**	0.234	0.360	-241.01**	0.360	0.447	155.68**
NRE × AE	0.160	0.346	550.12**	0.160	0.282	-290.51**	0.282	0.346	154.48**
NRE × EL	0.028	0.040	38.31**	0.028	0.082	-107.56**	0.082	0.040	-93.21**
NRE × ED	0.498	0.621	391.20**	0.498	0.578	-248.07**	0.578	0.621	224.23**
NKR × PH	0.234	0.568	1008.10**	0.234	0.402	-511.47**	0.402	0.568	534.93**
NKR × AE	0.206	0.519	942.17**	0.206	0.387	-482.23**	0.387	0.519	379.25**
NKR × EL	0.646	0.618	68.570**	0.646	0.659	-26.22**	0.659	0.618	-103.53**
NKR × ED	0.319	0.334	22.55**	0.319	0.394	-83.45**	0.394	0.334	-73.38**
NKR × NRE	0.067	0.092	58.63**	0.067	0.164	-124.62**	0.164	0.092	-95.29**
CD × PH	0.256	0.416	253.86**	0.256	0.376	-246.66**	0.376	0.416	59.46**
CD × AE	0.313	0.488	248.73**	0.313	0.439	-215.07**	0.439	0.488	66.05**
CD × EL	0.308	0.359	98.98**	0.308	0.351	-76.61**	0.351	0.359	14.34**
CD × ED	0.653	0.730	263.58**	0.653	0.729	-325.70**	0.729	0.730	2.64*
CD × NRE	0.269	0.259	23.550**	0.269	0.298	-59.45**	0.298	0.259	-69.79**
CD × NKR	0.069	0.086	294.860**	0.069	0.064	7.30**	0.064	0.087	-228.23**
CL × PH	0.222	0.485	555.35**	0.222	0.369	-339.33**	0.369	0.485	288.34**
CL × AE	0.170	0.449	552.07**	0.170	0.340	-388.77**	0.340	0.449	225.58**
CL × EL	0.908	0.936	171.82**	0.908	0.923	-69.04**	0.923	0.936	70.45**
CL × ED	0.430	0.479	79.17**	0.430	0.523	-106.13**	0.523	0.479	-56.79**
CL × NRE	0.023	0.003	53.69**	0.023	0.065	-82.83**	0.065	0.003	-152.69**
CL × NKR	0.639	0.592	136.66**	0.639	0.647	-26.52**	0.647	0.592	-160.26**
CL × CD	0.343	0.393	74.94**	0.343	0.391	-69.94**	0.391	0.393	3.76**
TNK × PH	0.303	0.642	998.75**	0.303	0.488	-556.69**	0.488	0.642	471.33**
TNK × AE	0.226	0.556	1051.70**	0.226	0.419	-564.68**	0.419	0.556	399.82**
TNK × EL	0.548	0.493	147.44**	0.548	0.540	11.17**	0.540	0.493	76.05**
TNK × ED	0.532	0.639	192.04**	0.532	0.594	-110.93**	0.594	0.639	70.25**
TNK × NRE	0.519	0.691	350.18**	0.519	0.625	-242.68**	0.625	0.691	121.83**
TNK × NKR	0.719	0.736	69.52**	0.719	0.777	-136.53**	0.777	0.736	110.79**
TNK × CD	0.191	0.116	180.57**	0.191	0.179	31.27**	0.179	0.116	144.95**
TNK × CL	0.535	0.428	274.75**	0.535	0.502	56.70**	0.502	0.428	148.45**
CD/ED × PH	-0.123	0.2273	235.70**	-0.123	-0.174	120.65**	-0.174	0.227	111.10**
CD/ED × AE	-0.034	0.0840	91.19**	-0.034	-0.051	33.37**	-0.051	0.084	65.52**
CD/ED × EL	0.002	0.0400	84.15**	0.002	-0.079	124.89**	-0.079	0.040	55.65**
CD/ED × ED	-0.121	0.0576	88.49**	-0.121	-0.078	-57.34**	-0.078	0.058	29.73**
CD/ED × NRE	-0.13	0.3127	210.45**	-0.130	-0.219	116.49**	-0.219	0.313	86.75**
CD/ED × NKR	-0.221	0.4987	717.16**	-0.221	-0.360	355.08**	-0.360	0.499	336.64**
CD/ED × CD	0.666	0.636	80.68**	0.666	0.620	120.02**	0.620	0.636	62.00**
CD/ED × CL	0.038	0.048	20.31**	0.038	-0.029	126.76**	-0.029	0.048	146.04**
CD/ED × TNK	-0.265	0.5475	504.87**	-0.265	-0.421	301.55**	0.421	0.547	190.86**
TKW × PH	0.405	0.638	505.66**	0.405	0.539	-329.62**	0.539	0.638	254.21**
TKW × AE	0.418	0.674	537.91**	0.418	0.553	-335.17**	0.553	0.674	286.18**
TKW × EL	0.364	0.594	591.29**	0.364	0.452	-214.08**	0.452	0.594	432.89**
TKW × ED	0.488	0.685	617.07**	0.488	0.623	-499.34**	0.623	0.685	214.27**

(Continua)

(Conclusão)

Trait pairs	ASO × AVT			ASO × AVP			AVP × AVT		
	Average r		<i>t</i>	Average r		<i>t</i>	Average r		<i>t</i>
	ASO	AVT		ASO	AVP		AVP	AVT	
TKW × NRE	-0.206	0.0141	626.90**	-0.206	-0.082	-223.54**	-0.082	0.014	130.71**
TKW × NKR	0.096	0.209	288.70**	0.096	0.14	-104.59**	0.14	0.209	159.98**
TKW × CD	0.482	0.738	564.25**	0.482	0.644	-472.08**	0.644	0.738	204.82**
TKW × CL	0.384	0.55	305.53**	0.384	0.471	-182.82**	0.471	0.55	164.41**
TKW × TNK	-0.102	0.13	528.63**	-0.102	0.013	-229.57**	0.013	0.13	277.06**
TKW × CD/ED	0.163	0.314	338.57**	0.163	0.236	-197.08**	0.236	0.314	172.82**

‘*’ and ‘**’ show the significances at 0.001 and 0.01 of probability level, respectively. ‘ns’ is not significant.

PH, Plant height; AE, ear height; EL, ear length; ED, ear diameter; NRE, number of rows per ear; NKR, number of kernels per row; CL, cob length; CD, cob diameter; CD/ED, cob diameter / ear diameter ratio; TNK, total number of kernels per ear; TKW, thousand-kernel weight.

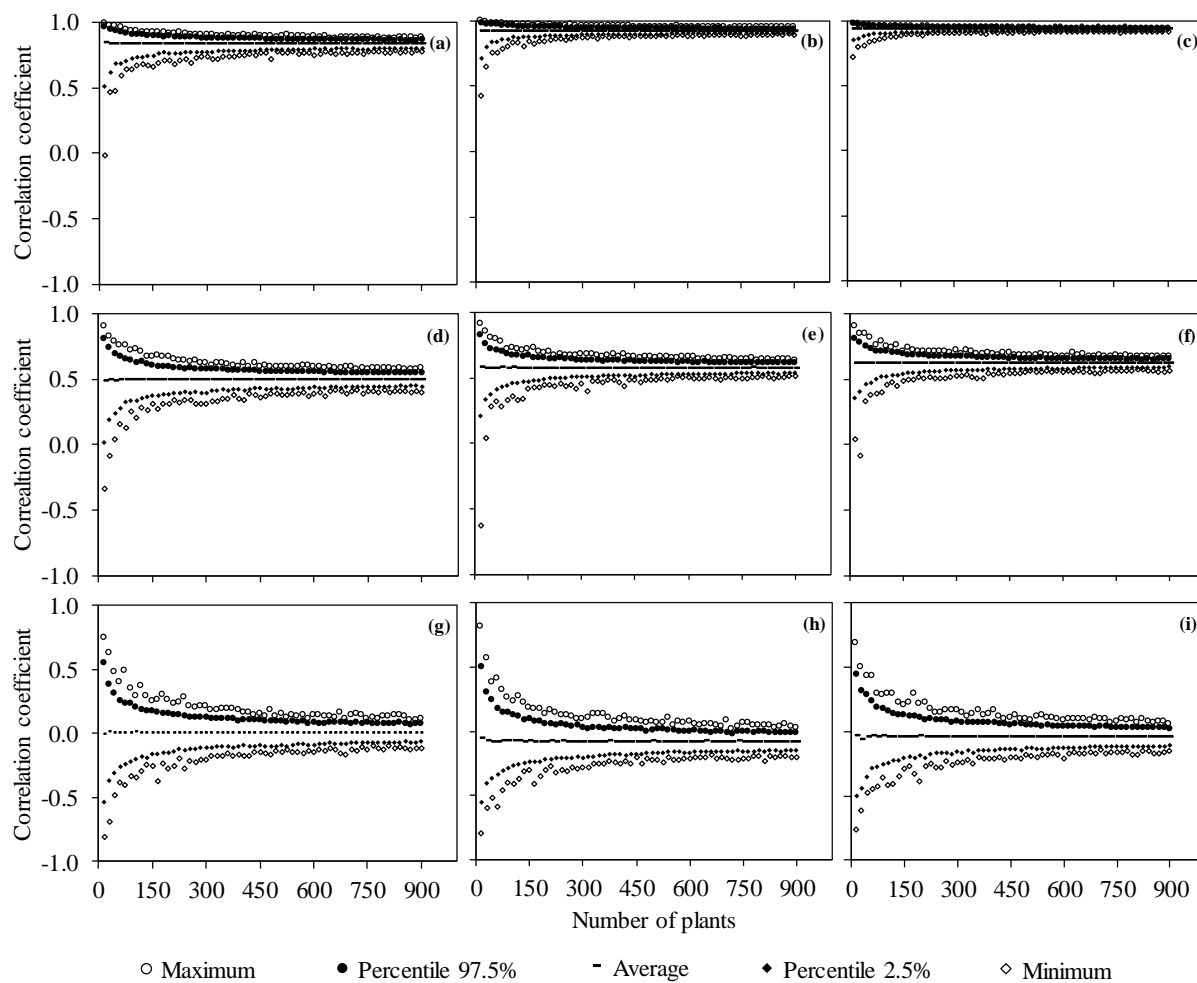


Figure 1. Descriptive analysis of 1000 bootstrap estimates of Pearson's correlation coefficient. Symbols represent the maximum values, percentile 97.5%, average, percentile 2.5% and minimum, obtained for the pair of traits plant height \times ear height estimated in ASO (a) in AVP (b) and AVT (c) scenarios; number of kernels row \times ear diameter estimated in ASO (d), AVP (e) and AVT (f) scenarios and cob diameter / ear diameter ratio \times ear length, estimated in ASO (g), AVP (h) and AVT (i) scenarios.

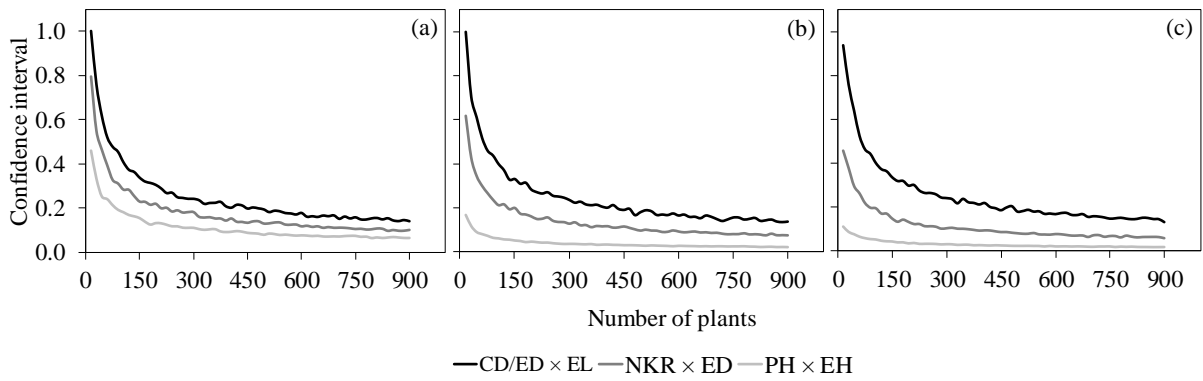


Figure 2. Confidence interval of the correlation coefficient for the confidence interval of 95%. (a) ASO scenario. (b) AVP scenario and (c) AVT scenario. Lines in greyscale represent the pair cob diameter / ear diameter ratio \times ear length (CD/ED \times EL), number of kernels rows \times ear diameter (NKR \times ED) and plant height \times ear height (PH \times EH).

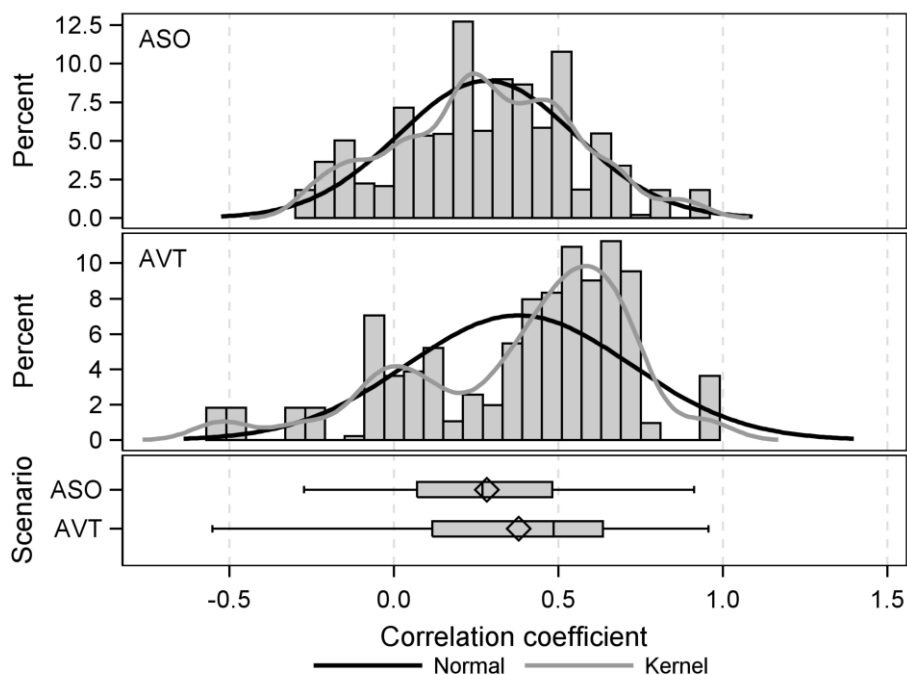


Figure 3. Distribution of average values of correlation coefficient in ASO \times AVT scenarios combination. Columns represent the observed values. Black and gray lines represent the normal distribution and Kernel density estimation, respectively. ASO and AVT scenarios represent the correlation coefficients estimated by all sampled observations, and by the average values of treatments, respectively. In the lower plot, the average (rhombus), the median (vertical line), the distance between the 25th and 75th percentiles (length of the box) and the maximum and minimum values (outer spread) of the estimated correlation coefficient are presented for each scenario.

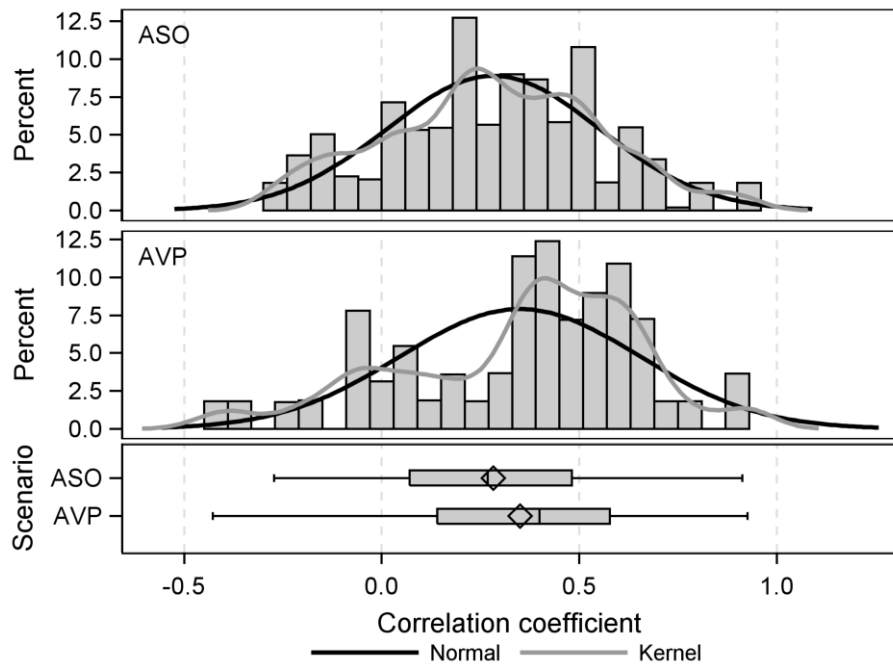


Figure 4. Distribution of average values of correlation coefficient in ASO \times AVP scenarios combination. Columns represent the observed values. Black and gray lines represent the normal distribution and Kernel density estimation, respectively. ASO and AVP scenarios represent the correlation coefficients estimated by all sampled observations, and by the average values of plots, respectively. In the lower plot, the average (rhombus), the median (vertical line), the distance between the 25th and 75th percentiles (length of the box) and the maximum and minimum values (outer spread) of the estimated correlation coefficient are presented for each scenario.

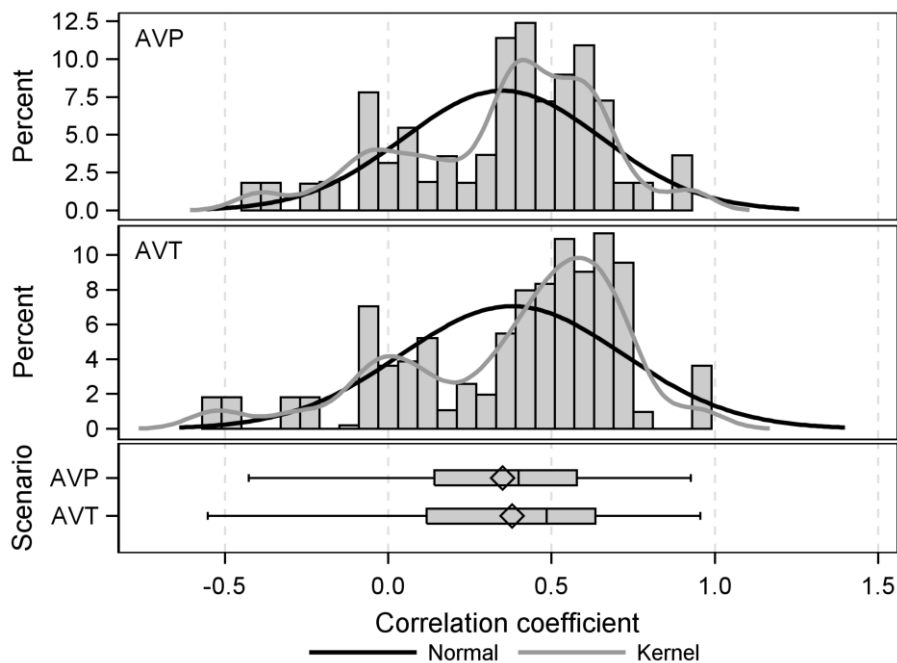


Figure 5. Distribution of average values of correlation coefficient in AVP \times AVT scenarios combination. Black and gray lines represent the normal distribution and Kernel density estimation, respectively. AVP and AVT scenarios represent the correlation coefficients estimated by average values of plots and treatments, respectively. In the lower plot, the average (rhombus), the median (vertical line), the distance between the 25th and 75th percentiles (length of the box) and the maximum and minimum values (outer spread) of the estimated correlation coefficient are presented for each scenario.

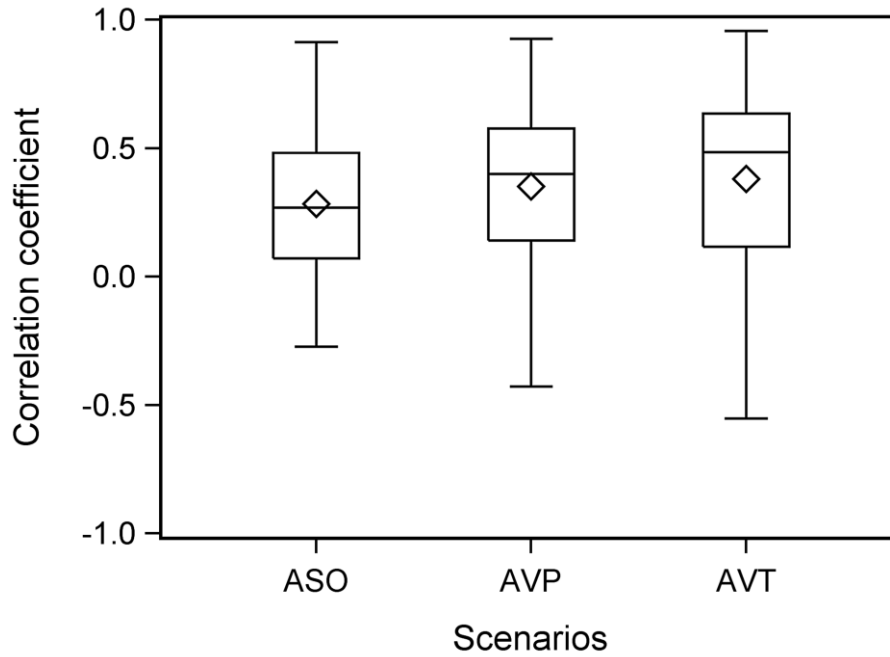


Figure 6. Descriptive analysis of correlation coefficients of 55 trait pairs estimated in 60 sample sizes by 1000 bootstrap simulations. Scenarios represent the original data coming from all sampled observations (ASO), coming from average values of each plot (AVP) and coming from average values of treatments (AVT). The rhombus within the box represents the average in the scenario. The horizontal line within the box represents the median value. The length of the box is the distance between the 25th and 75th percentiles. Outer spread represents the maximum and minimum values.

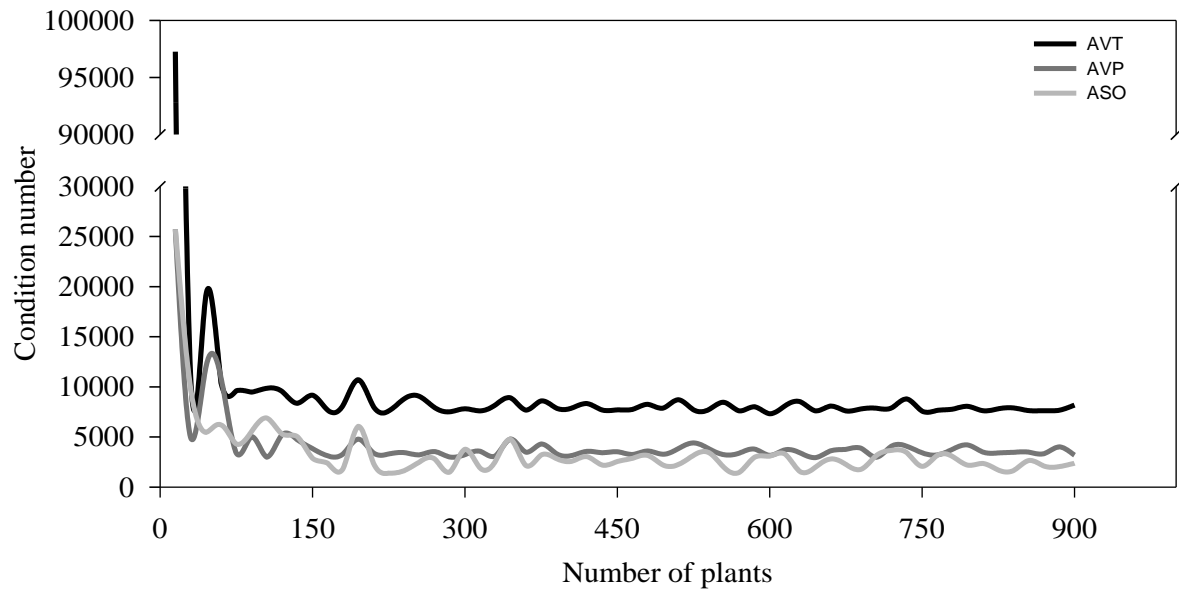


Figure 7. Condition number of correlation's matrices among explanatory traits estimated with 60 different sample sizes. For each sample size, the traits' values were estimated by average of 1000 bootstrap simulations of the original data coming from all sampled observations (ASO), coming from average values each plot (AVP) and coming from average values of treatments (AVT).

5 DISCUSSÃO GERAL

As hipóteses da presente pesquisa foram comprovadas. O método de arranjo de dados proposto, ou seja, estimar os coeficientes de correlação com todas as observações amostradas foi eficiente em reduzir o viés associado as estimativas, melhorar o condicionamento das matrizes e aumentar acurácia da análise de trilha em estudos agronômicos.

A utilização de médias nas estimativas dos coeficientes de correlação, prática comum evidenciada em diversos trabalhos, reduziu o desvio padrão de todos os caracteres analisados, superestimando os coeficientes de correlação em aproximadamente 90% das combinações. A redução na amplitude dos dados pode ser observada no Apêndice A. Este resultado refletiu diretamente no aumento da multicolinearidade da matriz de correlação das variáveis explicativas, na redução da eficiência dos métodos utilizados atualmente para ajustá-la e no aumento médio cerca de 8% no efeito residual da análise.

Os efeitos indiretos, principalmente os associados a variáveis com alto VIF, apresentaram elevadas magnitudes quando os coeficientes de trilha foram estimados sem o uso de métodos de ajuste da multicolinearidade (Apêndice B). Isto reforça ainda mais a necessidade de um diagnóstico confiável da multicolinearidade que permita ao pesquisador identificar, além de sua magnitude, quais são as variáveis associadas a este problema.

Metodologias que mascaram a real variância ou desvios de um conjunto de n -variáveis influenciarão as magnitudes de suas correlações, pois esta medida estatística é baseada na covariância e no desvio padrão dos caracteres. Em adição ao conceito estatístico metodologicamente tendencioso, a inferência da magnitude e sentido de associação entre caracteres quando a correlação é estimada com base em dados médios é equívoca, pois esta inferência é realizada para uma população com variância diferente da original (ex. de quando todas as observações são utilizadas para tal estimativa). Como grande parte dos estudos agronômicos realiza inferências populacionais baseados em amostragens (plantas), a utilização do valor médio destas plantas para estimar correlações e fazer inferência a população de interesse, é um equívoco que, sem dúvidas, não deve ser considerado.

As simulações realizadas foram eficientes em revelar o impacto dos diferentes cenários e tamanhos amostrais no comportamento estatístico do coeficiente de correlação. O aumento no número de plantas reduziu a magnitude do intervalo de confiança independentemente do cenário estudado e da magnitude da correlação avaliada, contudo o número ótimo de plantas dependeu destes fatores.

Foi observado também que o número de plantas necessário para estimativa do coeficiente de correlação no intervalo de confiança bootstrap de $95\% \leq 0,1$, aumentou no sentido de combinações com menor magnitude e que este número foi sempre maior quando todas as observações da amostra foram consideradas. Por exemplo, para uma combinação com alta magnitude de associação ($r > 0,8$), a utilização dos valores médios de parcelas e de híbridos necessitou 83 e 89% menos plantas, respectivamente, quando comparado às estimativas realizadas com todas as observações amostradas. Embora este fato leve a entender que a utilização de médias aumenta a acurácia das estimativas, a média do coeficiente de correlação para a mesma combinação foi superestimada em 11 e 15%, respectivamente. Estas correlações espúrias devem ser cautelosamente avaliadas, pois podem levar o pesquisador a uma tendenciosidade na interpretação e indicação de resultados.

A redução do número de plantas necessárias e a superestimativa do coeficiente de correlação observada quando valores oriundos de médias são utilizados está diretamente associado com a redução das variâncias individuais. Por um lado, valores médios, com menor variância e desvio padrão, necessitam menor tamanho amostral para estimativa de correlações com mesmo intervalo de confiança. Por outro, quando a variação individual é reduzida, a magnitude do coeficiente de correlação tende a aumentar devido as variáveis apresentarem covariância semelhante (dividendo da fórmula da correlação), contudo, um menor desvio padrão (divisor da fórmula da correlação), quando comparado a utilização de todas as observações amostradas.

O impacto da utilização de valores médios também foi observado no condicionamento das matrizes de variáveis explicativas. De fato, a multicolinearidade destas matrizes foi severa para os três cenários, apresentando os maiores problemas quando o tamanho amostral foi relativamente baixo ($n < 100$). A utilização dos valores médios (parcelas e tratamentos), no entanto, aumentou em 24 e 131% o valor do número de condição da matriz na média dos 60 tamanhos amostrais estudados. Este fato é preocupante, pois pesquisadores que fazem uso de médias para estimativas de correlação visando sua utilização na análise de trilha, bem como em outras que utilizam regressão múltipla, terão maiores problemas para ajustar a multicolinearidade das matrizes e, como visto, poderão ter uma redução na acurácia desta análise.

A melhor estratégia para mitigar os problemas que a multicolinearidade causa nas estimativas de coeficientes de regressão é reduzi-la, uma vez que eliminá-la completamente é praticamente impossível. Nesta pesquisa, técnicas estatísticas e biométricas foram eficazes no sentido de revelar a magnitude e a origem deste problema. Sugere-se um método simples para

reduzir a multicolinearidade e melhorar a acurácia da análise de trilha: estimar os coeficientes de correlação com todas as observações, mantendo a rastreabilidade da amostra a fim de não mascarar a real variância existente. Isto pode ser facilmente realizado sem acréscimos substanciais de tempo, mão-de-obra e recursos financeiros, pois parte-se do pressuposto que todas as plantas da amostra foram previamente analisadas.

6 CONCLUSÃO GERAL

A utilização de valores oriundos de médias reduz a variância individual de um conjunto de n -variáveis, superestima a magnitude do coeficiente de correlações entre os pares de combinação, aumenta a multicolinearidade desta matriz e reduz a acurácia das estimativas dos coeficientes de trilha.

O número de plantas necessário para estimativa de coeficientes de correlação com intervalo de confiança bootstrap de 95% é maior quando todas as observações da amostra são utilizadas e aumenta no sentido de pares de combinação com menor magnitude. No entanto, quando utilizado todas as informações amostradas, 210 plantas são suficientes para estimativa do coeficiente de correlação linear de Pearson entre caracteres de híbridos simples de milho no intervalo de confiança bootstrap de 95% $< 0,30$.

REFERÊNCIAS

ADESOJI, A. G.; ABUBAKAR, I. U.; LABE, D. A. Character association and path coefficient analysis of maize (*Zea mays* L.) grown under incorporated legumes and nitrogen. **Journal of Agronomy**, v. 14, n. 3, p. 158–163, 2015.

ALIN, A. Multicollinearity. **Wiley Interdisciplinary Reviews: Computational Statistics**, v. 2, n. 3, p. 370–374, 2010.

AUCOTT, L. S.; GARTHWAITE, P. H.; CURRALL, J. Regression methods for high dimensional multicollinear data. **Communications in Statistics–Simulation and Computation**, v. 29, n. 4, p. 1021–1037, 2000.

FARIA, L. A.; PELUZIO, J. M.; AFFÉRI, F. S.; CARVALHO, E. V. de; DOTTO, M. A.; FARIA, E. A. Análise de trilha para crescimento e rendimento de genótipos de milho sob diferentes doses nitrogenadas. **Journal of Bioenergy and Food Science**, v. 2, n. 1, p. 1–11, 2015.

FARRAR, D. E.; GLAUBER, R. R. Multicollinearity in regression analysis: the problem revisited. **The Review of Economic and Statistics**, v. 49, n. 1, p. 92–107, 1967.

GUNST, R. F.; MASON, R. L. Advantages of examining multicollinearities in regression analysis. **Biometrics**, v. 33, n. 1, p. 249–260, 1977.

HUANG, C.-C. L.; JOU, Y.-J.; CHO, H.-J. A new multicollinearity diagnostic for generalized linear models. **Journal of Applied Statistics**, v. 43, n. 11, p. 2029–2043, 2015.

JADHAV, N. H.; KASHID, D. N.; KULKARNI, S. R. Subset selection in multiple linear regression in the presence of outlier and multicollinearity. **Statistical Methodology**, v. 19, p. 44–59, 2014.

KHAMENAE, M. M.; BAHRAMINEJAD, S.; SADEGHI, F.; HONARMAND, S. J.; MANIEE, M. Path analysis and multivariate factorial analyses for determining interrelationships between grain yield and related characters in maize hybrids. **African Journal of Agricultural Research**, v. 7, n. 48, p. 6437–6446, 2012.

KIERS, H. A. L.; SMILDE, A. K. A comparison of various methods for multivariate regression with highly collinear variables. **Statistical Methods and Applications**, v. 16, n. 2, p. 193–228, 2006.

KUMAR, V.; SINGH, S. K.; BHATI, P. K.; SHARMA, A.; SHARMA, S. K.; MAHAJAN, V. Correlation, path and genetic diversity analysis in maize (*Zea mays* L.). **Environment & Ecology**, v. 33, n. 2, p. 971–975, 2015.

MANSFIELD, E. R.; HELMS, B. P. Detecting multicollinearity. **The American Statistician**, v. 36, n. 3, p. 158–160, 1982.

MA, Z.; QIN, Y.; WANG, Y.; ZHAO, X.; ZHANG, F.; TANG, J.; FU, Z. Proteomic analysis of silk viability in maize inbred lines and their corresponding hybrids. **PLoS ONE**, v. 10, n. 12, e0144050, 2015.

NARDINO, M.; SOUZA, V. Q. de; BARETTA, D.; KONFLANZ, V. A.; CARVALHO, I. R.; FOLLMANN, D. N.; CARON, B. O. Association of secondary traits with yield in maize F₁'s. **Ciência Rural**, v. 46, n. 5, p. 776–782, 2016.

NATARAJ, V.; SHAHI, J. P.; AGARWAL, V. Correlation and path analysis in certain inbred genotypes of maize (*Zea mays* L.) at Varanasi. **International Journal of Innovative Research and Development**, v. 3, n. 1, p. 14–17, 2014.

NATARAJ, V.; SHAHI, J. P.; VANDANA, D. Character association and path analyses in maize (*Zea mays* L.). **Environment and Ecology**, v. 33, n. 1, p. 78–81, 2015.

PEARSON, K. Notes on the history of correlation. **Biometrika**, v. 13, n. 1, p. 25–45, 1920.

RIGON, J. P. G.; CAPUANI, S.; BRITO NETO, J. F. de; da ROSA, G. M.; WASTOWSKI, A. D.; RIGON, C. A. G. Dissimilaridade genética e análise de trilha de cultivares de soja avaliada por meio de descritores quantitativos. **Revista Ceres**, v. 59, n. 2, p. 233–240, 2012.

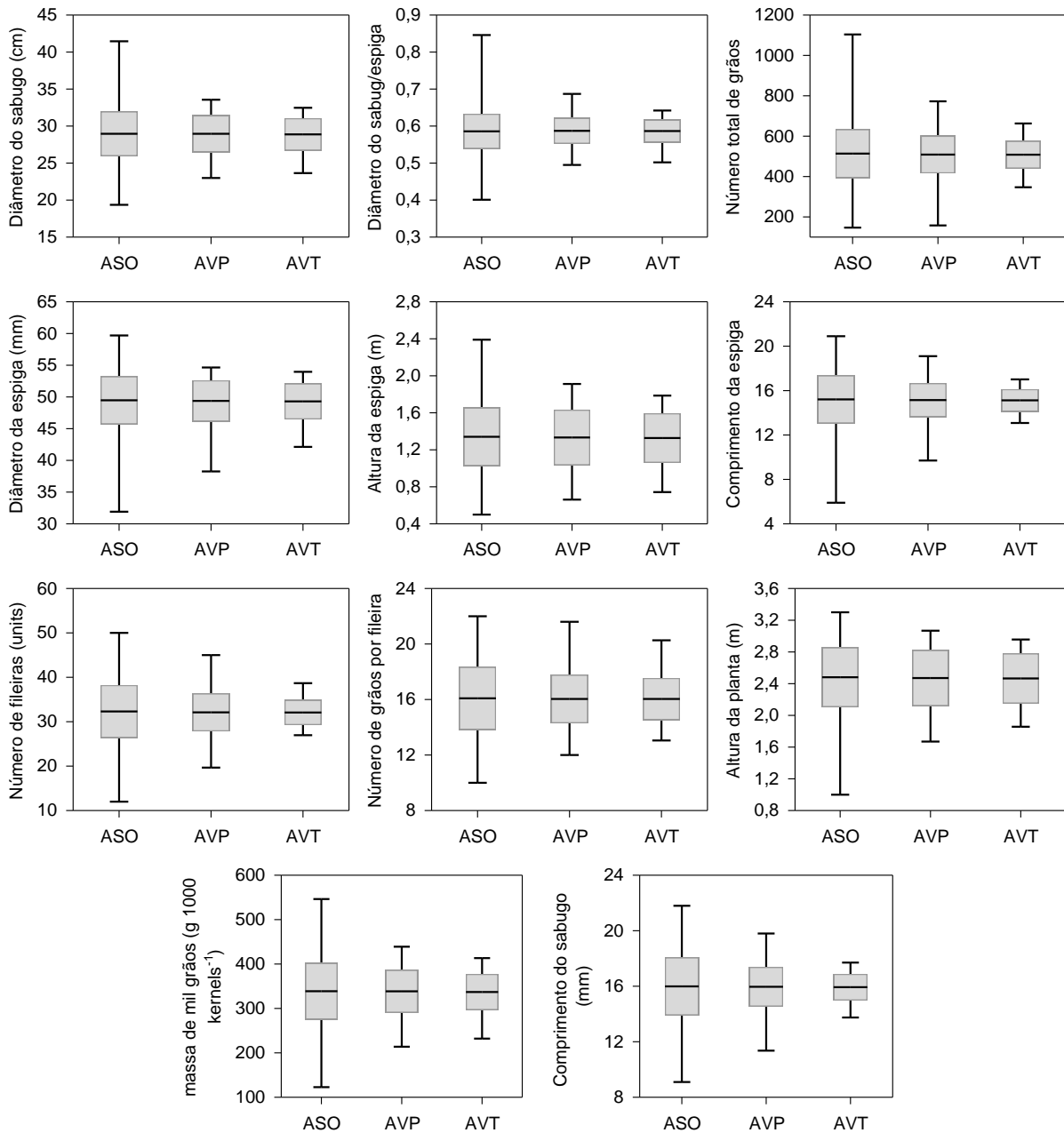
TOEBE, M.; CARGNELUTTI, A. Multicollinearity in path analysis of maize (*Zea mays* L.). **Journal of Cereal Science**, v. 57, n. 3, p. 453–462, 2013.

TORRES, F. E.; TEODORO, P. E.; RIBEIRO, L. P.; CORREA, C. C. G.; HERNANDES, F. B.; FERNANDES, R. L.; GOMES, A. C.; LOPES, K. V. Correlations and path analysis on oil content of castor genotypes. **Bioscience Journal**, v. 31, n. 5, p. 1363–1369, 2015.

WRIGHT, S. Correlation and causation. **Journal of Agricultural Research**, v. 20, n. 7, p. 557–585, 1921.

YU, H.; JIANG, S.; LAND, K. C. Multicollinearity in hierarchical linear models. **Social Science Research**, v. 53, n. 1 p. 118–136, 2015.

APÊNDICE A – ANÁLISE DESCRITIVA DAS VARIÁVEIS ANALISADAS EM CADA CENÁRIO DE ARRANJO DE DADOS.



APÊNDICE B – EFEITOS INDIRETOS ESTIMADOS EM DIFERENTES CENÁRIOS E MÉTODOS DE ANÁLISE DE TRILHA.

Material suplementar S1. Efeitos indiretos estimados tradicionalmente com todas as observações amostradas (ASO) e com os valores médios das parcelas (AVP).

VARIÁVEL†	Efeitos	Cenários	
		ASO	AVP
AP	EFEITO DIRETO SOBRE KWE	0.015	0.019
	EFEITO INDIRETO VIA AE	-0.023	-0.002
	EFEITO INDIRETO VIA CE	-0.003	-0.029
	EFEITO INDIRETO VIA DE	1.166	0.305
	EFEITO INDIRETO VIA NFG	-0.022	0.014
	EFEITO INDIRETO VIA NGF	0.010	0.043
	EFEITO INDIRETO VIA DS	-0.792	-0.227
	EFEITO INDIRETO VIA CS	-0.004	0.015
	EFEITO INDIRETO VIA NTG	0.249	0.318
	EFEITO INDIRETO VIA DS/DE	-0.296	-0.070
	EFEITO INDIRETO VIA MMG	0.215	0.332
	LINEAR	0.516	0.716
VARIÁVEL	AE		
	EFEITO DIRETO SOBRE KWE	-0.027	-0.003
	EFEITO INDIRETO VIA AP	0.013	0.018
	EFEITO INDIRETO VIA CE	-0.003	-0.028
	EFEITO INDIRETO VIA DE	1.128	0.290
	EFEITO INDIRETO VIA NFG	-0.016	0.011
	EFEITO INDIRETO VIA NGF	0.009	0.041
	EFEITO INDIRETO VIA DS	-0.951	-0.265
	EFEITO INDIRETO VIA CS	-0.003	0.014
	EFEITO INDIRETO VIA NTG	0.199	0.274
	EFEITO INDIRETO VIA DS/DE	-0.110	-0.021
	EFEITO INDIRETO VIA MMG	0.223	0.340
	LINEAR	0.461	0.671
VARIÁVEL	CE		
	EFEITO DIRETO SOBRE KWE	-0.011	-0.071
	EFEITO INDIRETO VIA AP	0.004	0.008
	EFEITO INDIRETO VIA AE	-0.007	-0.001
	EFEITO INDIRETO VIA DE	1.088	0.246
	EFEITO INDIRETO VIA NFG	-0.005	0.003
	EFEITO INDIRETO VIA NGF	0.026	0.070
	EFEITO INDIRETO VIA DS	-0.954	-0.210
	EFEITO INDIRETO VIA CS	-0.015	0.037
	EFEITO INDIRETO VIA NTG	0.441	0.354
	EFEITO INDIRETO VIA DS/DE	-0.076	-0.033
	EFEITO INDIRETO VIA MMG	0.193	0.277
	LINEAR	0.685	0.680
VARIÁVEL	DE		
	EFEITO DIRETO SOBRE KWE	2.379	0.476
	EFEITO INDIRETO VIA AP	0.007	0.012
	EFEITO INDIRETO VIA AE	-0.013	-0.002
	EFEITO INDIRETO VIA CE	-0.005	-0.036
	EFEITO INDIRETO VIA NFG	-0.045	0.022
	EFEITO INDIRETO VIA NGF	0.014	0.042
	EFEITO INDIRETO VIA DS	-1.907	-0.439
	EFEITO INDIRETO VIA CS	-0.007	0.021
	EFEITO INDIRETO VIA NTG	0.432	0.388
	EFEITO INDIRETO VIA DS/DE	-0.361	-0.032
	EFEITO INDIRETO VIA MMG	0.259	0.383
	LINEAR	0.754	0.835
VARIÁVEL	NFG		
	EFEITO DIRETO SOBRE KWE	-0.089	0.037
	EFEITO INDIRETO VIA AP	0.004	0.007
	EFEITO INDIRETO VIA AE	-0.005	-0.001
	EFEITO INDIRETO VIA CE	-0.001	-0.006
	EFEITO INDIRETO VIA DE	1.196	0.275
	EFEITO INDIRETO VIA NGF	0.004	0.017
	EFEITO INDIRETO VIA DS	-0.800	-0.180
	EFEITO INDIRETO VIA CS	-0.001	0.003

	EFEITO INDIRETO VIA NTG	0.402	0.406
	EFEITO INDIRETO VIA DS/DE	-0.320	-0.087
	EFEITO INDIRETO VIA MMG	-0.112	-0.050
VARIÁVEL	LINEAR	0.279	0.422
	NGF		
	EFEITO DIRETO SOBRE KWE	0.040	0.106
	EFEITO INDIRETO VIA AP	0.004	0.008
	EFEITO INDIRETO VIA AE	-0.006	-0.001
	EFEITO INDIRETO VIA CE	-0.007	-0.047
	EFEITO INDIRETO VIA DE	0.843	0.188
	EFEITO INDIRETO VIA NFG	-0.009	0.006
	EFEITO INDIRETO VIA DS	-0.282	-0.039
	EFEITO INDIRETO VIA CS	-0.011	0.026
	EFEITO INDIRETO VIA NTG	0.557	0.507
	EFEITO INDIRETO VIA DS/DE	-0.522	-0.146
	EFEITO INDIRETO VIA MMG	0.050	0.085
VARIÁVEL	LINEAR	0.658	0.694
	DS		
	EFEITO DIRETO SOBRE KWE	-2.903	-0.602
	EFEITO INDIRETO VIA AP	0.004	0.007
	EFEITO INDIRETO VIA AE	-0.009	-0.001
	EFEITO INDIRETO VIA CE	-0.004	-0.025
	EFEITO INDIRETO VIA DE	1.563	0.347
	EFEITO INDIRETO VIA NFG	-0.024	0.011
	EFEITO INDIRETO VIA NGF	0.004	0.007
	EFEITO INDIRETO VIA CS	-0.006	0.016
	EFEITO INDIRETO VIA NTG	0.166	0.117
	EFEITO INDIRETO VIA DS/DE	1.420	0.250
	EFEITO INDIRETO VIA MMG	0.259	0.397
VARIÁVEL	LINEAR	0.470	0.525
	CS		
	EFEITO DIRETO SOBRE KWE	-0.016	0.041
	EFEITO INDIRETO VIA AP	0.004	0.007
	EFEITO INDIRETO VIA AE	-0.005	-0.001
	EFEITO INDIRETO VIA CE	-0.010	-0.065
	EFEITO INDIRETO VIA DE	1.103	0.249
	EFEITO INDIRETO VIA NFG	-0.004	0.002
	EFEITO INDIRETO VIA NGF	0.026	0.069
	EFEITO INDIRETO VIA DS	-1.054	-0.234
	EFEITO INDIRETO VIA NTG	0.427	0.329
	EFEITO INDIRETO VIA DS/DE	0.023	-0.013
	EFEITO INDIRETO VIA MMG	0.205	0.289
VARIÁVEL	LINEAR	0.698	0.672
	NTG		
	EFEITO DIRETO SOBRE KWE	0.763	0.651
	EFEITO INDIRETO VIA AP	0.005	0.009
	EFEITO INDIRETO VIA AE	-0.007	-0.001
	EFEITO INDIRETO VIA CE	-0.006	-0.038
	EFEITO INDIRETO VIA DE	1.347	0.283
	EFEITO INDIRETO VIA NFG	-0.047	0.023
	EFEITO INDIRETO VIA NGF	0.029	0.082
	EFEITO INDIRETO VIA DS	-0.632	-0.108
	EFEITO INDIRETO VIA CS	-0.009	0.020
	EFEITO INDIRETO VIA DS/DE	-0.650	-0.170
	EFEITO INDIRETO VIA MMG	-0.057	0.007
VARIÁVEL	LINEAR	0.737	0.760
	DS/DE		
	EFEITO DIRETO SOBRE KWE	2.241	0.403
	EFEITO INDIRETO VIA AP	-0.002	-0.003
	EFEITO INDIRETO VIA AE	0.001	0.000
	EFEITO INDIRETO VIA CE	0.000	0.006
	EFEITO INDIRETO VIA DE	-0.383	-0.037
	EFEITO INDIRETO VIA NFG	0.013	-0.008
	EFEITO INDIRETO VIA NGF	-0.009	-0.038
	EFEITO INDIRETO VIA DS	-1.840	-0.374
	EFEITO INDIRETO VIA CS	0.000	-0.001
	EFEITO INDIRETO VIA NTG	-0.221	-0.275
	EFEITO INDIRETO VIA MMG	0.087	0.146
VARIÁVEL	LINEAR	-0.114	-0.183
	MMG		

EFEITO DIRETO SOBRE KWE	0.536	0.615
EFEITO INDIRETO VIA AP	0.006	0.010
EFEITO INDIRETO VIA AE	-0.011	-0.001
EFEITO INDIRETO VIA CE	-0.004	-0.032
EFEITO INDIRETO VIA DE	1.150	0.297
EFEITO INDIRETO VIA NFG	0.019	-0.003
EFEITO INDIRETO VIA NGF	0.004	0.015
EFEITO INDIRETO VIA DS	-1.400	-0.388
EFEITO INDIRETO VIA CS	-0.006	0.019
EFEITO INDIRETO VIA NTG	-0.080	0.008
EFEITO INDIRETO VIA DS/DE	0.362	0.096
LINEAR	0.575	0.635
Coefficiente de determinação	1.02	0.973
Efeito residual	0	0.161

† AP, altura de planta; AE, altura da espiga; CE, comprimento da espiga; DE, diâmetro da espiga; NFG, número de fileira de grãos; NGF, número de grãos por fileira; DS, diâmetro do sabugo; CS, comprimento do sabugo; NTG, número total de grãos por espiga; DS/DE, relação diâmetro do sabugo/diâmetro da espiga; MMG, massa de mil grãos.

Material suplementar S2. Efeitos indiretos estimados com todas as observações amostradas (ASO) e com os valores médios das parcelas (AVP) com a inclusão de k na diagonal da matriz $X'X$ das características explicativas.

VARIÁVEL†	Efeitos	Cenários	
		ASO	AVP
VARIÁVEL†	AP		
	EFEITO DIRETO SOBRE KWE	0.039	0.081
	EFEITO INDIRETO VIA AE	-0.021	0.016
	EFEITO INDIRETO VIA CE	-0.005	0.004
	EFEITO INDIRETO VIA DE	0.075	0.083
	EFEITO INDIRETO VIA NFG	0.005	0.021
	EFEITO INDIRETO VIA NGF	0.024	0.075
	EFEITO INDIRETO VIA DS	-0.036	0.017
	EFEITO INDIRETO VIA CS	0.012	0.017
	EFEITO INDIRETO VIA NTG	0.208	0.176
	EFEITO INDIRETO VIA DS/DE	-0.015	0.007
	EFEITO INDIRETO VIA MMG	0.227	0.210
	LINEAR	0.516	0.716
VARIÁVEL	AE		
	EFEITO DIRETO SOBRE KWE	-0.025	0.018
	EFEITO INDIRETO VIA AP	0.033	0.075
	EFEITO INDIRETO VIA CE	-0.004	0.004
	EFEITO INDIRETO VIA DE	0.072	0.079
	EFEITO INDIRETO VIA NFG	0.004	0.016
	EFEITO INDIRETO VIA NGF	0.022	0.073
	EFEITO INDIRETO VIA DS	-0.044	0.020
	EFEITO INDIRETO VIA CS	0.010	0.015
	EFEITO INDIRETO VIA NTG	0.166	0.151
	EFEITO INDIRETO VIA DS/DE	-0.005	0.002
	EFEITO INDIRETO VIA MMG	0.234	0.216
	LINEAR	0.461	0.671
VARIÁVEL	CE		
	EFEITO DIRETO SOBRE KWE	-0.018	0.011
	EFEITO INDIRETO VIA AP	0.011	0.034
	EFEITO INDIRETO VIA AE	-0.006	0.007
	EFEITO INDIRETO VIA DE	0.070	0.067
	EFEITO INDIRETO VIA NFG	0.001	0.005
	EFEITO INDIRETO VIA NGF	0.061	0.124
	EFEITO INDIRETO VIA DS	-0.044	0.016
	EFEITO INDIRETO VIA CS	0.044	0.042
	EFEITO INDIRETO VIA NTG	0.368	0.196
	EFEITO INDIRETO VIA DS/DE	-0.004	0.003
	EFEITO INDIRETO VIA MMG	0.203	0.176
	LINEAR	0.685	0.680
VARIÁVEL	DE		
	EFEITO DIRETO SOBRE KWE	0.152	0.129
	EFEITO INDIRETO VIA AP	0.019	0.052

	EFEITO INDIRETO VIA AE	-0.012	0.011
	EFEITO INDIRETO VIA CE	-0.008	0.006
	EFEITO INDIRETO VIA NFG	0.010	0.033
	EFEITO INDIRETO VIA NGF	0.033	0.074
	EFEITO INDIRETO VIA DS	-0.087	0.034
	EFEITO INDIRETO VIA CS	0.023	0.024
	EFEITO INDIRETO VIA NTG	0.361	0.214
	EFEITO INDIRETO VIA DS/DE	-0.018	0.003
	EFEITO INDIRETO VIA MMG	0.273	0.243
	LINEAR	0.754	0.835
VARIÁVEL	NFG		
	EFEITO DIRETO SOBRE KWE	0.020	0.058
	EFEITO INDIRETO VIA AP	0.010	0.029
	EFEITO INDIRETO VIA AE	-0.004	0.005
	EFEITO INDIRETO VIA CE	-0.001	0.001
	EFEITO INDIRETO VIA DE	0.076	0.075
	EFEITO INDIRETO VIA NGF	0.009	0.031
	EFEITO INDIRETO VIA DS	-0.037	0.014
	EFEITO INDIRETO VIA CS	0.002	0.003
	EFEITO INDIRETO VIA NTG	0.335	0.225
	EFEITO INDIRETO VIA DS/DE	-0.016	0.009
	EFEITO INDIRETO VIA MMG	-0.118	-0.032
	LINEAR	0.279	0.422
VARIÁVEL	NGF		
	EFEITO DIRETO SOBRE KWE	0.093	0.188
	EFEITO INDIRETO VIA AP	0.010	0.033
	EFEITO INDIRETO VIA AE	-0.006	0.007
	EFEITO INDIRETO VIA CE	-0.012	0.007
	EFEITO INDIRETO VIA DE	0.054	0.051
	EFEITO INDIRETO VIA NFG	0.002	0.009
	EFEITO INDIRETO VIA DS	-0.013	0.003
	EFEITO INDIRETO VIA CS	0.032	0.029
	EFEITO INDIRETO VIA NTG	0.465	0.280
	EFEITO INDIRETO VIA DS/DE	-0.026	0.014
	EFEITO INDIRETO VIA MMG	0.053	0.054
	LINEAR	0.658	0.694
VARIÁVEL	DS		
	EFEITO DIRETO SOBRE KWE	-0.133	0.046
	EFEITO INDIRETO VIA AP	0.011	0.031
	EFEITO INDIRETO VIA AE	-0.008	0.008
	EFEITO INDIRETO VIA CE	-0.006	0.004
	EFEITO INDIRETO VIA DE	0.100	0.094
	EFEITO INDIRETO VIA NFG	0.006	0.017
	EFEITO INDIRETO VIA NGF	0.009	0.012
	EFEITO INDIRETO VIA CS	0.018	0.018
	EFEITO INDIRETO VIA NTG	0.139	0.065
	EFEITO INDIRETO VIA DS/DE	0.070	-0.025
	EFEITO INDIRETO VIA MMG	0.272	0.252
	LINEAR	0.470	0.525
VARIÁVEL	CS		
	EFEITO DIRETO SOBRE KWE	0.049	0.045
	EFEITO INDIRETO VIA AP	0.010	0.030
	EFEITO INDIRETO VIA AE	-0.005	0.006
	EFEITO INDIRETO VIA CE	-0.016	0.010
	EFEITO INDIRETO VIA DE	0.071	0.067
	EFEITO INDIRETO VIA NFG	0.001	0.004
	EFEITO INDIRETO VIA NGF	0.061	0.122
	EFEITO INDIRETO VIA DS	-0.048	0.018
	EFEITO INDIRETO VIA NTG	0.357	0.182
	EFEITO INDIRETO VIA DS/DE	0.001	0.001
	EFEITO INDIRETO VIA MMG	0.216	0.183
	LINEAR	0.698	0.672
VARIÁVEL	NTG		
	EFEITO DIRETO SOBRE KWE	0.637	0.360
	EFEITO INDIRETO VIA AP	0.013	0.040
	EFEITO INDIRETO VIA AE	-0.006	0.007
	EFEITO INDIRETO VIA CE	-0.010	0.006
	EFEITO INDIRETO VIA DE	0.086	0.077
	EFEITO INDIRETO VIA NFG	0.011	0.036
	EFEITO INDIRETO VIA NGF	0.068	0.146

	EFEITO INDIRETO VIA DS	-0.029	0.008
	EFEITO INDIRETO VIA CS	0.027	0.023
	EFEITO INDIRETO VIA DS/DE	-0.032	0.017
	EFEITO INDIRETO VIA MMG	-0.059	0.005
	LINEAR	0.737	0.760
VARIÁVEL	DS/DE		
	EFEITO DIRETO SOBRE KWE	0.110	-0.040
	EFEITO INDIRETO VIA AP	-0.005	-0.014
	EFEITO INDIRETO VIA AE	0.001	-0.001
	EFEITO INDIRETO VIA CE	0.001	-0.001
	EFEITO INDIRETO VIA DE	-0.025	-0.010
	EFEITO INDIRETO VIA NFG	-0.003	-0.012
	EFEITO INDIRETO VIA NGF	-0.022	-0.068
	EFEITO INDIRETO VIA DS	-0.084	0.029
	EFEITO INDIRETO VIA CS	0.000	-0.001
	EFEITO INDIRETO VIA NTG	-0.185	-0.152
	EFEITO INDIRETO VIA MMG	0.091	0.092
	LINEAR	-0.114	-0.183
VARIÁVEL	MMG		
	EFEITO DIRETO SOBRE KWE	0.565	0.390
	EFEITO INDIRETO VIA AP	0.016	0.044
	EFEITO INDIRETO VIA AE	-0.010	0.010
	EFEITO INDIRETO VIA CE	-0.006	0.005
	EFEITO INDIRETO VIA DE	0.074	0.080
	EFEITO INDIRETO VIA NFG	-0.004	-0.005
	EFEITO INDIRETO VIA NGF	0.009	0.026
	EFEITO INDIRETO VIA DS	-0.064	0.030
	EFEITO INDIRETO VIA CS	0.019	0.021
	EFEITO INDIRETO VIA NTG	-0.067	0.004
	EFEITO INDIRETO VIA DS/DE	0.018	-0.010
	LINEAR	0.575	0.635
	<i>k</i> value	0.050	0.100
	Coefficiente de determinação	0.931	0.922
	Efeito residual	0.261	0.278

† AP, altura de planta; AE, altura da espiga; CE, comprimento da espiga; DE, diâmetro da espiga; NFG, número de fileira de grãos; NGF, número de grãos por fileira; DS, diâmetro do sabugo; CS, comprimento do sabugo; NTG, número total de grãos por espiga; DS/DE, relação diâmetro do sabugo/diâmetro da espiga; MMG, massa de mil grãos.

Material suplementar S3. Efeitos indiretos estimados com todas as observações amostradas (ASO) e com os valores médios das parcelas (AVP) excluindo as variáveis responsáveis pela multicolinearidade.

VARIÁVEL†	Efeitos	Cenários	
		ASO	AVP
	AP		
	EFEITO DIRETO SOBRE KWE	0.012	-
	EFEITO INDIRETO VIA AE	-0.033	-
	EFEITO INDIRETO VIA CE	-0.015	-
	EFEITO INDIRETO VIA DE	-0.033	-
	EFEITO INDIRETO VIA NFG	-	-
	EFEITO INDIRETO VIA NGF	0.001	-
	EFEITO INDIRETO VIA DS	-	-
	EFEITO INDIRETO VIA CS	-	-
	EFEITO INDIRETO VIA NTG	0.291	-
	EFEITO INDIRETO VIA DS/DE	-0.002	-
	EFEITO INDIRETO VIA MMG	0.294	-
	LINEAR	0.516	-
VARIÁVEL	AE		
	EFEITO DIRETO SOBRE KWE	-0.040	-
	EFEITO INDIRETO VIA AP	0.010	-
	EFEITO INDIRETO VIA CE	-0.014	-
	EFEITO INDIRETO VIA DE	-0.032	-
	EFEITO INDIRETO VIA NFG	-	-
	EFEITO INDIRETO VIA NGF	0.001	-
	EFEITO INDIRETO VIA DS	-	-
	EFEITO INDIRETO VIA CS	-	-
	EFEITO INDIRETO VIA NTG	0.232	-
	EFEITO INDIRETO VIA DS/DE	-0.001	-

	EFEITO INDIRETO VIA MMG	0.304	-
	LINEAR	0.461	-
VARIÁVEL	CE		-
	EFEITO DIRETO SOBRE KWE	-0.056	-0.073
	EFEITO INDIRETO VIA AP	0.003	-
	EFEITO INDIRETO VIA AE	-0.010	-
	EFEITO INDIRETO VIA DE	-0.031	-0.010
	EFEITO INDIRETO VIA NFG	-	-
	EFEITO INDIRETO VIA NGF	0.000	0.005
	EFEITO INDIRETO VIA DS	-	0.073
	EFEITO INDIRETO VIA CS	-	0.038
	EFEITO INDIRETO VIA NTG	0.516	0.355
	EFEITO INDIRETO VIA DS/DE	0.000	0.001
	EFEITO INDIRETO VIA MMG	0.263	0.291
	LINEAR	0.685	0.680
VARIÁVEL	DE		
	EFEITO DIRETO SOBRE KWE	-0.068	-0.019
	EFEITO INDIRETO VIA AP	0.006	-
	EFEITO INDIRETO VIA AE	-0.019	-
	EFEITO INDIRETO VIA CE	-0.026	-0.037
	EFEITO INDIRETO VIA NFG	-	0.035
	EFEITO INDIRETO VIA NGF	0.003	0.044
	EFEITO INDIRETO VIA DS	-	-
	EFEITO INDIRETO VIA CS	-	0.021
	EFEITO INDIRETO VIA NTG	0.505	0.389
	EFEITO INDIRETO VIA DS/DE	-0.002	0.001
	EFEITO INDIRETO VIA MMG	0.354	0.402
	LINEAR	0.754	0.835
VARIÁVEL	NFG		
	EFEITO DIRETO SOBRE KWE	-	0.061
	EFEITO INDIRETO VIA AP	-	-
	EFEITO INDIRETO VIA AE	-	-
	EFEITO INDIRETO VIA CE	-	-0.006
	EFEITO INDIRETO VIA DE	-	-0.011
	EFEITO INDIRETO VIA NGF	-	0.018
	EFEITO INDIRETO VIA DS	-	-
	EFEITO INDIRETO VIA CS	-	0.003
	EFEITO INDIRETO VIA NTG	-	0.408
	EFEITO INDIRETO VIA DS/DE	-	0.003
	EFEITO INDIRETO VIA MMG	-	-0.053
	LINEAR	-	0.422
VARIÁVEL	NGF		
	EFEITO DIRETO SOBRE KWE	0.005	0.111
	EFEITO INDIRETO VIA AP	0.003	-
	EFEITO INDIRETO VIA AE	-0.007	-
	EFEITO INDIRETO VIA CE	-0.003	-0.048
	EFEITO INDIRETO VIA DE	-0.034	-0.008
	EFEITO INDIRETO VIA NFG	-	0.010
	EFEITO INDIRETO VIA DS	-	-
	EFEITO INDIRETO VIA CS	-	0.026
	EFEITO INDIRETO VIA NTG	0.470	0.508
	EFEITO INDIRETO VIA DS/DE	-0.002	0.005
	EFEITO INDIRETO VIA MMG	-0.153	0.090
	LINEAR	0.279	0.694
VARIÁVEL	DS		
	EFEITO DIRETO SOBRE KWE	-	-
	EFEITO INDIRETO VIA AP	-	-
	EFEITO INDIRETO VIA AE	-	-
	EFEITO INDIRETO VIA CE	-	-
	EFEITO INDIRETO VIA DE	-	-
	EFEITO INDIRETO VIA NFG	-	-
	EFEITO INDIRETO VIA NGF	-	-
	EFEITO INDIRETO VIA CS	-	-
	EFEITO INDIRETO VIA NTG	-	-
	EFEITO INDIRETO VIA DS/DE	-	-
	EFEITO INDIRETO VIA MMG	-	-
	LINEAR	-	-
VARIÁVEL	CS		
	EFEITO DIRETO SOBRE KWE	-	0.041
	EFEITO INDIRETO VIA AP	-	-

	EFEITO INDIRETO VIA AE	-	-
	EFEITO INDIRETO VIA CE	-	-0.067
	EFEITO INDIRETO VIA DE	-	-0.010
	EFEITO INDIRETO VIA NFG	-	0.004
	EFEITO INDIRETO VIA NGF	-	0.072
	EFEITO INDIRETO VIA DS	-	-
	EFEITO INDIRETO VIA NTG	-	0.330
	EFEITO INDIRETO VIA DS/DE	-	0.000
	EFEITO INDIRETO VIA MMG	-	0.303
	LINEAR	-	0.672
VARIÁVEL	NTG		
	EFEITO DIRETO SOBRE KWE	0.892	0.653
	EFEITO INDIRETO VIA AP	0.004	-
	EFEITO INDIRETO VIA AE	-0.010	-
	EFEITO INDIRETO VIA CE	-0.032	-0.039
	EFEITO INDIRETO VIA DE	-0.038	-0.012
	EFEITO INDIRETO VIA NFG	-	0.038
	EFEITO INDIRETO VIA NGF	0.003	0.086
	EFEITO INDIRETO VIA DS	-	-
	EFEITO INDIRETO VIA CS	-	0.021
	EFEITO INDIRETO VIA DS/DE	-0.004	0.006
	EFEITO INDIRETO VIA MMG	-0.077	0.008
	LINEAR	0.737	0.760
VARIÁVEL	DS/DE		
	EFEITO DIRETO SOBRE KWE	0.014	-0.013
	EFEITO INDIRETO VIA AP	-0.002	-
	EFEITO INDIRETO VIA AE	0.002	-
	EFEITO INDIRETO VIA CE	0.002	0.006
	EFEITO INDIRETO VIA DE	0.011	0.002
	EFEITO INDIRETO VIA NFG	-	-0.013
	EFEITO INDIRETO VIA NGF	-0.001	-0.040
	EFEITO INDIRETO VIA DS	-	-
	EFEITO INDIRETO VIA CS	-	-0.001
	EFEITO INDIRETO VIA NTG	-0.259	-0.276
	EFEITO INDIRETO VIA MMG	0.118	0.153
	LINEAR	-0.114	-0.183
VARIÁVEL	MMG		
	EFEITO DIRETO SOBRE KWE	0.733	0.645
	EFEITO INDIRETO VIA AP	0.005	-
	EFEITO INDIRETO VIA AE	-0.017	-
	EFEITO INDIRETO VIA CE	-0.020	-0.033
	EFEITO INDIRETO VIA DE	-0.033	-0.012
	EFEITO INDIRETO VIA NFG	-	-0.005
	EFEITO INDIRETO VIA NGF	-0.001	0.015
	EFEITO INDIRETO VIA DS	-	-
	EFEITO INDIRETO VIA CS	-	0.019
	EFEITO INDIRETO VIA NTG	-0.094	0.008
	EFEITO INDIRETO VIA DS/DE	0.002	-0.003
	LINEAR	0.575	0.635
	Coefficiente de determinação	0.977	0.973
	Efeito residual	0.161	0.165

† AP, altura de planta; AE, altura da espiga; CE, comprimento da espiga; DE, diâmetro da espiga; NFG, número de fileira de grãos; NGF, número de grãos por fileira; DS, diâmetro do sabugo; CS, comprimento do sabugo; NTG, número total de grãos por espiga; DS/DE, relação diâmetro do sabugo/diâmetro da espiga; MMG, massa de mil grãos.