UNIVERSIDADE FEDERAL DE SANTA MARIA
CENTRO DE CIÊNCIAS RURAIS
PROGRAMA DE PÓS-GRADUAÇÃO EM AGRONOMIA

Fernando Machado Haesbaert

TESTES DE MULTICOLINEARIDADE EM VARIÁVEIS MORFOLÓGICAS E PRODUTIVAS DE TOMATEIRO

Santa Maria, RS, Brasil
2016

Fernando Machado Haesbaert

TESTES DE MULTICOLINEARIDADE EM VARIÁVEIS MORFOLÓGICAS E PRODUTIVAS DE TOMATEIRO

Tese apresentada ao Curso de Pós-Graduação em Agronomia, Área de Concentração em Produção Vegetal, da Universidade Federal de Santa Maria (UFSM, RS), como requisito parcial para obtenção do título de **Doutor em Agronomia**

Orientador: Prof. Dr. Sidinei José Lopes

Santa Maria, RS
2016

Ficha catalográfica

Fernando Machado Haesbaert

TESTES DE MULTICOLINEARIDADE EM VARIÁVEIS MORFOLÓGICAS E PRODUTIVAS DE TOMATEIRO

Tese apresentada ao Curso de Pós-Graduação em Agronomia, da Universidade Federal de Santa Maria (UFSM, RS), como requisito parcial para a obtenção do título de **Doutor em Agronomia.**

Aprovado em 15 de abril de 2016:

_____

Sidinei José Lopes, Dr. (UFSM)
(Presidente/Orientador)

_____

Alessandro Dal'Col Lúcio, Dr. (UFSM)

_____

Alberto Cargnelutti Filho, Dr. (UFSM)

_____

Lindolfo Storck, Dr. (UTFPR)

_____

Betania Brum, Dra. (UTFPR)

Santa Maria, RS
2016

Dedicatória

A minha esposa, Frankiele
Aos meus pais, Christiano e Elizabet
Aos meus irmãos, Cristian e Gabriel

AGRADECIMENTOS

À minha família, pai Christiano e mãe Elizabet, irmãos Cristian e Gabriel e minha amada esposa Frankiele, pelo apoio incondicional, pelo incentivo, amor e compreensão de vocês.
À Universidade Federal de Santa Maria e ao Programa de Pós-Graduação em Agronomia pela oportunidade de realização do curso de doutorado.
À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior(CAPES) pela concessão da bolsa de doutorado.
Aos professores do Departamento de Fitotecnia, em especial aos professores Sidinei, meu orientador, Alessandro e Alberto pelos ensinamentos e auxílio no desenvolvimento do trabalho e aos professores Lindolfo Storck e Betania Brum pela colaboração.
Agradeço aos colegas de mestrado e doutorado e bolsistas de iniciação científica pelo apoio e amizade.
Agradeço aos funcionários técnico administrativos pelo pronto atendimento das demandas da pesquisa.

Muito Obrigado!

“A alegria está na luta, na tentativa, no sofrimento envolvido e não na vitória propriamente dita”

(Mahatma Gandhi)

RESUMO

TESTES DE MULTICOLINEARIDADE EM VARIÁVEIS MORFOLÓGICAS E PRODUTIVAS DE TOMATEIRO

AUTOR: Fernando Machado Haesbaert
ORIENTADOR: Sidinei José Lopes

Este trabalho apresenta um estudo comparativo de metodologias de identificação da multicolinearidade em análises multivariadas. A multicolinearidade é ocasionada pelo intenso relacionamento linear entre variáveis em estudo e pode prejudicar a interpretação dos resultados de várias técnicas de estatística multivariada. Os objetivos deste trabalho foram comparar metodologias de identificação da multicolinearidade em diversos cenários de número de variáveis, tamanho de amostra e grau de correlação entre variáveis, bem como, identificar técnicas mais adequadas para o identificação da multicolinearidade. Foram utilizados dados de variáveis morfológicas e produtivas de um experimento com tomateiro para gerar as amostras aleatórias com distribuição normal multivariada em cenários de números de variáveis e tamanhos de amostra em três níveis de correlação entre as variáveis (baixa, média e alta). Para cada um dos cenários foram obtidas 1000 amostras multivariadas e quantificado o percentual de indicação de presença de multicolinearidade pelos critérios do determinante da matriz de correlação, número de condição e fator de inflação de variância e pelos testes de Farrar e Glauber e de Haitovsky. Os critérios e testes de avaliação da multicolinearidade apresentam resultados diferentes conforme são alterados o número de variáveis, tamanho de amostra e grau de correlação entre as variáveis. Tamanho de amostra pouco superior ao número de variáveis aumenta a ocorrência de multicolinearidade. Os critérios do número de condição e fator de inflação de variância são eficientes na identificação de multicolinearidade entre variáveis de tomateiro.

Palavras-chave: Análise Multivariada. Correlação. Pressupostos. Olericultura.

ABSTRACT

MULTICOLLINEARITY TESTS IN VARIABLE MORPHOLOGICAL AND PRODUCTION OF TOMATO

AUTHOR: FERNANDO MACHADO HAESBAERT
ADVISER: SIDINEI JOSÉ LOPES

This work presents a comparative study of multicollinearity identification methodologies in multivariate analyzes. Multicollinearity is caused by intense linear relationship between the study variables and can interfere with the interpretation of the results of various multivariate statistical techniques. The objectives of this study were to compare multicollinearity identification methodologies in different settings number of variables, sample size and degree of correlation between variables and identify the most appropriate techniques for the identification of multicollinearity. Morphological and productive variables of an experiment with tomato data were used to generate random samples with multivariate normal distribution in scenario variables numbers and sample sizes in three levels of correlation between variables (low, medium and high). For each scenario were obtained in 1000 multivariate samples and quantified the percentage of presence of multicollinearity statement by the criteria of determining the correlation matrix, condition number and factor inflation variance and the test Farrar and Glauber and Haitovsky. The criteria and the evaluation tests multicollinearity have different results are amended as the number of variable sample size and the degree of correlation between variables. Sample size slightly higher than the number of variables increases the occurrence of multicollinearity. The criteria of the condition number and variance inflation factor is effective in identifying multicollinearity among tomato variables.

Keywords: Multivariate Analysis. Correlation. Assumptions. Olericulture.

LISTA DE TABELAS

LISTA DE FIGURAS

LISTA DE ANEXO

SUMÁRIO

1 INTRODUÇÃO

O propósito da ciência moderna é testar teorias, observando e mensurando dados, com base no método científico. O método científico é norteado em quatro princípios fundamentais: i) base empírica: busca contrapor ideias a fatos observáveis; ii) amostragem: como geralmente não podemos estudar o todo, tomamos uma amostra e aceitamos que o todo se comporte como ela, desde que essa seja representativa da população; iii) controle de variáveis: quando existem variáveis que interfiram uma à outra é necessário isolar a ação de uma variável sobre a outra durante o estudo, sendo estas interações entre variáveis o objeto de estudo da estatística multivariada; e, iv) teste de hipóteses: os resultados da pesquisa são confrontados com uma hipótese empírica, podendo esta ser sustentada ou negada pelos resultados observados (VOLPATO, 2013).

No processo de aplicação do método científico, a análise estatística das informações é uma poderosa ferramenta para validar os resultados. Em pesquisas experimentais, geralmente são mensuradas diversas variáveis nas mesmas unidades experimentais. O estudo estatístico destas variáveis pode ser realizado a partir da análise de cada umas das variáveis isoladamente, utilizando para este fim as técnicas da estatística univariada. No entanto, quando um determinado fenômeno em estudo, por conta de sua complexidade depende de mais de uma variável, a análise univariada pode ser insuficiente, sendo necessário a utilização de técnicas de estatística multivariada para conhecer de forma mais ampla as relações entre as variáveis e a correta interpretação do fenômeno em estudo. Segundo Mingoti (2005), à medida que o grau de correlação entre as variáveis e o número de variáveis aumenta, a análise torna-se mais complexa e dificulta a obtenção de resultados satisfatórios com a estatística univariada.

1.1 A CULTURA DE TOMATEIRO

O tomate (*Solanum lycopersicum L.*) é uma cultura de grande importância econômica, representada pela extensa área cultivada, sendo considerada a principal cultura olerícola nacional. O cultivo do tomateiro no ano agrícola de 2014/2015 teve volume de produção no Brasil de 4,3 milhões de toneladas, em uma área de 64 mil hectares, sendo os estados maiores produtores, Goiás, Minas Gerais e São Paulo (IBGE, 2016). No Rio Grande do Sul, o cultivo do tomateiro em campo ocorre nos meses mais quentes do ano, em função das suas exigências climáticas, em que a faixa de temperatura ideal para a cultura é de 21 a 24°C, sendo a mínima de 10°C e a máxima de 38°C (FILGUEIRA, 2002). Para a realização do cultivo fora da época

ideal, se faz necessário a utilização de ambiente protegido, adequando o ambiente às condições de cultivo e possibilitando incrementos na produtividade (ANDRIOLO, 1999). A produção em ambiente protegido proporciona incremento produtivo e de qualidade dos frutos, maior precocidade de produção, melhor controle de pragas e doenças e economia da água de irrigação (CERMEÑO, 1990).

1.2 PESQUISA EXPERIMENTAL

A pesquisa na área da agronomia influenciada, no início de seu desenvolvimento, pelo aprimoramento das pesquisa e avanços na área da química que encontrou vasta aplicação na área agronômica, sendo os primeiros trabalhos realizados por Liebig (1840). No entanto, nesta época, a pesquisa na área agrícola era realizada de forma bastante empírica. O desenvolvimento das técnicas de pesquisa científica aplicadas a agricultura foram desenvolvidas somente no século XX, quando Ronald Aylmer Fisher desenvolveu os princípios da experimentação agrícola na Rothamsted Experimental Station, na Inglaterra. A partir desses estudos pioneiros e estabelecimentos das técnicas experimentais, com seus conceitos aceitos e praticados até hoje, como o princípio da repetição experimental, da casualização das repetições e estabelecimentos de delineamentos experimentais como o mais utilizado atualmente na agronomia o delineamento de blocos ao acaso. Assim, a partir de 1970 a experimentação passa a ser amplamente aceito como método adequado para a realização de pesquisas na área agrícola, tendo a parcela a unidade básica de medida dos fenômenos a serem analisados (ALMEIDA, 2004).

A pesquisa científica na área agrícola está alicerçada nos conceitos de estatística experimental, onde inúmeras metodologias estatísticas estão disponíveis para auxiliar os pesquisadores na tomada de decisões e obtenção de conclusões válidas para os estudos. A estatística experimental aplicada às pesquisas agrícolas ou experimentação agrícola, aborda desde o planejamento experimental, a condução e avaliação dos experimentos até a análise dos dados experimentais e elaboração das conclusões.

Dentre os aspectos de planejamento experimental, a amostragem tem papel fundamental da qualidade dos dados obtidos, uma vez que visa representar a população como um todo. Os métodos de análise estatística aplicada à experimentação agrícola envolve desde métodos univariados, como a análise de variância, até métodos multivariados, como regressão múltipla, análise de trilha, dentre outros.

1.3 ESTATÍSTICA MULTIVARIADA

A estatística multivariada baseia-se na observação de diversas variáveis, nas mesmas unidades experimentais ou amostrais, e tem como objetivo obter interpretações e/ou inferências com base nas respostas conjunta das diversas variáveis envolvidas no fenômeno em estudo. A necessidade de medir diversas variáveis concomitantemente é pautada pela incapacidade de variáveis isoladas abrangerem adequadamente a variabilidade dos dados obtidos de fenômenos biológico complexos.

Um grande número de técnicas de estatísticas multivariadas foi desenvolvido ao longo dos anos. No entanto, foi com a popularização dos computadores que essas técnicas puderam ser amplamente utilizadas na comunidade acadêmica, científica e também empresarial. A análise multivariada compreende várias técnicas e podem ser separadas em dois grupos (KENDALL, 1980): i) técnicas de avaliação da interdependência - estudam as relações de conjuntos de variáveis entre si, tais como as técnicas de análise de agrupamento, componentes principais, correlações canônicas e análise fatorial; ii) técnicas de avaliação da dependência - estuda a dependência de uma ou mais variáveis em relação as outras, tais como a regressão múltipla e a análise discriminante. As técnicas de estatística multivariada são potencialmente úteis em diversas áreas do conhecimento, tais como: Economia, Biologia, Administração e Agronomia.

Estudos com diversas culturas agrícolas utilizam as potencialidades das técnicas de estatística multivariada. O estudo das relações entre caracteres é fundamental em diversas áreas de interesse agrícola, principalmente no melhoramento de plantas, para a seleção indireta, quando a variável de interesse apresenta baixa herdabilidade ou quando existe dificuldade de mensuração da variável de interesse para a seleção precoce de plantas e seleção simultânea para mais de um caractere (CRUZ; REGAZZI, 1997), no estudo das relações entre características das plantas.

Embora a complexidade das técnicas multivariadas seja maior, quando comparada as técnicas univariadas, o rápido desenvolvimento de softwares proporcionou o suporte tecnológico necessário para a difusão e aplicação em larga escala, por diversos setores. No entanto, faz-se necessário a capacitação dos profissionais envolvidos a fim de proporcionar o conhecimento estatístico necessário para a correta escolha e utilização da técnica, bem como, observar as premissas dos modelos utilizados.

1.4 PRESSUPOSIÇÕES NAS ANÁLISES MULTIVARIADAS

As pressuposições das técnicas multivariadas podem ser entendidas como sendo todos os cuidados referentes a obtenção e a avaliação dos dados para atender as exigências estatísticas necessárias para cada técnica. Assim, um pressuposto inicial refere-se a obtenção de amostras representativas de uma população, quando não é possível avaliar o todo. Os cuidados referentes à obtenção da amostra vão desde o planejamento da forma de coleta, do número de elementos amostrados e da coleta em si, cuidados esses referentes a manipulação dos dados, para evitar erros grosseiros, tais como, erros de medição, anotação ou de calibração de equipamentos. A amostragem necessita de especial atenção quanto ao planejamento e execução, de modo que os procedimentos de amostragem sejam conduzidos cuidadosamente, pois será a partir dos dados da amostra que serão realizadas as análises e obtidos os resultados, que serão generalizados para toda a população. Segundo Hair et al. (2009), tamanhos de amostra pequenos podem resultar em bom ajuste do modelo, mas com baixíssimo poder estatístico do teste, ou seja, quando amostras pequenas são adotadas para ajustes de modelos multivariados pode ocorrer bom ajuste da amostra ao modelo, mas sem poder de generalização dos resultados. Por outro lado, os mesmos autores acrescentam que tamanhos de amostra muito grandes ($n > 400$) tornam os testes muito sensíveis e praticamente todo efeito é significante.

Outro importante procedimento de avaliação inicial dos dados é a identificação de dados perdidos ou de dados atípicos (outliers). A identificação de dados perdidos tem grande importância na estatística multivariada. Mingoti (2005) recomenda que somente dados completos devem ser utilizados em análises multivariadas, ou seja, se para um dado elemento amostral não foi possível mensurar ou foi perdido o valor referente a alguma variável avaliada, este elemento amostral deve ser eliminado do conjunto de dados. Da mesma forma, os valores atípicos devem ser cuidadosamente identificados, pois podem causar prejuízos na interpretação dos resultados. Valores atípicos são observações que apresentam um grande afastamento das demais observações sendo inconsistente com o conjunto de dados por possuírem valores extremos. Muitas vezes esses dados podem ser oriundos de erros de medição, anotação ou digitação e sempre que possível, devem ser verificados quanto ao seu verdadeiro valor e corrigido. Quando não é possível de ser verificado, o procedimento mais prudente a ser utilizado deve ser a exclusão do elemento amostral (MINGOTI, 2005).

Outro passo da avaliação dos dados envolve a verificação do atendimento das pressuposições estatísticas do modelo multivariado a ser utilizado. Segundo Hair et al. (2009), a necessidade da avaliação dos pressupostos aumenta em análises multivariadas, quando comparada com técnicas univariadas, devido à complexidade das relações entre as variáveis, o

que torna as distorções e o viés mais significativo quando alguma pressuposição é violada. A avaliação das pressuposições dos modelos multivariados deve ser realizada tanto para as variáveis isoladamente, quanto coletivamente.

Para Hair et al. (2009), quatro são as pressuposições estatísticas necessárias para aplicação das técnicas multivariadas, sendo a primeira delas a normalidade dos dados. Para realizar a maioria das técnicas multivariadas, é necessário que as variáveis envolvidas tenham distribuição normal univariada e também multivariada (JOHNSON; WICHERN, 2007; HAIR et al., 2009). Esta distribuição é uma generalização da normal univariada, para conjuntos com duas ou mais variáveis aleatórias (MINGOTI, 2005; FERREIRA, 2008). É importante que a distribuição das variáveis aleatórias sigam distribuição normal multivariada pois, na estatística multivariada, a maioria das técnicas tem como pressuposto que os dados sigam normalidade multivariada para sua efetivação e obtenção de qualidade nas inferências. Um estudo aprofundado das propriedades da distribuição normal multivariada pode ser obtido em Ferreira (2008).

Segundo Hair et al. (2009), se um conjunto de variáveis tem distribuição normal multivariada, as variáveis isoladamente também serão normalmente distribuídas, no entanto, a recíproca não necessariamente é verdadeira, pois mesmo se as variáveis isoladamente são normais univariadas, conjuntamente as variáveis podem não seguir a distribuição normal multivariada. Com isso, torna-se necessário a avaliação da normalidade multivariada por testes específicos, tais como o teste de Shapiro-Wilks multivariado (ROYSTON, 1983, 1993) e uma extensão do teste univariado Shapiro-Francia (SHAPIRO; FRANCIA, 1972) para o caso multivariado proposto por Silva (2009).

Outro pressuposto é homocedasticidade, o qual se refere à suposição de que as variáveis apresentam níveis iguais de variâncias ao longo do domínio, ou seja, que as variâncias sejam homogêneas no conjunto das variáveis. Um teste que avalia a dependência das variáveis conjuntamente é o teste de M de Box multivariado (HAIR et al., 2009). A linearidade das relações também é necessária de ser avaliada quando técnicas multivariadas baseadas em medidas correlacionais entre as variáveis são utilizadas, tais como regressão múltipla, regressão logística, análise fatorial e modelagem de equações estruturais. A necessidade desta avaliação provém do fato de a correlação quantificar somente a relação linear entre as variáveis, não identificando os efeitos não lineares. A avaliação visual das relações entre as variáveis, através de diagramas de dispersão, é a forma mais prática de identificação do padrão de relacionamento entre as variáveis (HAIR et al., 2009).

A ausência de erros correlacionados também é uma exigência das técnicas multivariadas. Entende-se por erros correlacionados, segundo Kendall e Buckland (1971) a condição de correlação entre amostras observadas no tempo ou no espaço. A ocorrência de erros correlacionados é comum quando fatores não incluídos no modelo interferem no resultado, ou mesmo devido aos procedimentos de coleta dos dados que podem causar interferências em algum grupo de variáveis e não em outros. A forma mais comum de redução do correlacionamento entre os erros é a inclusão de variáveis que represente o fator omitido.

Outra característica necessária aos dados refere-se à ausência ou baixo grau de multicolinearidade entre as variáveis. Multicolinearidade é o grau em que uma variável pode ser explicada pelas outras variáveis do modelo (HAIR et al., 2009), e essa relação de interdependência é comum de ser encontrada entre variáveis independentes. A necessidade de ausência de multicolinearidade em graus elevados se dá pelo fato de que, quando presente, prejudica a interpretação dos resultados das técnicas multivariadas.

Efeitos como estimação de parâmetros, em modelos de regressão múltipla, sem sentido prático e com alto erro padrão dos coeficientes (HAIR et al., 2009), aumento nos erros de estimação dos efeitos diretos da análise de trilha, com resultados sem sentido biológico e sem interpretação prática (TOEBE; CARGNELUTTI FILHO, 2013), são comuns em análises multivariadas com presença de multicolinearidade. Em análise conjunta, a presença de multicolinearidade em grau elevado pode causar incapacidade de obter estimativas confiáveis (HAIR et al., 2009). Na análise de correlação canônica, a presença de multicolinearidade em um dos grupos, pode afetar os coeficientes canônicos, não representando a verdadeira relação entre os grupos.

1.5 MULTICOLINEARIDADE

O termo multicolinearidade foi utilizado pela primeira vez por FRISCH (1934) e originalmente significava uma relação linear perfeita entre duas ou mais variáveis. Este conceito clássico não inclui outras relações não lineares entre variáveis o que, segundo Gujarati e Porter (2011), também pode prejudicar a aplicação de técnicas multivariadas, pois variáveis com relação não linear, em geral, apresentam alta correlação.

Em análise de regressão múltipla, a inclusão de variáveis multicolineares, geralmente, melhora os modelos de regressão, mas a inclusão de variáveis redundantes pode ter efeitos prejudiciais nas estimativas dos parâmetros do modelo. Além dos efeitos na explicação, a multicolinearidade pode ter sérios efeitos nas estimativas dos coeficientes de regressão e na

aplicabilidade geral do modelo estimado (HAIR et al., 2009). Nesse sentido, a existência de multicolinearidade em grau elevado entre as variáveis independentes, pode causar erros na análise e interpretação dos dados. Portanto, ao escolher as variáveis independentes que farão parte do modelo de regressão é necessário que antes seja realizado o diagnóstico de multicolinearidade.

A multicolinearidade é devida a vários fatores, como problemas na amostragem, restrições do modelo ou da amostra, especificação do modelo que pode ocorrer quando se adiciona termos polinomiais no modelo de regressão, ou quando ocorre sobredeterminação do modelo, que acontece quando se tem mais variáveis explanatórias do que número de amostras ou observações (MONTGOMERY et al., 2012).

Para Achen (1982), a multicolinearidade tem como efeito, ou problema, causar a obtenção de erros padrão muito altos nas estimativas dos coeficientes, mas que este problema também pode ser causado por micronumerosidade (tamanho de amostra muito pequeno), como também, a falta ou pouca variância das variáveis independentes do modelo. Para Kmenta (1986) é natural ocorrer a multicolinearidade entre as variáveis independentes. Porém deve-se detectar o grau ou intensidade da multicolinearidade, uma vez que é uma característica das variáveis explanatórias amostradas.

Kennedy (2009) argumenta que o estimador da regressão linear pelo método dos mínimos quadrados ordinários, na presença de multicolinearidade, permanece não viesado e, de fato, ainda é o melhor estimador linear não viesado. Assim, se os pressupostos da regressão linear continuam a ser atendidos, o estimador pelo método dos mínimos quadrados mantém todas as suas propriedades desejáveis. Figueiredo Filho et al. (2011) citam que a maior dificuldade de modelos com problemas de multicolinearidade é o aumento da magnitude da variância dos parâmetros estimados. Isso porque a presença de altos níveis de correlação entre as variáveis independentes impossibilita estimar, com precisão, o efeito de cada variável sobre a variável dependente. Para Kennedy (2009), além de criar altas variações nas estimativas dos coeficientes, a multicolinearidade está associada a problemas indesejáveis nos cálculos com base na matriz de dados que sejam instáveis, ou seja, nos quais pequenas variações na matriz de dados, tais como a adição ou supressão de uma única observação, pode levar a grandes mudanças nas estimativas dos parâmetros.

Garson (2011) conceitua a multicolinearidade como uma correlação excessiva entre as variáveis preditoras. Quando a correlação é excessiva ($r \geq 0{,}90$), o erro padrão dos coeficientes se torna grande, tornando difícil ou impossível avaliar a importância relativa das variáveis preditoras. A multicolinearidade é menos importante quando a finalidade da pesquisa é a

predição, já que os valores preditos da variável dependente permanecem estáveis, mas a multicolinearidade é um problema grave quando a finalidade da pesquisa inclui a modelagem causal.

A avaliação da multicolinearidade é fundamental no melhoramento de plantas, principalmente para a seleção indireta, quando a variável de interesse apresenta baixa herdabilidade ou quando existe dificuldade de mensuração de determinada variável, para a seleção precoce de plantas e seleção simultânea para mais de um caractere (CRUZ; REGAZZI, 1997). Também foi constatado na cultura do milho, que o elevado grau de multicolinearidade e em menor escala, a não-normalidade multivariada, aumentam os erros de estimação dos efeitos diretos da análise de trilha, com resultados sem sentido biológico e sem interpretação prática (TOEBE; CARGNELUTTI FILHO, 2013). Nesse estudo, ficou evidenciado que a transformação de dados e a eliminação de variáveis altamente correlacionadas, permitem a estimação de efeitos diretos estáveis e confiáveis.

Por outro lado, Goldberger (1991) destaca que erros cometidos na estimação de coeficientes de correlação e nas análises complementares (como por exemplo, análise de trilha) em casos de micronumerosidade (utilização de amostras pequenas) podem ser maiores em relação aos erros cometidos nessas estimativas, em caso de alto grau de multicolinearidade. O termo micronumerosidade refere-se a tamanhos de amostras muito pequenos, ou quando o tamanho de amostra é pouco superior ao número de variáveis.

A multicolinearidade também pode ser entendida como a extensão em que uma variável pode ser explicada por outras variáveis explicativas (HAIR et al., 2009). Já o termo colinearidade significa a relação linear entre duas variáveis independentes. Já a multicolinearidade se refere à relação linear entre mais de duas variáveis explicativas, ou seja, a interdependência ocorre entre três ou mais variáveis independentes, embora, ambos os termos são usados como sinônimos (GUJARATI; PORTER, 2011; HAIR et al., 2009).

Outros autores (NETER; WASSERMAN, 1974) afirmam que só deve-se utilizar o termo multicolinearidade nos casos em que a correlação entre as variáveis é muito elevada ou perfeita (1 ou -1). Por outro lado, Schneider et al. (2009) salientam que o problema da multicolinearidade se manifesta quando a correlação entre as variáveis independentes é significativa e maior do que a correlação destas com a variável dependente. Os mesmos autores enfatizam que o termo multicolinearidade é empregado erroneamente como sinônimo de correlação alta ou perfeita entre as variáveis independentes.

Em análise de regressão, a avaliação de multicolinearidade requer a existência de uma medida para expressar o grau em que cada variável independente é explicada pelo conjunto das

demais variáveis independentes (HAIR et al., 2009). Portanto, a identificação da presença de multicolinearidade não é suficiente para saber se isso afetará ou não o modelo de regressão. É necessário conhecer a magnitude da multicolinearidade, que quando forte ou severa, exige um tratamento adequado para reduzi-la ou eliminá-la (GUJARATI; PORTER, 2011). Deste modo, a escolha da técnica a ser utilizada na identificação e quantificação da magnitude da multicolinearidade é necessária para verificar a necessidade de correção da multicolinearidade de variáveis (GUJARATI; PORTER, 2011).

A medida que a multicolinearidade é identificada no conjunto das variáveis, algumas soluções são propostas conforme a técnica utilizada. Em análise de regressão múltipla por exemplo, Hair et al. (2009) propõem algumas medidas, tais como: excluir variáveis independentes altamente correlacionadas; substituir variáveis altamente correlacionadas por outras variáveis potencialmente úteis como relações entre variáveis; utilização de modelos de regressão múltipla apenas para a finalidade de previsão, pois mesmo com presença de variáveis independentes altamente correlacionadas, o modelo de regressão é o melhor estimador não viesado; e adotar métodos de análise mais elaborados como a regressão Bayesiana. Pimentel et al. (2006) propõe o uso de técnicas viesadas de estimação dos coeficientes de regressão, como a regressão de cumeeira.

Em análise de trilha, ao se detectar multicolinearidade, a análise em crista permite a manutenção de todas as variáveis no modelo (CRUZ; CARNEIRO, 2003). Na análise de agrupamento, em presença de multicolinearidade, o uso de todos os caracteres não é um procedimento adequado, pois os caracteres multicolineares serão ponderados com maior peso (BARROSO; ARTES, 2003; CRUZ; CARNEIRO, 2003; HAIR et al., 2005; CORRAR et al., 2007). A utilização de componentes principais pode ser utilizado para resolver a multicolinearidade (KUTNER et al., 2003), pois os componentes principais são combinações lineares independentes.

Para contornar o problema da micronumerosidade relatado por Goldberger (1991), que provoca aumento dos erros padrão e aumento da incerteza nos resultados, Judge et al. (1988) recomendam tamanhos de amostra maiores para utilizar técnicas multivariadas em banco de dados com variáveis com problema de multicolinearidade. No entanto, não existe consenso na bibliografia quanto ao tamanho de amostra mínimo. Para Gujarati e Porter (2011) o número de observações amostradas deve ser pelo menos maior que o número de variáveis utilizadas. Para Hair et al. (2009) é necessário pelo menos cinco vezes mais observações que o número de variáveis envolvidas na análise.

Outra técnica utilizada frequentemente, para contornar a presença inconveniente da multicolinearidade, é a transformação de dados. Segundo Hair et al. (2009), a transformação de dados pode contribuir, tanto no atendimento da normalidade, quanto na melhoria das relações entre variáveis, favorecendo a redução do grau de multicolinearidade. Toebe e Cargnelutti Filho (2013) concluíram que a transformação de dados reduz o grau de multicolinearidade e a variabilidade das estimativas dos efeitos diretos, na análise de trilha tradicional com alto grau de multicolinearidade.

1.6 TESTES PARA IDENTIFICAÇÃO DE MULTICOLINEARIDADE

A forma mais empírica de identificação da multicolinearidade é através do coeficiente de correlação linear de Pearson entre variáveis explicativas. A análise do coeficiente de correlação linear de Pearson de cada par de variáveis pode indicar alta colinearidade quando a correlação entre duas variáveis explicativas é alta (VASCONCELLOS; ALVES, 2000; CRUZ; CARNEIRO, 2006).

A presença de correlação elevada entre as variáveis pode proporcionar elevado grau de multicolinearidade, o que resulta na obtenção de estimativas incoerentes. Em análise de regressão múltipla, por exemplo, pode causar estimativas de parâmetros com sinal inverso e sem sentido prático. A análise da multicolinearidade busca verificar se existe dependência entre as variáveis pois, caso exista, essa dependência provoca degenerações no modelo, limitando a utilização.

Por vezes a multicolinearidade é relacionada apenas à correlação linear entre as variáveis. Dormann et al. (2003) relatam que o coeficiente de correlação linear de Pearson superior a 0,7 é um indicador para multicolinearidade, pois distorce as estimativas em modelos de previsão. De acordo como Mason e Perreault (1991) e Tabachnick e Fidell (2013), valores de correlação lineares maiores que 0,80 são críticos. Já para Pasquali (2004), somente correlações superiores a 0,9 são indicativos de multicolinearidade. No entanto, é sabido que multicolinearidade e correlação não são sinônimos (BELSLEY et al., 1980), pois baixos valores de correlação podem mascarar altos níveis de multicolinearidade. Segundo Moore et al. (1984), a matriz de correlação linear entre as variáveis pode não ser eficiente no diagnóstico de multicolinearidade quando as dependências lineares não são em função das correlações simples e sim das inter-relações entre grupos de variáveis.

O determinante da matriz de correlação entre as variáveis é um indicador de multicolinearidade e pode ser obtido pelo produto dos valores próprios ou autovalores da

matriz. O determinante da matriz de correlação varia entre 0 e 1, sendo que 0 indica que a matriz é singular (correlações perfeitas) e 1 indica que não há qualquer tipo de correlação entre as variáveis, ou seja, as variáveis são ortogonais. Deste modo, o determinante da matriz de correlação pode ser utilizado como indicador de multicolinearidade.

Outra medida muito utilizada para identificar a presença de multicolinearidade é o número de condição (MONTGOMERY et al., 2012). O número de condição de uma matriz de dados mede a sensibilidade das estimativas dos parâmetros a pequenas mudanças na matriz de dados (BELSLEY et al., 1980; BELSLEY, 1982). O teste do número de condição possibilita classificar a multicolinearidade em três níveis: fraca, moderada e severa.

A multicolinearidade pode ser avaliada, segundo Fávero et al. (2009), pelo fator de inflação de variância (FIV), que indica a quantidade de aumento da variância do coeficiente de regressão estimado, em função da presença da multicolinearidade. Segundo Hair et al. (2009), a raiz quadrada do fator de inflação de variância indica o grau de aumento no erro padrão devido à existência de multicolinearidade, ou seja, a raiz quadrada do FIV de uma determinada variável indica o aumento esperado no erro padrão do coeficiente da variável em comparação ao coeficiente esperado na ausência de multicolinearidade. Esse aumento no erro padrão não é desejado, pois quanto maior for o erro padrão, o intervalo de confiança também será maior e assim, maior será a dificuldade em detectar a significância estatística dos parâmetros do modelo de regressão estimados. Relacionada com o fator de inflação de variância a tolerância também é amplamente utilizada para avaliação da multicolinearidade. A tolerância representa quanto de variabilidade de cada variável independente não é explicada pelas outras variáveis independentes (HAIR et al., 2009). A Tolerância = $1 - R_x^2$, em que, R_x^2 é o coeficiente de determinação do modelo da variável explicativa X, em função das demais variáveis explicativas presentes (FÁVERO et al., 2009).

A multicolinearidade também pode ser avaliada através do índice FG, proposto por Farrar e Glauber (1967). O índice FG é uma adaptação do teste de esfericidade de Bartlett (BARTLETT, 1937) para avaliar a multicolinearidade. Esse índice é uma estatística que considera em sua equação o número de variáveis envolvidas, o tamanho de amostra e o determinante da matriz de correlação, além de ser uma estatística que tem distribuição aproximadamente qui-quadrado e assim, segundo os autores, essa estatística é útil para identificar a multicolinearidade pois oferece uma medida generalizada, por meio da padronização do tamanho de amostra (n) e do número de variáveis (p). Porém, existem poucas informações sobre a utilização desse índice com a finalidade de identificação da presença de multicolinearidade.

O teste de Farrar e Glauber (1967) foi criticado por Haitovsky (1969) que propôs outro teste, o teste de significância de Haitovsky. Esse teste verifica se a matriz de correlação é singular devido a muitas correlações altas entre as variáveis. Uma matriz não singular é aquela que mostra valores relativamente baixos de correlação entre as variáveis. A matriz não singular tem um determinante que está próximo de 1,0. O teste de significância de Haitovisky é um teste para multicolinearidade que examina a hipótese nula de que a matriz de correlações entre variáveis é singular, com determinante igual a zero.

Segundo Cortina (1993), o procedimento de simulação Monte Carlo também poderia ser usado para determinar se existe ou não um nível de multicolinearidade. Pereira et al. (2014) realizaram estudos que utilizem esta técnica para avaliar estimadores de regressão ridge em experimentos na área de entomologia com diferentes graus de multicolinearidade entre as variáveis. Pereira (2014) também utilizou a técnica de simulação Monte Carlo para estudos de modelagem de equações diferenciais e regressão ridge sob diferentes níveis de multicolinearidade.

Para a utilização de técnicas de análises multivariada é recomendado a avaliação rigorosa dos dados, evitando a violação dos pressupostos dos modelos utilizados (HAIR et al., 2009). Como relatado acima, existem diversos testes possíveis de utilização na identificação da multicolinearidade, no entanto, não existe um único teste recomendado, ou indicação de qual o mais eficiente na identificação e quantificação da multicolinearidade. Segundo Gunasekara et al. (2008) não existe, de forma bem definida, um teste de magnitude da multicolinearidade.

A hipótese do presente estudo é que os resultados das diferentes técnicas de avaliação da multicolinearidade apresentam respostas diferentes quanto à presença ou ausência de multicolinearidade, e que são influenciadas pelo número de variáveis, o tamanho de amostra e o grau de correlação entre as variáveis morfológicas e produtivas na cultura do tomateiro.

Os objetivos deste trabalho foram comparar metodologias de identificação da multicolinearidade em diversos cenários de número de variáveis, tamanho de amostra e grau de correlação entre as variáveis morfológicas e produtivas de tomateiro, bem como, identificar qual(is) a(s) técnica(s) que pode(m) ser mais adequada(s) para identificação da multicolinearidade em variáveis independentes de experimentos com a cultura do tomateiro.

2 MATERIAL E MÉTODOS

2.1 BANCO DE DADOS

Foram utilizados os dados de um experimento com tomateiro conduzido na primavera-verão do ano de 2010 sob ambiente protegido (túnel alto) no Departamento de Fitotecnia da Universidade Federal de Santa Maria (UFSM) em Santa Maria, RS. Foi utilizado o híbrido Grandeur, do tipo salada e de crescimento indeterminado. O ambiente de cultivo em túnel alto, utilizado no experimento, possui cobertura plástica de polietileno transparente de 150 micras com aditivo anti UV. As plantas foram dispostas em três fileiras espaçadas de um metro e 0,75 m entre plantas, totalizando 22 plantas por fileira. Os tratos culturais adotados para condução do cultivo foram aplicados de acordo com a recomendação da cultura (FILGUEIRA, 2002).

As necessidades hídricas da cultura foram atendidas utilizando o sistema de irrigação por gotejamento e as adubações foram realizadas buscando produção de 75 t ha^{-1}. A adubação de base foi realizada conforme análise de solo, com 65 kg ha^{-1} de N, 230 kg ha^{-1} de P_2O_5 e 65 kg ha^{-1}de K_2O. As adubações de cobertura, com mais 30 kg ha^{-1} de N e K_2O, foram realizadas em intervalos de 15 dias, iniciando aos 20 dias após transplante e em seguida a cada adubação de cobertura realizou-se a amontoa para incorporar a adubação e eliminar plantas daninhas (SOCIEDADE BRASILEIRA DE CIÊNCIA DO SOLO, 2004; FILGUEIRA, 2002).

A cultura foi conduzida em haste única, realizando o desbaste das brotações não desejadas e o tutoramento feito por fios de ráfia. O manejo do ambiente protegido consistiu da abertura diária entre sete e oito horas e fechamento entre 16 e 18 horas, observando as condições climáticas de vento e chuva. O manejo fitossanitário foi realizado preventivamente com fungicidas e inseticidas recomendados para a cultura.

As variáveis mensuradas nas 66 plantas deste experimento foram: variáveis morfológicas - altura de planta (AP) em centímetros, número de folhas (NF), diâmetro do caule no colo da planta (DC) em milímetros, e variáveis produtivas - número de inflorescências/infrutescências (NI) e número de frutos por planta (NFR) em diferentes dias após o transplante (47, 61, 75 e 88 DAT), totalizando vinte variáveis.

2.2 SIMULAÇÃO DE DADOS E CENÁRIOS

As variáveis apresentadas na tabela 1 foram utilizadas como referência para gerar as amostras aleatórias multivariadas. As variáveis morfológicas: altura de planta, número de folhas

e diâmetro do caule apresentaram a menor variabilidade e as variáveis produtivas, número de inflorescências e número de frutos as maiores variabilidades. Deste modo, as simulações foram realizadas para variáveis morfológicas e produtivas separadamente, assim, em cada simulação, p variáveis foram selecionadas aleatoriamente dentre as 12 variáveis morfológicas e posteriormente para as oito variáveis produtivas.

Tendo como base as 12 variáveis morfológicas e as oito variáveis produtivas, foram elaboradas rotinas de programação no software R (R CORE TEAM, 2015) que geram amostras aleatórias multivariadas com distribuição normal multivariada para diferentes números de variáveis (p) e tamanhos de amostra (n) em três cenários de correlação (baixa, média e alta) entre as variáveis. Para as variáveis morfológicas os valores de número de variáveis (p) foram: 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 e 12 e para as variáveis produtivas os valores de p são: 2, 3, 4, 5, 6, 7 e 8.

Os diferentes tamanhos de amostra (n) simulados foram: 25, 30, 40, 50, 100, 150, 200, 500 e 1000, utilizando três níveis de correlação entre as variáveis, sendo o nível baixo para valores de correlação de todas as variáveis entre 0 e 0,3, para a correlação intermediária entre as variáveis a correlação foi estabelecida entre 0,4 e 0,7 e para alta correlação entre as variáveis, valores maiores do que 0,8. Os cenários de diferentes níveis de correlação foram utilizados com o objetivo de obter diferentes níveis de multicolinearidade. Tal estratégia também foi utilizada por Mason e Perreault (1991), Grewal et al. (2004) em estudos de marketing e também por Adenomon e Oyejola (2014) em estudos de modelos de previsão de modelos de séries temporais em dados econométricos e Pereira et al. (2014) em estudo entomológicos.

O primeiro passo da programação foi selecionar p variáveis de forma aleatória e obter as estatísticas descritivas de média e variância destas variáveis, selecionadas no banco de dados de plantas de tomateiro. A seguir, gerar a matriz sigma, contendo na diagonal principal a média das variâncias das variáveis selecionadas e no restante da matriz a multiplicação da variância média das variáveis selecionadas pelo valor da correlação desejada. Conforme apresentado na matriz abaixo:

$$\begin{bmatrix} \overline{S}^2 & \cdots & \rho.\overline{S}^2 \\ \vdots & \ddots & \vdots \\ \rho.\overline{S}^2 & \cdots & \overline{S}^2 \end{bmatrix}$$, onde $\overline{S}^2$ é a variância média das variáveis selecionadas para compor a simulação e ρ o coeficiente de correlação desejado para a matriz conforme o cenário a ser simulado.

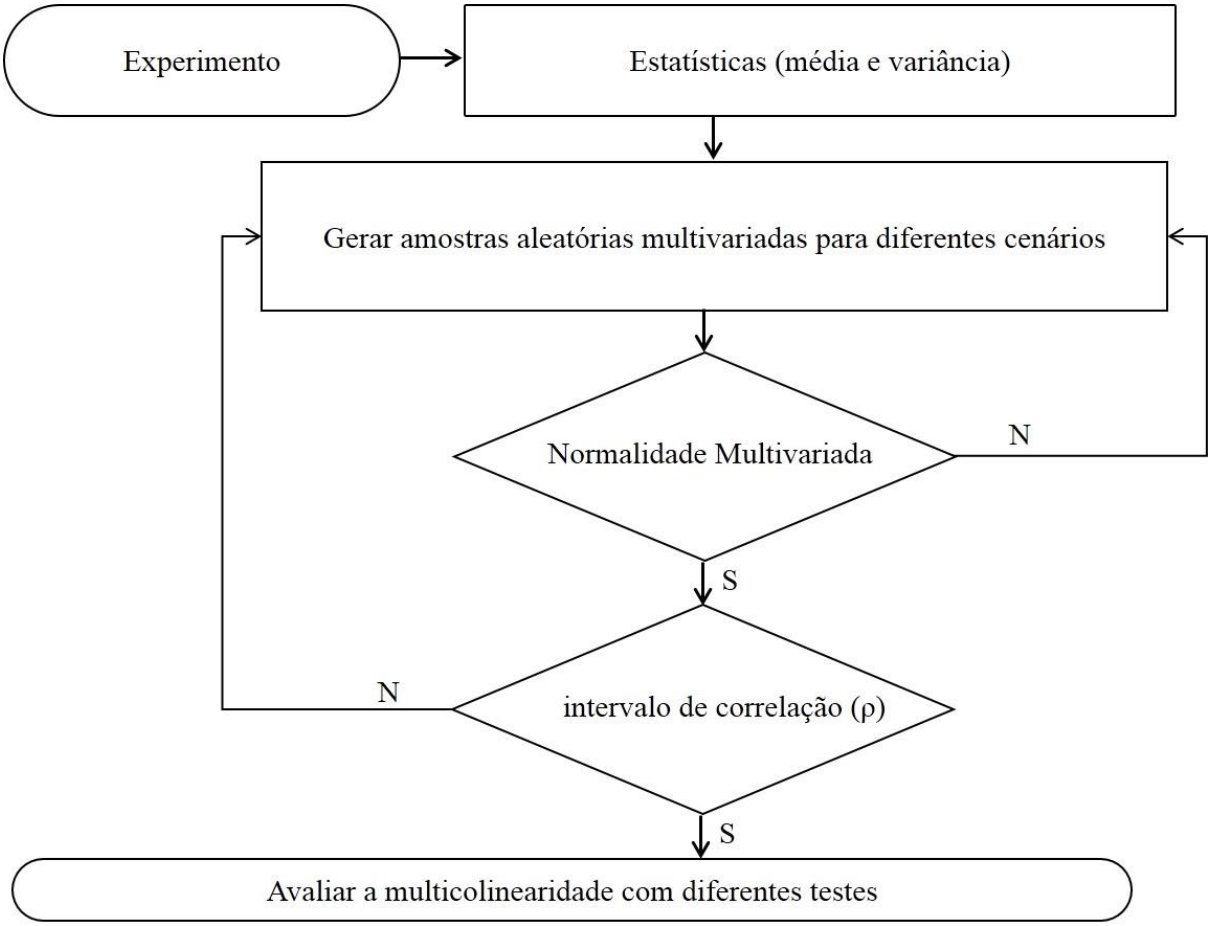
A sequência de execução está apresentada no fluxograma da figura 1. Os valores de correlação escolhidos para compor a matriz sigma foram: zero para a condição de baixa

correlação; 0,5 para a condição de correlação intermediária; e 0,9 para a condição de correlação alta.

O próximo passo da programação foi gerar amostras multivariadas de tamanho n, com p variáveis utilizando a função “mvrnorm” do pacote MASS, calculada para um determinado nível de correlação. A seguir, a amostra simulada foi verificada se está de acordo com as condições de cada cenário, sendo a condição de normalidade multivariada avaliada pelo teste de Shapiro-Francia’s multivariado (SILVA, 2009), bem como, verificar se as correlações entre as variáveis estão dentro dos níveis estabelecidos. Sendo uma destas condições não atendida, o processo reinicia e uma nova amostra é gerada.

A partir das verificações de que a amostra foi gerada corretamente, inicia-se a avaliação da amostra quanto ao diagnóstico de presença ou ausência de multicolinearidade através do determinante da matriz de correlação (DET), número de condição (NC), fator de inflação de variância (FIV), teste de Farrar e Glauber (FG) e teste de Haitovsky (H). Para cada um dos cenários foram obtidas 1000 amostras multivariadas e a seguir quantificado o percentual de indicação de presença de multicolinearidade pelos testes avaliados. O determinante da matriz de correlação foi escolhido em função da sua facilidade de obtenção e interpretação, o número de condição e fator de inflação de variância pela sua ampla utilização para a avaliação da multicolinearidade. Os teste de Farrar e Glauber e Haitovsky por serem teste estatísticos que consideram em suas equações o número de variáveis envolvidas e também o tamanho de amostra utilizada.

Figura 1. Fluxograma de execução da pesquisa.



2.3 IDENTIFICAÇÃO DE MULTICOLINEARIDADE

Os testes para avaliação da presença ou ausência de multicolinearidade, utilizados foram: o determinante da matriz de correlação (DET); número de condição (NC); fator de inflação de variância (FIV); teste de Farrar e Glauber (FG); e Teste de Haitovsky (H). Os testes de Farrar e Glauber (FG) e Teste de Haitovsky (H) são testes estatísticos e avaliam a multicolinearidade através de teste de hipóteses, indicando assim se existe ou não a multicolinearidade com base em uma probabilidade de erro. Os demais testes foram utilizados critérios de decisão estabelecidos pela literatura para determinar a presença ou não da multicolinearidade.

A correlação linear de Pearson pode ser obtida para um par de caracteres x e y, por meio da expressão, adaptada de Cruz e Regazzi (1997) e Ferreira (2009):

$$r = \frac{n.\sum_{i=1}^{n} X_{1i}.X_{2i} - (\sum_{i=1}^{n} X_{1i}).(\sum_{i=1}^{n} X_{2i})}{\sqrt{[n.\sum_{i=1}^{n} X_{1i}^2 - (\sum_{i=1}^{n} X_{1i})^2].[n.\sum_{i=1}^{n} X_{2i}^2 - (\sum_{i=1}^{n} X_{2i})^2]}}$$

O primeiro teste avaliado foi o determinante da matriz de correlação, que é utilizado como um indicador de multicolinearidade. O determinante varia entre 0 e 1, sendo que 0 indica que a matriz é singular (correlações perfeitas) e 1 indica que não há qualquer tipo de correlação entre as variáveis. Como regra geral, o valor do determinante deve ser superior a 0,00001, para indicar ausência de multicolinearidade (FIELD, 2009).

O número de condição, que é obtido pela divisão do maior pelo menor autovalor da matriz de correlação X'X, tem sido largamente utilizado no diagnóstico do grau de multicolinearidade (CRUZ; CARNEIRO, 2006; GUJARATI; PORTER, 2011).

$NC = \frac{\lambda_{máx}}{\lambda_{mín}}$, onde $\lambda_{máx}$ é o maior autovalor e $\lambda_{mín}$ o menor auto valor da matriz de correlação entre as variáveis. Esse critério de diagnóstico avalia o grau de multicolinearidade sobre toda a matriz de correlação. Quando existe uma ou mais relações de dependências lineares entre as variáveis, um ou mais autovalores (λ) da matriz de correlação serão muito próximos a zero ou nulo se a relação for perfeita (CRUZ; CARNEIRO, 2006). O critério do número de condição (NC), apresenta três classificações segundo Montgomery et al. (2012): número de condição menor do que 100, a multicolinearidade é considerada fraca, entre 100 e 1000 é moderada e maior do que 1000, a multicolinearidade é severa. Neste trabalho foram consideradas as duas faixas de multicolinearidade, moderada (NCM) e severa (NCS).

Outro critério de diagnóstico de multicolinearidade é o exame dos fatores de inflação de variância (FIV), definidos por:

$FIV = \frac{1}{1-R_x^2}$, em que: R_x^2 é o coeficiente de determinação da regressão da variável explicativa X em função das demais variáveis presentes (FÁVERO et al., 2009). Diferentemente do número de condição, o FIV apresenta um valor para cada variável, sendo a variável que apresenta o maior valor de FIV que deve ser avaliada quanto à condição de multicolinearidade. O FIV pode ser obtido alternativamente, na diagonal da inversa da matriz de correlação X'X, sendo cada elemento da diagonal, o valor de FIV para cada variável. Assim, para a variável X1, obtêm-se o FIV na primeira linha e primeira coluna da inversa da matriz de correlação X'X, de forma semelhante, para as demais variáveis, busca-se os valores de FIV de cada variável na diagonal da inversa da matriz de correlação X'X (CRUZ; CARNEIRO, 2006; HAIR et al., 2009). Segundo Fávero et al. (2009), Freund e Wilson (1998), Hair et al. (2009), Kutner et al. (2003)

e Gujarati e Porter (2011), Marquardt (1970) e Kleinbaum (1988 apud GUJARATI; PORTER, 2011), quando FIV é maior do que 10, indica a ocorrência de multicolinearidade.

A multicolinearidade também pode ser avaliada através do teste FG, desenvolvido por Farrar e Glauber (1967). O teste consiste de uma adaptação do teste de esfericidade de Bartlett (BARTLETT, 1937) para avaliar a multicolinearidade. O índice FG é obtido pela seguinte equação:

$FG = -\left[n - 1 - \frac{2p+5}{6}\right] * ln|X'X|$, onde n é o tamanho de amostra, p é o número de variáveis e $ln|X'X|$ é o logaritmo natural da matriz de correlação entre as variáveis do modelo. A matriz $X'X = V\Lambda V'$, onde V é a matriz dos autovetores, Λ é a matriz diagonal dos autovalores e V' é a matriz transposta dos autovetores. Essa estatística tem distribuição aproximadamente qui-quadrado (χ^2) com ($\frac{p*(p-1)}{2}$) graus de liberdade. E segundo os autores, essa estatística é útil para identificar a multicolinearidade pois oferece uma medida generalizada, por meio da padronização do tamanho de amostra (n) e do número de variáveis (p).

O teste de Haitovsky (1969) possibilita testar se o determinante da matriz de correlação difere ou não de zero. Sendo que o determinante diferente de zero (p-valor < 0,05) indica que a matriz $X'X$ é possível de inversão, possibilitando assim o cálculo dos parâmetros do modelo.

O teste de Haitovsky (1969) verifica se a matriz de correlação é singular devido a muitas correlações altas entre as variáveis. Uma matriz não singular é aquela que mostra relativamente baixas correlações entre as variáveis preditoras. A matriz não singular tem um determinante que está perto de 1,0. O teste de Haitovsky verifica se o determinante da matriz de correlação difere ou não de zero. Deste modo, a hipótese nula do teste de Haitovsky afirma que o determinante não difere de zero, e nesta condição têm-se problemas de multicolinearidade. O teste de Haitovsky pode ser obtido através da equação:

$H = -\left[1 + \frac{2p+5}{6} - n\right] * \ln(1 - |X'X|)$, onde p é o número de variáveis, n é o tamanho de amostra e $|X'X|$ é a matriz de correlação entre as variáveis do modelo. Essa estatística também tem distribuição aproximadamente qui-quadrado (χ^2) com ($\frac{p*(p-1)}{2}$) graus de liberdade. Assim, a diferença estatisticamente significativa indica que multicolinearidade não é um problema porque a matriz de correlação de variáveis de previsão não é singular (LANGASKENS, 1975).

Tanto para o índice FG quanto para o teste de Haitovsky, o nível de significância do teste considerado foi de 5% de probabilidade de erro. Com os dois testes estatísticos foi realizado a verificação da taxa de erro tipo I e poder do teste. A taxa de erro tipo I foi verificada através do percentual de multicolinearidade em condições de baixa correlação entre variáveis e

poder do teste, pelo percentual de indicação de multicolinearidade em condições de alta correlação entre as variáveis.

Com os valores médios das 1000 simulações do determinante da matriz de correlação e número de condição serão elaborados gráficos de superfície. A fim de observar o padrão de resposta destes critérios de avaliação da multicolinearidade à medida que são alterados o número de variáveis e tamanho de amostra.

3 RESULTADOS E DISCUSSÃO

Dentre as variáveis analisadas no experimento de tomateiro, as variáveis produtivas de número de inflorescências e número de frutos apresentam maior variabilidade, ao passo que as variáveis morfológicas, altura de planta, número de folhas e diâmetro do caule, apresentam variabilidade menor (Tabela 1). A variabilidade das variáveis produtivas é maior nas primeiras avaliações em função da baixa médias de frutos, sendo que muitas plantas ainda não tem frutos, proporcionando um grande número de plantas com valor zero, o que aumenta o coeficiente de variação em função da baixa média da variável (LÚCIO et al. 2010). Essa diferenciação de variabilidade entre os caracteres morfológicos e produtivos também foram encontrados por Brunes (2013) na cultura do pimentão. Deste modo, as variáveis foram separadas em dois grupos: morfológicas e produtivas para a realização do trabalho.

Tabela 1. Estatísticas descritivas das variáveis morfológicas e produtivas de 66 plantas de tomateiro, cultivadas sob túnel plástico.

(continua)

Variáveis (unidade de medida)	Dias após transplante	Sigla	Média	Variância
Altura de planta (cm)	47	AP47	55,64	50,23
Número de folhas (unidades)	47	NF47	14,65	1,62
Diâmetro do caule (cm)	47	DC47	8,35	1,12
Número de inflorescências (unidades)	47	NI47	2,64	0,36
Número de frutos (unidades)	47	NFR47	0,21	0,29
Altura de planta (cm)	61	AP61	83,38	112,52
Número de folhas (unidades)	61	NF61	18,56	5,76
Diâmetro do caule (cm)	61	DC61	9,98	1,12
Número de inflorescências (unidades)	61	NI61	3,80	0,31
Número de frutos (unidades)	61	NFR61	1,12	1,59
Altura de planta (cm)	75	AP75	116,52	159,73
Número de folhas (unidades)	75	NF75	20,18	4,83
Diâmetro do caule (cm)	75	DC75	11,38	1,04

(conclusão)

Número de inflorescências (unidades)	75	NI75	4,97	1,14
Número de frutos (unidades)	75	NFR75	11,95	25,98
Altura de planta (cm)	88	AP88	136,86	171,20
Número de folhas (unidades)	88	NF88	21,24	4,49
Diâmetro do caule (cm)	88	DC88	13,20	2,75
Número de inflorescências (unidades)	88	NI88	6,70	2,28
Número de frutos (unidades)	88	NFR88	19,83	45,59

Nas tabelas 2 e 3 estão apresentados os valores de correlação linear de Pearson paras as variáveis morfológicas e produtivas, respectivamente. A avaliação destas correlações originais dos dados coletados no experimentos permite visualizar que as correlações variam de -0,17 até 0,90, mostrando grande amplitude de graus de correlacionamento entre as variáveis avaliadas.

Tabela 2. Correlação linear de Pearson acima da diagonal principal e p-valor do teste t abaixo da diagonal principal para variáveis morfológicas de tomateiro.

	AP47	AP61	AP75	AP88	DC47	DC61	DC75	DC88	NF47	NF61	NF75	NF88
AP47	1	0,85*	0,80*	0,69*	0,52*	0,24^{ns}	0,09^{ns}	0,00^{ns}	0,63*	0,56*	0,37*	0,41*
AP61	0,000	1	0,90*	0,76*	0,68*	0,51*	0,39*	0,11^{ns}	0,68*	0,75*	0,52*	0,48*
AP75	0,000	0,000	1	0,89*	0,56*	0,40*	0,30*	0,07^{ns}	0,64*	0,72*	0,53*	0,48*
AP88	0,000	0,000	0,000	1	0,48*	0,28*	0,27*	0,06^{ns}	0,48*	0,59*	0,49*	0,51*
DC47	0,000	0,000	0,000	0,000	1	0,58*	0,40*	0,21^{ns}	0,46*	0,60*	0,42*	0,43*
DC61	0,055	0,000	0,001	0,021	0,000	1	0,69*	0,49*	0,42*	0,46*	0,23^{ns}	0,17^{ns}
DC75	0,495	0,001	0,015	0,028	0,001	0,000	1	0,57*	0,36*	0,32*	0,08^{ns}	0,05^{ns}
DC88	0,973	0,382	0,593	0,631	0,084	0,000	0,000	1	0,18^{ns}	0,11^{ns}	-0,05^{ns}	0,04^{ns}
NF47	0,000	0,000	0,000	0,000	0,000	0,000	0,003	0,150	1	0,63*	0,36*	0,34*
NF61	0,000	0,000	0,000	0,000	0,000	0,000	0,009	0,374	0,000	1	0,72*	0,58*
NF75	0,002	0,000	0,000	0,000	0,000	0,060	0,530	0,677	0,003	0,000	1	0,88*
NF88	0,001	0,000	0,000	0,000	0,000	0,165	0,694	0,757	0,005	0,000	0,000	1

* Significativo pelo teste t a 5% de probabilidade de erro. ^{ns} Não significativo pelo teste t a 5% de probabilidade de erro.

Tabela 3. Correlação linear de Pearson acima da diagonal principal e p-valor do teste t abaixo da diagonal principal para variáveis produtivas de tomateiro.

	NFR47	NFR61	NFR75	NFR88	NI47	NI61	NI75	NI88
NFR47	1	0,46*	-0,10^{ns}	-0,14^{ns}	-0,04^{ns}	-0,16^{ns}	-0,10^{ns}	-0,17^{ns}
NFR61	0,000	1	0,31*	0,16^{ns}	0,20^{ns}	0,14^{ns}	0,22^{ns}	0,18^{ns}
NFR75	0,439	0,010	1	0,82*	0,49*	0,70*	0,75*	0,77*
NFR88	0,256	0,205	0,000	1	0,54*	0,73*	0,79*	0,85*
NI47	0,731	0,103	0,000	0,000	1	0,56*	0,54*	0,35*
NI61	0,187	0,251	0,000	0,000	0,000	1	0,71*	0,67*
NI75	0,446	0,075	0,000	0,000	0,000	0,000	1	0,81*
NI88	0,185	0,144	0,000	0,000	0,004	0,000	0,000	1

* Significativo pelo teste t a 5% de probabilidade de erro. ^{ns} Não significativo pelo teste t a 5% de probabilidade de erro.

A análise dos resultados dos testes de avaliação da multicolinearidade através do percentual de casos com presença de multicolinearidade em diferentes cenários de grau de correlação entre as variáveis, número de variáveis avaliadas e tamanho de amostra considerado, permite avaliar o desempenho do teste quanto a coerência em cada cenário. Na tabela 4, são apresentados as porcentagens de casos em que os testes indicaram a presença de multicolinearidade em condições de alta correlação entre as variáveis ($r > 0,8$), na qual se verifica que o único teste que não foi capaz de identificar a condição de multicolinearidade, foi o teste do determinante. Não houve indicação de multicolinearidade quando, apenas duas variáveis estavam presentes, independentemente do tamanho de amostra utilizado. Esse mesmo teste também não foi capaz de indicar a presença de multicolinearidade em todos os casos simulados, quando três variáveis foram utilizadas e tamanhos de amostras foram menores do que 100.

O número de condição foi avaliado em duas classes de indicação de multicolinearidade, sendo elas, moderada para o número de condição entre 100 e 1000 (NCM) e severo quando o número de condição é maior do que 1000 (NCS) (MONTGOMERY et al., 2012). Este critério de avaliação da multicolinearidade apresentou 100% de indicação de multicolinearidade moderada nos cenários com alta correlação entre as variáveis. Já para a classe de multicolinearidade severa, o número de condição indicou percentual diferente de 100% apenas quando duas variáveis estão presentes e o tamanho de amostra é menor ou igual a 100, embora os percentuais de indicação de multicolinearidade nessa condição sejam elevados ($\geq 96,3\%$). O FIV e os testes FG e H identificaram corretamente a presença de multicolinearidade em todos os cenários simulados. Os testes FG e H evidenciaram elevado poder em todos os cenários simulados, pois indicaram a condição de multicolinearidade quando esta estava presente.

Tabela 4. Porcentagem de multicolinearidade detectada pelo determinante (DET), número de condição moderado (NCM), número de condição severo (NCS), fator de inflação de variância (FIV), teste de Farrar e Glauber (FG) e teste de Haitovsky (H) sob alta correlação (r > 0,8), em 1000 simulações para variáveis morfológicas. (continua)

Número de variáveis (p)	Teste	Tamanho de amostra (n)								
		25	30	40	50	100	150	200	500	1000
2	DET	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	NCM	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	NCS	96,3	97,5	98,0	99,4	99,9	100,0	100,0	100,0	100,0
	FIV	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	FG	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	H	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
3	DET	96,7	97,4	98,6	99,1	100,0	100,0	100,0	100,0	100,0
	NCM	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	NCS	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	FIV	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	FG	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	H	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
4	DET	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	NCM	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	NCS	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	FIV	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	FG	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	H	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
5	DET	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	NCM	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	NCS	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	FIV	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	FG	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	H	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
6	DET	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	NCM	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	NCS	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	FIV	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	FG	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	H	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
7	DET	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	NCM	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	NCS	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	FIV	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	FG	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	H	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
8	DET	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	NCM	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	NCS	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	FIV	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	FG	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	H	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
9	DET	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	NCM	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	NCS	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	FIV	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	FG	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	H	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
10	DET	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	NCM	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	NCS	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0

(conclusão)

	FIV	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	FG	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	H	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
11	DET	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	NCM	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	NCS	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	FIV	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	FG	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	H	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
12	DET	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	NCM	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	NCS	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	FIV	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	FG	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	H	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0

Na tabela 5, estão apresentados os resultados dos testes de multicolinearidade também para a condição de alta correlação entre as variáveis produtivas de tomateiro. O padrão de resposta dos testes foi semelhante ao apresentado para as variáveis morfológicas (Tabela 4) discutidos anteriormente. O teste do determinante da matriz de correlação entre as variáveis não foi capaz de indicar multicolinearidade, quando apenas duas variáveis estavam presentes, independentemente do tamanho de amostra. O número de condição, na classe severo, também não indicou multicolinearidade severa, em 100% de casos quando o número de variáveis é apenas duas e o tamanho de amostra é menor ou igual a 50 plantas avaliadas. Os demais testes indicaram presença de multicolinearidade em todas as simulações, indicando assim que os testes de FG e H, que são testes de hipóteses e, portanto, que seguem uma distribuição de probabilidade conhecida, no caso a distribuição Qui-Quadrado, apresentam alto poder do teste, pois indicam multicolinearidade quando esta está presente.

Tabela 5. Porcentagem de multicolinearidade detectada pelo determinante (DET, número de condição moderado (NCM), número de condição severo (NCS), fator de inflação de variância (FIV), teste de Farrar e Glauber (FG) e teste de Haitovsky (H) sob alta correlação (r > 0,8), em 1000 simulações para variáveis produtivas. (continua)

Número de variáveis (p)	Teste	Tamanho de amostra (n)								
		25	30	40	50	100	150	200	500	1000
2	DET	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	NCM	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	NCS	95,6	97,0	98,1	99,5	100,0	100,0	100,0	100,0	100,0
	FIV	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	FG	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	H	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
3	DET	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	NCM	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	NCS	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	FIV	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0

(conclusão)

	FG	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	H	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
4	DET	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	NCM	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	NCS	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	FIV	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	FG	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	H	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
5	DET	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	NCM	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	NCS	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	FIV	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	FG	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	H	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
6	DET	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	NCM	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	NCS	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	FIV	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	FG	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	H	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
7	DET	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	NCM	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	NCS	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	FIV	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	FG	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	H	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
8	DET	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	NCM	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	NCS	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	FIV	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	FG	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	H	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0

Na tabela 6 e 7 estão apresentados os resultados dos testes de avaliação de multicolinearidade para a condição de baixa correlação entre as variáveis morfológicas e produtivas, respectivamente. Para correlação baixa entre as variáveis envolvidas, variando de 0 até 0,3, verificaram-se que os testes do determinante da matriz de correlação (DET) entre as variáveis, o teste do número de condição (NC) e o teste do fator de inflação da variância (FIV) foram corretos em não indicar presença de multicolinearidade, pois as variáveis estão pouco correlacionadas.

Os teste de Farrar e Glauber (FG) e Haitovsky (H) apresentaram casos com indicativo de multicolinearidade, mesmo com correlação baixa entre as variáveis. Porém, o teste FG apresentou erro tipo I, ou seja, indicou presença de multicolinearidade quando esta não está presente, principalmente em condições de tamanhos de amostra maiores e muito próximos ao limite de significância do teste, tanto para as variáveis morfológicas quanto para as variáveis produtivas (Tabelas 6 e 7).

O teste H indicou presença de multicolinearidade principalmente para tamanhos de amostra menores e número grande de variáveis envolvidas. A taxa de erro tipo I do teste de Haitovisky foi bastante elevada, chegando a 100% em alguns casos, tanto para variáveis morfológicas quanto produtivas (Tabelas 6 e 7).

Tabela 6. Porcentagem de multicolinearidade detectada pelo determinante (DET), número de condição moderado (NCM), número de condição severo (NCS), fator de inflação de variância (FIV), teste de Farrar e Glauber (FG) e teste de Haitovsky (H) sob baixa correlação ($0 < r < 0,3$), em 1000 simulações para variáveis morfológicas. (continua)

Número de variáveis (p)	Teste	Tamanho de amostra (n)								
		25	30	40	50	100	150	200	500	1000
2	DET	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	NCM	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	NCS	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	FIV	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	FG	0,0	0,0	0,0	1,0	4,1	6,1	5,3	4,4	3,4
	H	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
3	DET	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	NCM	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	NCS	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	FIV	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	FG	0,0	0,0	0,2	0,6	3,7	5,9	6,2	5,8	4,7
	H	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
4	DET	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	NCM	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	NCS	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	FIV	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	FG	0,0	0,0	0,0	0,3	3,9	4,6	5,3	4,2	4,1
	H	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
5	DET	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	NCM	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	NCS	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	FIV	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	FG	0,4	0,0	0,2	0,3	3,7	5,1	5,0	6,2	5,1
	H	0,4	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
6	DET	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	NCM	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	NCS	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	FIV	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	FG	0,0	0,1	0,0	0,2	3,5	6,2	4,6	4,8	4,7
	H	53,5	10,7	0,0	0,0	0,0	0,0	0,0	0,0	0,0
7	DET	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	NCM	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	NCS	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	FIV	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	FG	0,0	0,0	0,0	0,5	3,0	4,4	5,2	4,2	6,1
	H	99,8	94,5	20,4	0,9	0,0	0,0	0,0	0,0	0,0
8	DET	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	NCM	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	NCS	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	FIV	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	FG	0,0	0,0	0,0	0,1	3,0	5,5	4,1	5,0	4,7
	H	100,0	100,0	95,1	43,3	0,0	0,0	0,0	0,0	0,0
9	DET	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0

(conclusão)

	NCM	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	NCS	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	FIV	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	FG	0,0	0,0	0,0	0,0	5,0	5,2	4,8	4,8	6,1
	H	100,0	100,0	100,0	98,3	0,0	0,0	0,0	0,0	0,0
10	DET	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	NCM	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	NCS	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	FIV	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	FG	0,0	0,0	0,0	0,0	3,0	4,4	4,0	5,9	5,5
	H	100,0	100,0	100,0	100,0	0,3	0,0	0,0	0,0	0,0
11	DET	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	NCM	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	NCS	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	FIV	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	FG	0,0	0,0	0,0	0,0	3,5	4,5	5,3	5,9	5,0
	H	100,0	100,0	100,0	100,0	31,9	0,0	0,0	0,0	0,0
12	DET	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	NCM	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	NCS	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	FIV	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	FG	0,0	0,0	0,0	0,1	3,5	5,9	4,7	4,2	5,9
	H	100,0	100,0	100,0	100,0	94,3	0,0	0,0	0,0	0,0

Tabela 7. Porcentagem de multicolinearidade detectada pelo determinante (DET), número de condição moderado (NCM), número de condição severo (NCS), fator de inflação de variância (FIV), teste de Farrar e Glauber (FG) e teste de Haitovsky (H) sob baixa correlação ($0 < r < 0,3$) em 1000 simulações para variáveis produtivas. (continua)

Número de variáveis (p)		Tamanho de amostra (n)								
		25	30	40	50	100	150	200	500	1000
2	DET	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	NCM	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	NCS	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	FIV	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	FG	0,0	0,0	0,0	2,9	4,6	4,0	3,9	4,9	4,2
	H	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
3	DET	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	NCM	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	NCS	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	FIV	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	FG	0,0	0,0	0,0	0,1	0,8	4,1	4,6	4,0	5,0
	H	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
4	DET	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	NCM	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	NCS	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	FIV	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	FG	0,0	0,0	0,5	0,2	5,2	5,3	5,3	5,3	5,3
	H	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
5	DET	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	NCM	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	NCS	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	FIV	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	FG	0,0	0,0	0,2	0,3	4,2	4,7	5,6	7,1	5,2
	H	0,2	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
6	DET	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0

(conclusão)

	NCM	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	NCS	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	FIV	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	FG	0,0	0,0	0,5	3,3	5,1	4,6	6,5	4,4	5,2
	H	51,4	10,3	0,0	0,0	0,0	0,0	0,0	0,0	0,0
7	DET	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	NCM	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	NCS	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	FIV	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	FG	0,0	0,0	0,0	0,5	2,7	4,9	4,7	5,2	4,6
	H	99,6	93,6	23,2	0,6	0,0	0,0	0,0	0,0	0,0
8	DET	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	NCM	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	NCS	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	FIV	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	FG	0,0	0,0	0,0	0,2	3,9	5,6	4,9	5,5	4,9
	H	100,0	100,0	94,7	45,4	0,0	0,0	0,0	0,0	0,0

Nas tabelas 8 e 9, estão os resultados comparando os testes para os cenários com grau de correlação intermediária entre as variáveis para as variáveis morfológicas e produtivas respectivamente. Verificam-se que os testes do determinante da matriz de correlação (DET), número de condição (NC) e fator de inflação da variância (FIV) não indicaram multicolinearidade em nenhuma das combinações de número de variáveis e tamanho de amostra, tanto para as variáveis morfológicas quanto para as variáveis produtivas.

Os testes FG e H indicaram presença de multicolinearidade em alguns cenários, principalmente para maior número de variáveis e tamanhos de amostra pequenos, tanto para variáveis produtivas quanto morfológicas. O teste FG indicou presença de multicolinearidade, em 100% dos casos nessa condição de correlação intermediária entre as variáveis, em todos os cenários e em ambos os tipos de variáveis. Esse é um teste mais sensível, já que indica a presença de multicolinearidade, mesmo quando outros testes não indicam esta condição. O teste H, assim como o teste FG, indicou grande percentual de multicolinearidade, em vários cenários, com correlação intermediária entre as variáveis. O teste H não indicou multicolinearidade apenas nos cenários com duas variáveis, e a partir de três variáveis passou a indicar multicolinearidade, principalmente, em cenários de tamanhos de amostra pequenos, aumentando o percentual de multicolinearidade com o aumento do tamanho de amostra e número de variáveis. Assim, o teste H também apresenta maior sensibilidade que o teste do determinante da matriz de correlação, número de condição e fator de inflação de variância.

Tabela 8. Porcentagem de multicolinearidade detectada pelo determinante (DET), número de condição moderado (NCM), número de condição severo (NCS), fator de inflação de variância (FIV), teste de Farrar e Glauber (FG) e teste de Haitovsky (H) sob média correlação (0,4 < r < 0,7) em 1000 simulações para variáveis morfológicas. (continua)

Número de variáveis (p)	Teste	Tamanho de amostra (n)								
		25	30	40	50	100	150	200	500	1000
2	DET	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	NCM	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	NCS	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	FIV	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	FG	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	H	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
3	DET	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	NCM	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	NCS	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	FIV	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	FG	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	H	6,5	0,4	0,0	0,0	0,0	0,0	0,0	0,0	0,0
4	DET	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	NCM	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	NCS	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	FIV	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	FG	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	H	100,0	99,8	88,9	53,5	0,0	0,0	0,0	0,0	0,0
5	DET	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	NCM	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	NCS	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	FIV	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	FG	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	H	100,0	100,0	100,0	100,0	80,7	8,2	0,1	0,0	0,0
6	DET	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	NCM	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	NCS	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	FIV	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	FG	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	H	100,0	100,0	100,0	100,0	100,0	100,0	89,3	0,0	0,0
7	DET	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	NCM	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	NCS	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	FIV	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	FG	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	H	100,0	100,0	100,0	100,0	100,0	100,0	100,0	65,2	0,0
8	DET	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	NCM	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	NCS	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	FIV	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	FG	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	H	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	85,9
9	DET	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	NCM	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	NCS	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	FIV	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	FG	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	H	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
10	DET	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	NCM	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	NCS	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	FIV	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0

(conclusão)

	FG	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	H	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
11	DET	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	NCM	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	NCS	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	FIV	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	FG	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	H	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
12	DET	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	NCM	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	NCS	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	FIV	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	FG	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	H	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0

Tabela 9. Porcentagem de multicolinearidade detectada pelo determinante (DET), número de condição moderado (NCM), número de condição severo (NCS), fator de inflação de variância (FIV), teste de Farrar e Glauber (FG) e teste de Haitovsky (H) sob média correlação (0,4 < r < 0,7) em 1000 simulações para variáveis produtivas. (continua)

Número de variáveis (p)	Teste	Tamanho de amostra (n)								
		25	30	40	50	100	150	200	500	1000
2	DET	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	NCM	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	NCS	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	FIV	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	FG	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	H	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
3	DET	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	NCM	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	NCS	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	FIV	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	FG	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	H	7,3	0,5	0,0	0,0	0,0	0,0	0,0	0,0	0,0
4	DET	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	NCM	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	NCS	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	FIV	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	FG	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	H	100,0	99,4	88,7	54,6	0,0	0,0	0,0	0,0	0,0
5	DET	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	NCM	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	NCS	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	FIV	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	FG	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	H	100,0	100,0	100,0	100,0	83,9	10,9	0,2	0,0	0,0
6	DET	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	NCM	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	NCS	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	FIV	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	FG	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	H	100,0	100,0	100,0	100,0	100,0	100,0	88,8	0,0	0,0
7	DET	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	NCM	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0

(conclusão)

	NCS	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	FIV	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	FG	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	H	100,0	100,0	100,0	100,0	100,0	100,0	100,0	63,8	0,0
8	DET	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	NCM	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	NCS	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	FIV	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	FG	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	H	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	83,9

Os testes do número de condição moderado (NCM) e fator de inflação de variância (FIV) concordam na indicação de ausência de multicolinearidade em todos os cenários simulados para os casos de correlação alta, baixa e intermediária entre as variáveis, tanto para as variáveis morfológicas, quanto para as produtivas. Ambos os testes são amplamente utilizados para a finalidade de identificação da multicolinearidade e no presente estudo apresentaram desempenho equivalentes.

Nas figuras 2 e 3 estão representados graficamente os valores médios do teste do número de condição (NC) das 1000 simulações na condição de alta correlação entre as variáveis, em cada uma das condições analisadas de número de variáveis e tamanhos de amostra, para as variáveis com baixa variabilidade (morfológicas) e alta variabilidade (produtivas), respectivamente. Nota-se um padrão de aumento do número médio de NC para cenários com maior número de variáveis e tamanhos de amostra pequenos. Esse comportamento do número de condição ser mais elevado quando se têm muitas variáveis e tamanho de amostra pequeno e pouco superior ao número de variáveis, foi relatado por Goldberger (1991), que denomina este problema de "micronumerosidade", ou seja, amostras pequenas. Nessa condição, a multicolinearidade está associada a micronumerosidade, ou seja, quando o tamanho de amostra é pequeno e pouco superior ao número de variáveis, tem-se o problema de multicolinearidade agravado.

Para Hair et al. (2009) o recomendado é que sejam amostrados pelo menos cinco vezes mais observações, que o número de variáveis envolvidas na análise. Tabachnick e Fidell (2013) sugerem utilizar tamanho de amostra igual ou superior a 50 + 80X, onde X representa o número de variáveis independentes. Já Stevens (1996) recomenda uma proporção de 15 observações para cada variável observada a fim de ampliar a confiabilidade das estimativas. Alabi et al. (2007) relatou que o aumento do tamanho de amostra reduz a taxa de erro tipo II do estimador de mínimos quadrados em todos os níveis de multicolinearidade estudados. Clements e Hendry (1995) relataram que as estimativas dos parâmetros podem ser mal determinados devido ao

grande número de variáveis, devido ao agravamento do grau de multicolinearidade. Para Mason e Perreault (1991) a avaliação da multicolinearidade deve ser interpretada em conjunto com o tamanho de amostra e número de variáveis envolvidas na análise.

Ao analisar os valores médios do teste do número de condição (NC) na condição de correlação baixa, apresentados nas figuras 4 e 5 para variáveis com baixa variabilidade (variáveis morfológicas) e alta variabilidade (variáveis produtivas) respectivamente; e, para a condição de correlação intermediária, apresentadas nas figuras 6 e 7 para variáveis com baixa variabilidade (variáveis morfológicas) e alta variabilidade (variáveis produtivas), respectivamente, nota-se o mesmo padrão de aumento do valor médio do NC para cenários com maior número de variáveis e tamanhos de amostra pequenos, conforme pode ser visualizado nas figuras 2 e 3 para a condição de alta correlação. Assim, pode-se concluir que independentemente da condição de correlação entre as variáveis, o comportamento do valor do número de condição ser mais elevado, quando têm-se muitas variáveis e tamanho de amostra pequeno, permanece o mesmo.

De forma semelhante pode se observar que a média do determinante da matriz de correlação das 1000 simulações de cada cenário também segue um padrão de comportamento à medida que o número de variáveis e tamanho de amostra são alterados (Figuras 8 e 9). A medida que o tamanho de amostra diminui e o número de variáveis aumenta o valor do determinante da matriz de correlação tende a diminuir, provocando maior tendência a multicolinearidade.

A variabilidade presente nas variáveis não influenciou nos testes de multicolinearidade estudados, não sendo possível verificar grande diferença na identificação de multicolinearidade pelos testes analisados quanto ao grau de variabilidade das variáveis. No entanto, o número de faixas de variabilidade foi de apenas duas. São necessários novos estudos utilizando maior número de faixas de variabilidade das variáveis para comprovação deste resultado.

Figura 2. Média do número de condição (NC) para 1000 simulações em cenários de alta correlação entre as variáveis em diferentes números de variáveis (p) e tamanhos de amostra (n) de variáveis morfológicas.

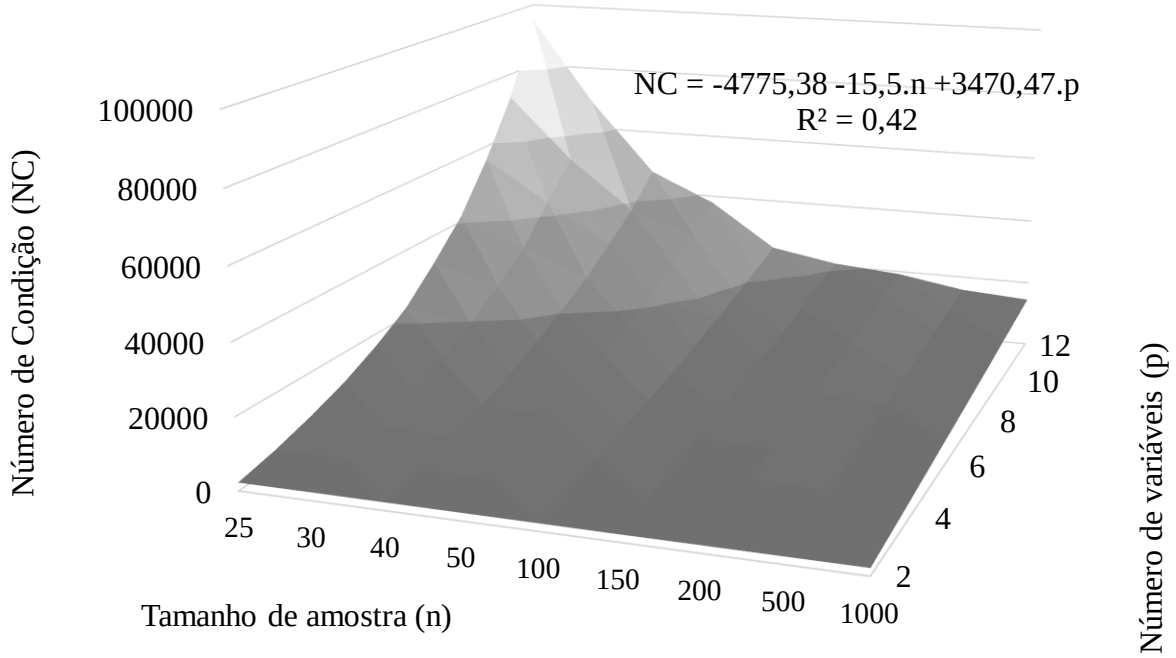


Figura 3. Média do número de condição (NC) para 1000 simulações em cenários de alta correlação entre as variáveis em diferentes números de variáveis (p) e tamanhos de amostra (n) de variáveis produtivas.

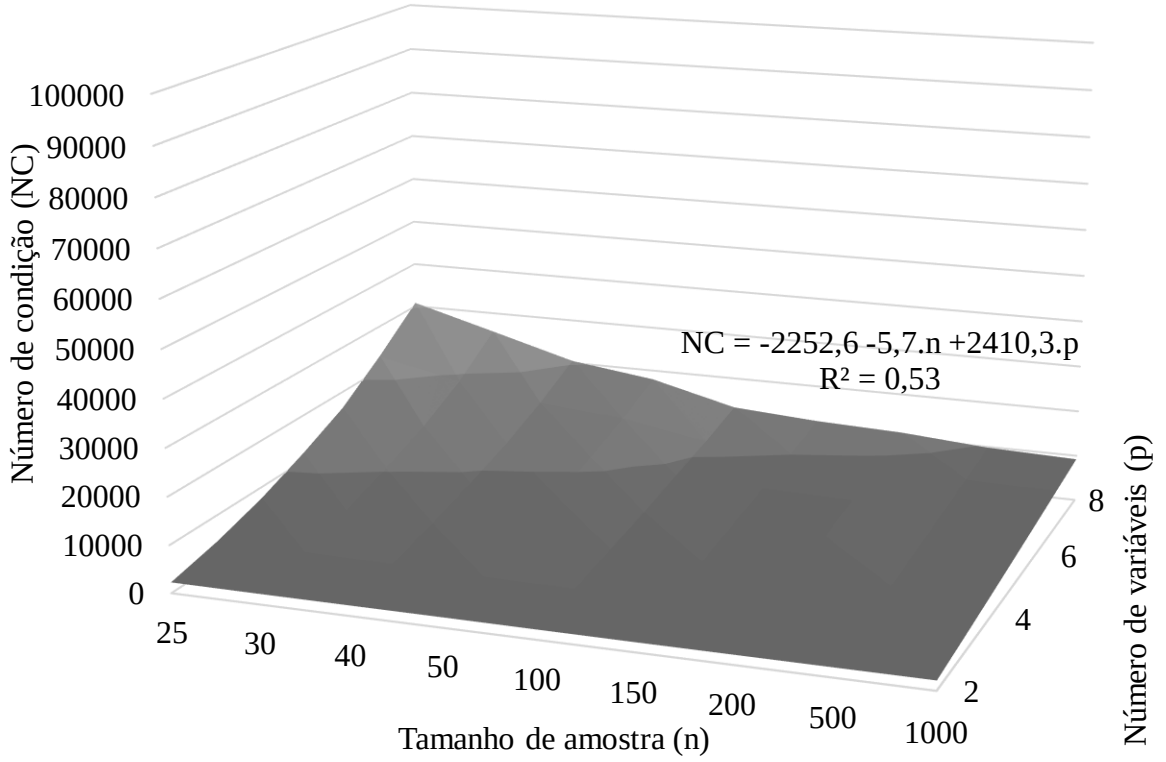


Figura 4. Média do número de condição (NC) para 1000 simulações em cenários de baixa correlação entre as variáveis em diferentes números de variáveis (p) e tamanhos de amostra (n) de variáveis morfológicas.

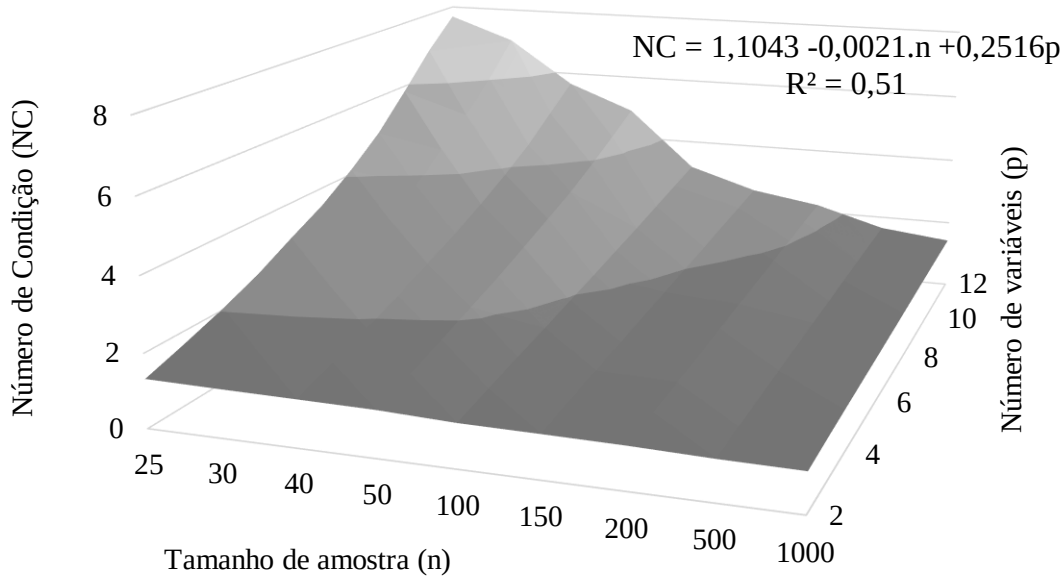


Figura 5. Média do número de condição (NC) para 1000 simulações em cenários de baixa correlação entre as variáveis em diferentes números de variáveis (p) e tamanhos de amostra (n) de variáveis produtivas.

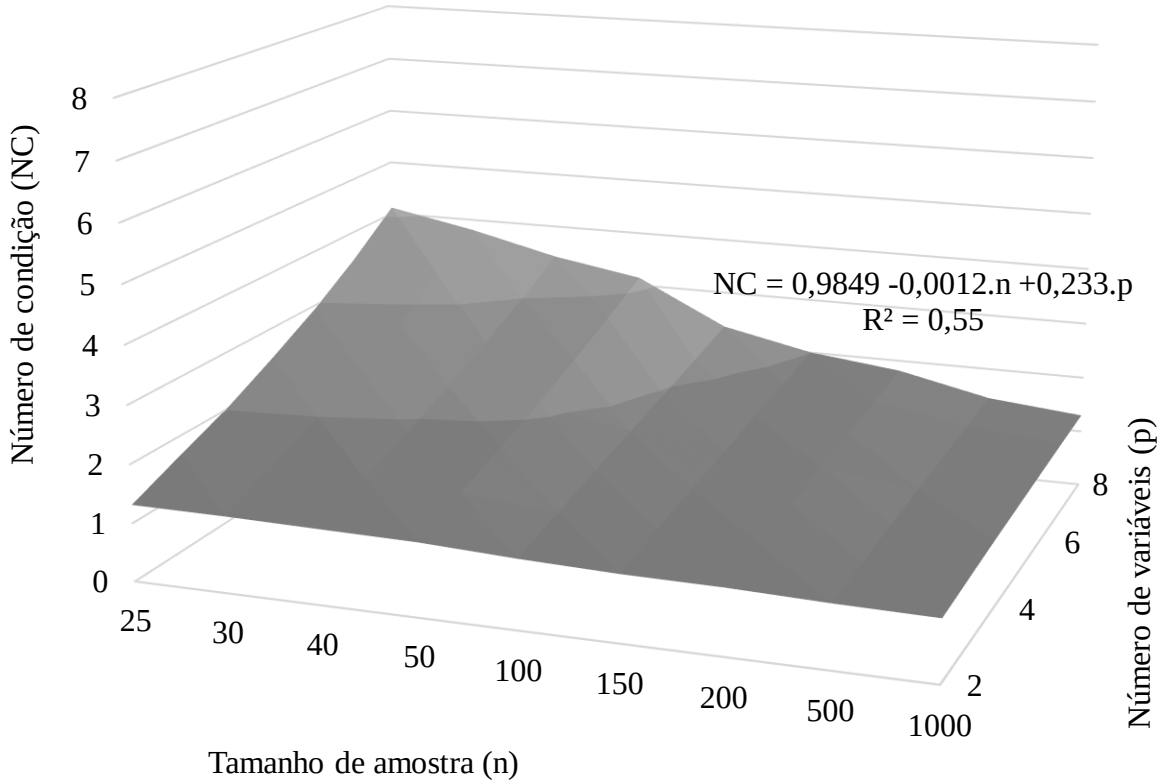


Figura 6. Média do número de condição (NC) para 1000 simulações em cenários com correlação intermediária entre as variáveis (0,4 < r < 0,7), diferentes números de variáveis (p) e tamanhos de amostra (n) de variáveis morfológicas.

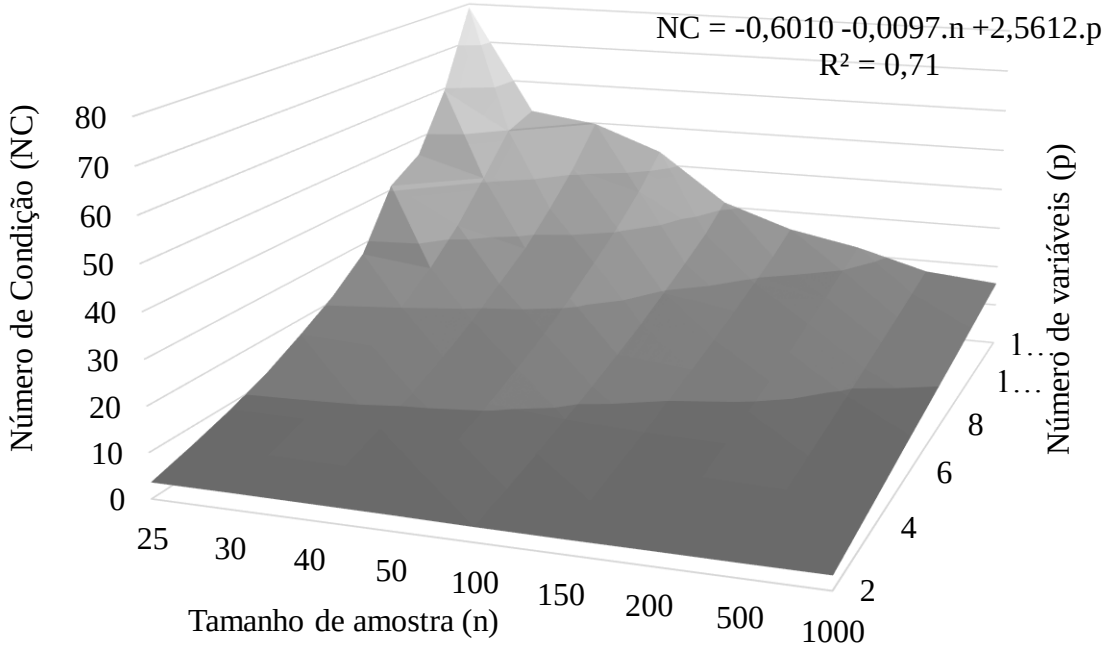
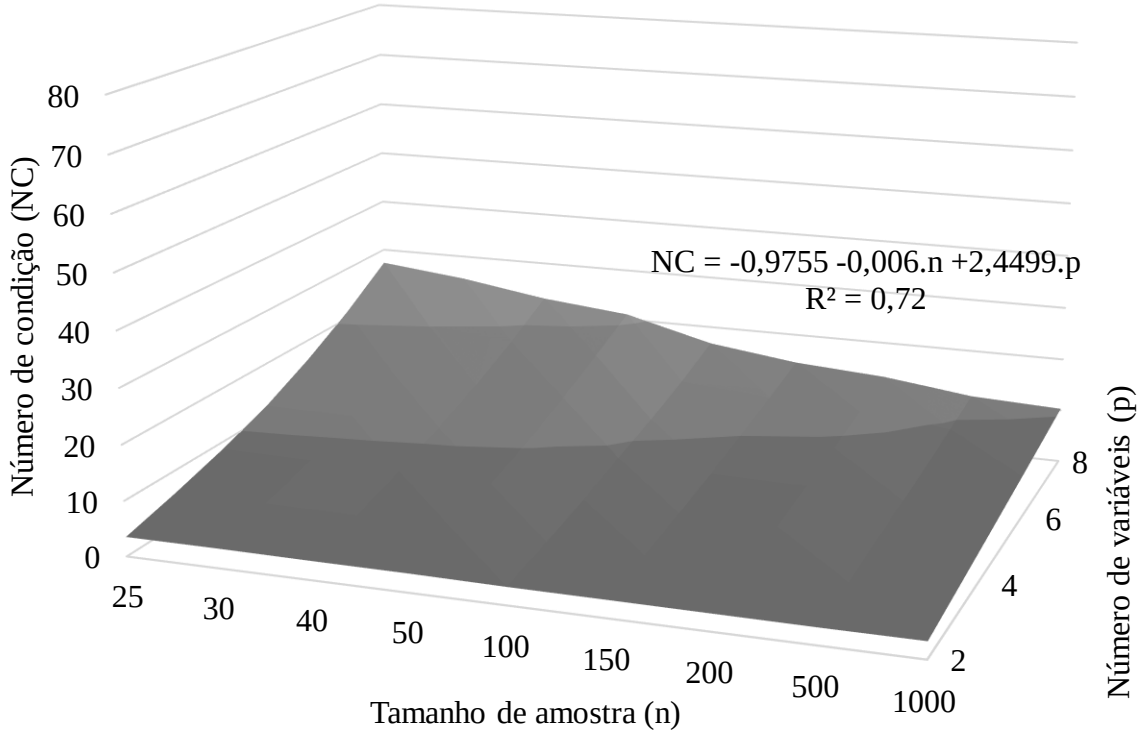


Figura 7. Média do número de condição (NC) para 1000 simulações em cenários de correlação intermediária entre as variáveis (0,4 < r < 0,7) em diferentes números de variáveis (p) e tamanhos de amostra (n) de variáveis produtivas.



Nas figuras 8 e 9 estão representados os valores médios do teste do determinante da matriz de correlação (DET), das 1000 simulações na condição de alta correlação entre as variáveis em cada uma das condições analisadas de número de variáveis e tamanhos de amostra, para as variáveis com baixa variabilidade (variáveis morfológicas) e alta variabilidade (variáveis produtivas), respectivamente. Nota-se um padrão de redução drástica do valor médio do determinante, quando o número de variáveis envolvidas passa de duas para três, sendo que o tamanho de amostra tem pouca interferência sob o valor do determinante.

Figura 8. Média do determinante da matriz de correlação (DET) para 1000 simulações em cenários de alta correlação entre as variáveis em diferentes números de variáveis (p) e tamanhos de amostra (n) de variáveis morfológicas.

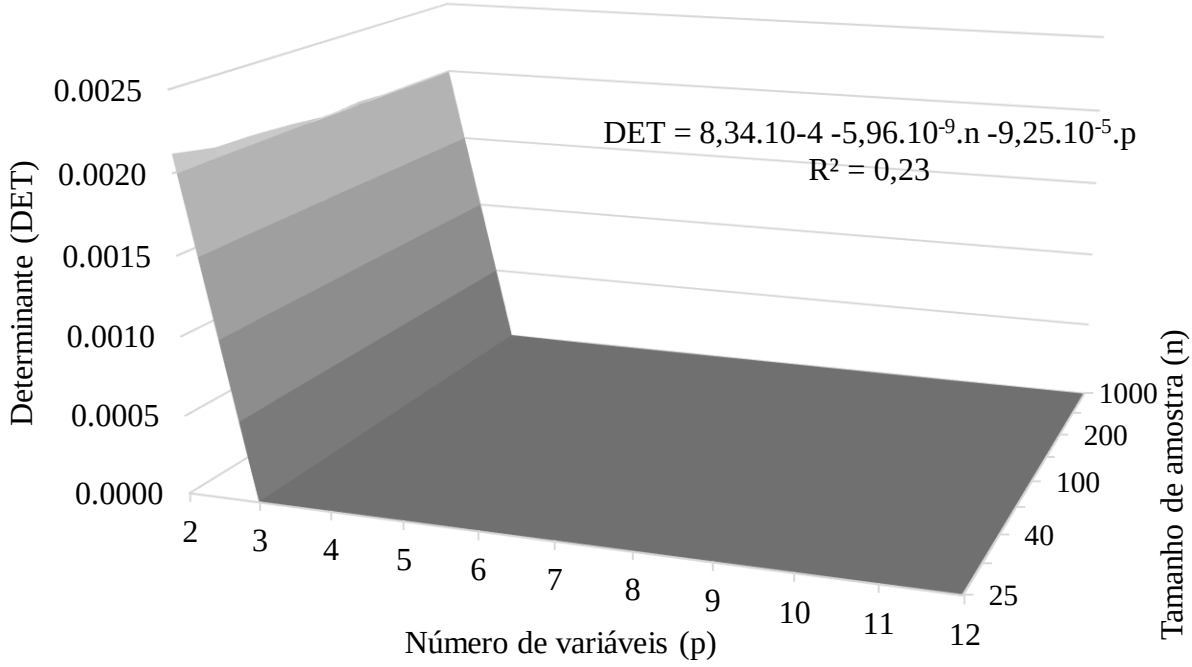
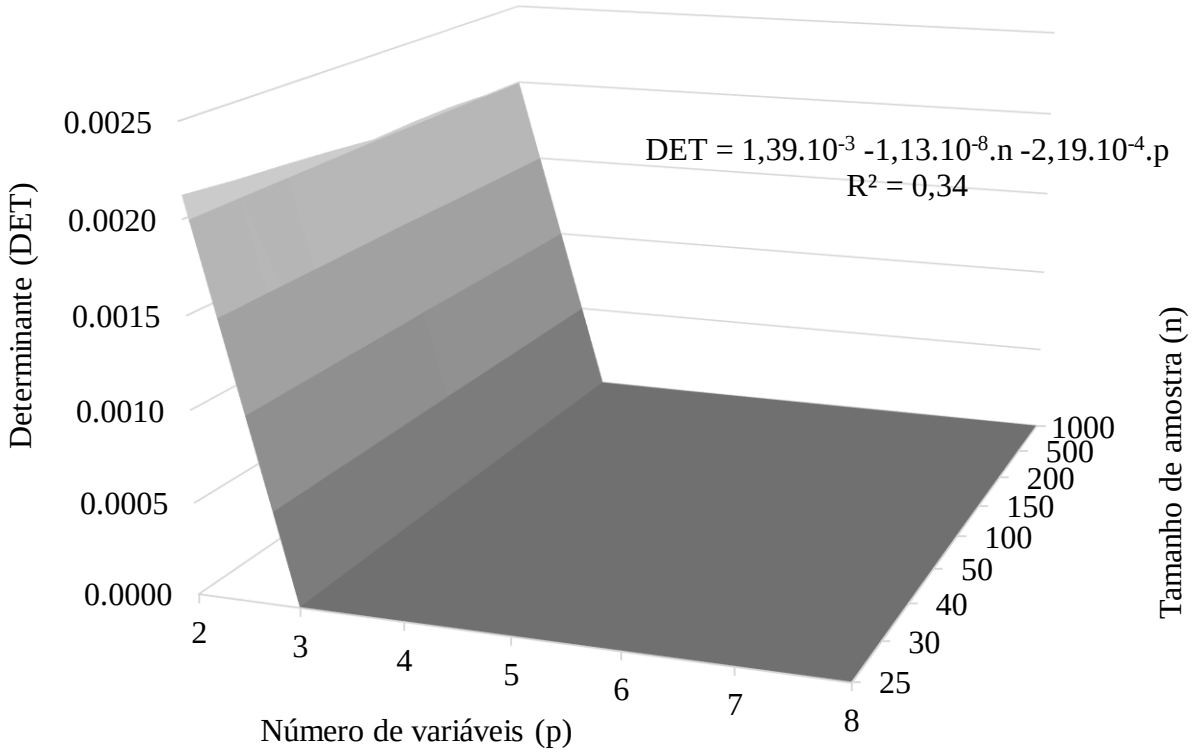


Figura 9. Média do determinante da matriz de correlação (DET) para 1000 simulações em cenários de alta correlação entre as variáveis em diferentes números de variáveis (p) e tamanhos de amostra (n) de variáveis produtivas.



Nas figuras 10 e 11 estão representados os valores médios do teste do determinante da matriz de correlação (DET), das 1000 simulações na condição de baixa correlação entre as variáveis em cada uma das condições analisadas de número de variáveis e tamanhos de amostra, para as variáveis com baixa variabilidade e alta variabilidade, respectivamente. Há um padrão de redução do valor médio do determinante quando o número de variáveis envolvidas aumenta, porém essa redução é gradual e combinada com o tamanho de amostra. O tamanho de amostra tem interferência no valor do determinante sendo que menor tamanho de amostra, combinado com grande número de variáveis, proporciona a redução do determinante da matriz de correlação e maior probabilidade de ocorrência de multicolinearidade nos dados. Esse comportamento corrobora com os resultados observados para o teste do número de condição.

As técnicas do número de condição (NC) e determinante da matriz de correlação (DET) são técnicas que informam o grau de multicolinearidade presente na matriz de correlação X'X de uma forma global, bem como os procedimentos baseados em teste de hipóteses. Já o fator de inflação da variância informa a contribuição de cada variável explicativa na multicolinearidade. Deste modo, é importante utilizar técnicas que informem a situação de multicolinearidade da matriz de correlação de forma ampla, pois assim, é possível verificar se

a multicolinearidade está ou não presenta nas variáveis. Assim, identificando a presença da multicolinearidade torna-se necessário identificar quais variáveis estão altamente relacionadas e provocando a multicolinearidade. Para a identificação das variáveis multicolineares, o teste do fator de inflação de variância torna-se bastante útil, pois este quantifica quanto cada uma das variáveis contribuem para a multicolinearidade. Outra forma de identificação das variáveis que estão altamente relacionadas é através do exame dos autovetores associados ao menor autovalor, assim, a variável associada ao maior autovetor do menor autovalor é a variável com maior relacionamento linear entre as variáveis e que está provocando a multicolinearidade (CRUZ; CARNEIRO, 2006).

Figura 10. Média do determinante da matriz de correlação (DET) para 1000 simulações em cenários de baixa correlação entre as variáveis em diferentes números de variáveis (p) e tamanhos de amostra (n) de variáveis morfológicas.

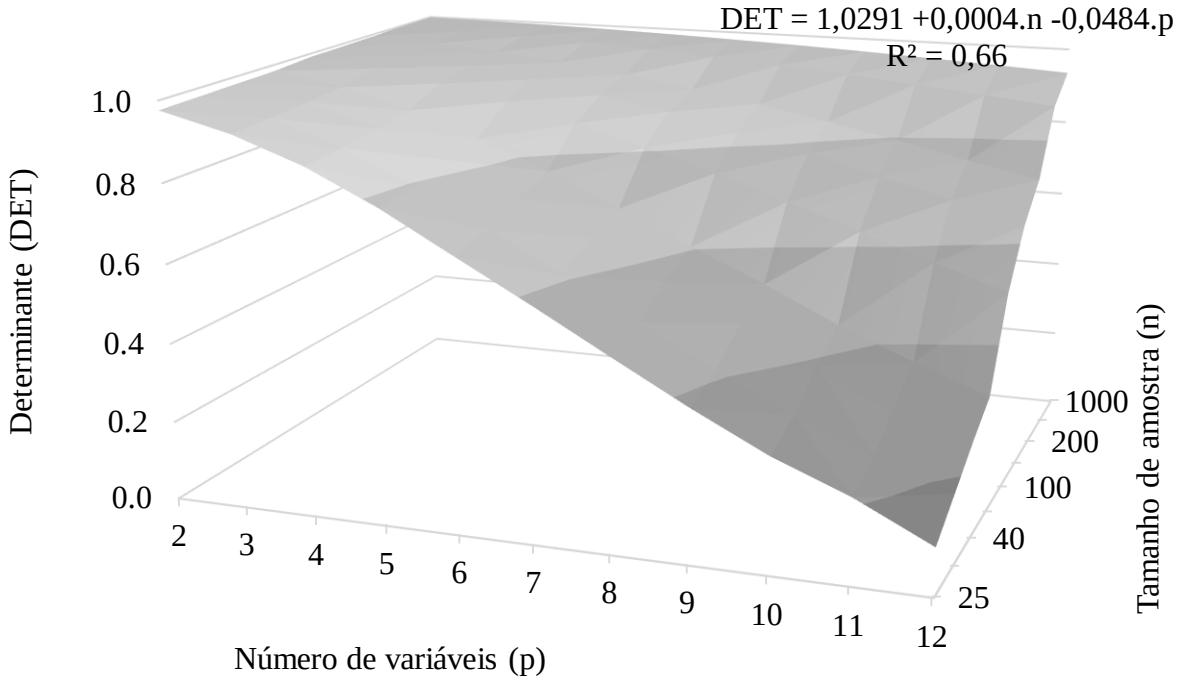


Figura 11. Média do determinante da matriz de correlação (DET) para 1000 simulações em cenários de baixa correlação entre as variáveis em diferentes números de variáveis (p) e tamanhos de amostra (n) de variáveis produtivas.

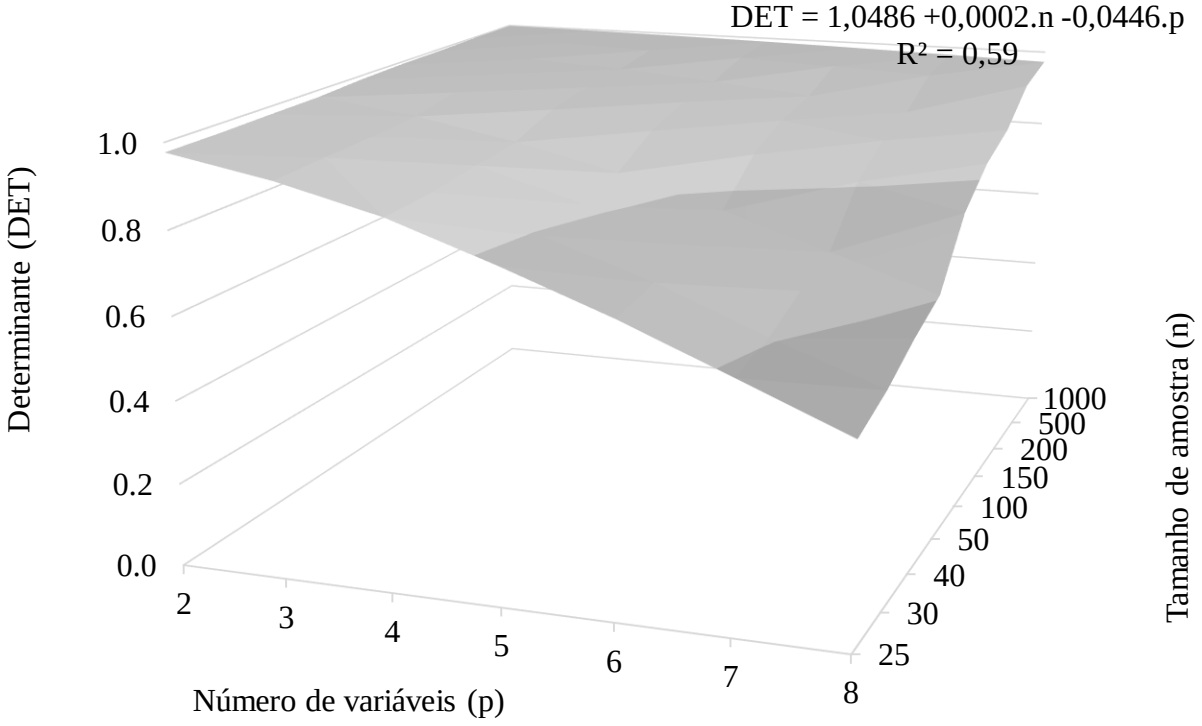


Figura 12. Média do determinante da matriz de correlação (DET) para 1000 simulações em cenários com correlação intermediária entre as variáveis (0,4 < r < 0,7), diferentes números de variáveis (p) e tamanhos de amostra (n) de variáveis morfológicas.

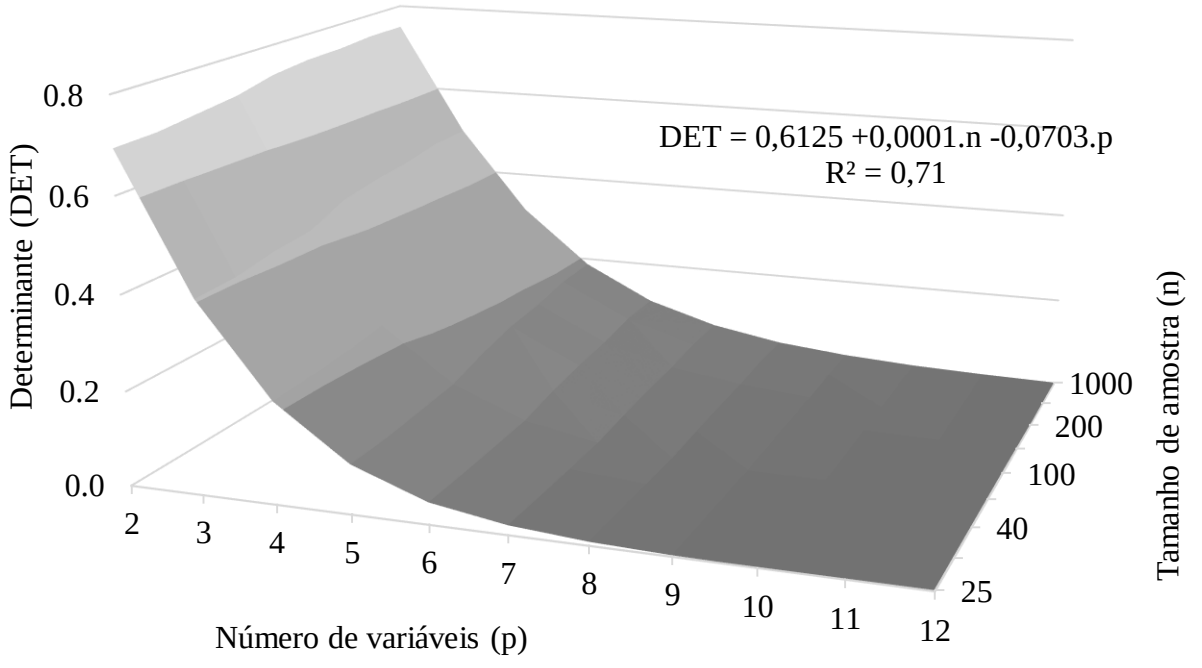
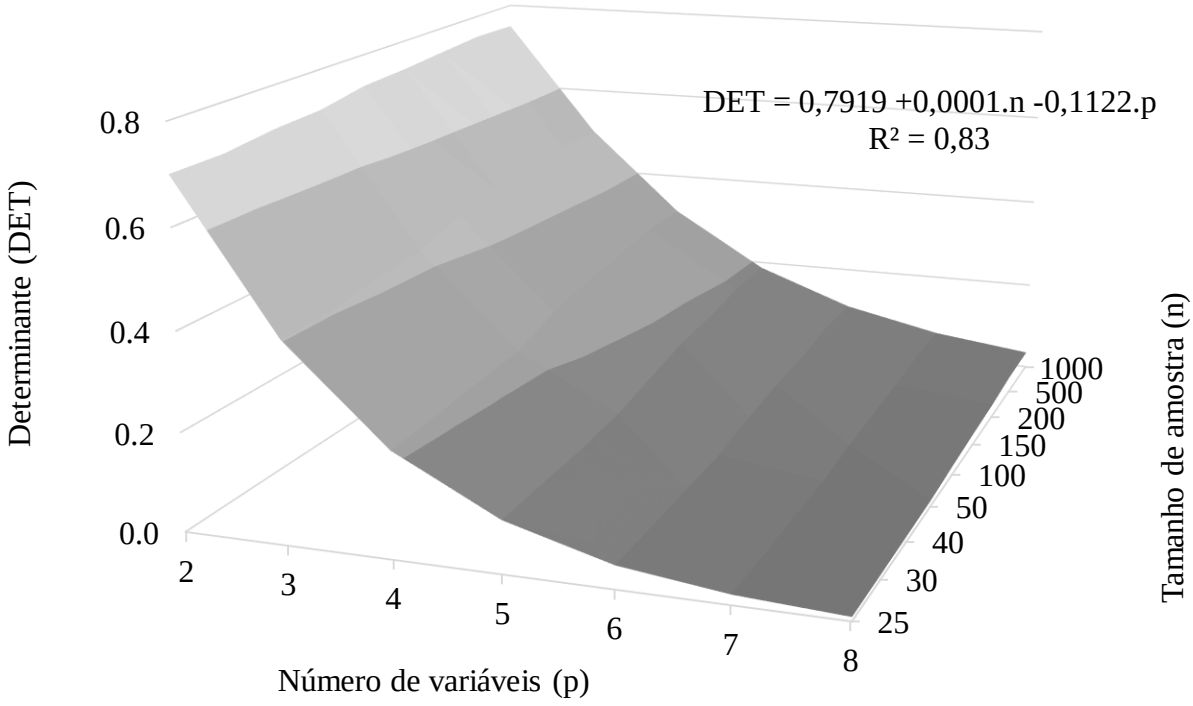


Figura 13. Média do determinante da matriz de correlação (DET) para 1000 simulações em cenários com correlação intermediária entre as variáveis (0,3 < r < 0,7), diferentes números de variáveis (p) e tamanhos de amostra (n) de variáveis produtivas.



4 CONCLUSÃO

Os testes de avaliação da multicolinearidade apresentam resultados diferentes conforme são alterados o número de variáveis, tamanho de amostra e grau de correlação entre as variáveis.

Tamanhos de amostra pequenos e número de variáveis grande aumentam a ocorrência de multicolinearidade.

Os testes do número de condição e fator de inflação de variância são eficientes na identificação de multicolinearidade entre as variáveis.

REFERÊNCIAS BIBLIOGRÁFICAS

ACHEN, C. H. **Interpreting and using regression**. Beverly: Sage Publications, 1982.

ADENOMON, M. O.; OYEJOLA, B. A. A Simulation study of effects of collinearity on forecasting of bivariate time series data. **Scholars Journal of Physics, Mathematics and Statistics**, v. 1, n. 1 p. 4-21, 2014.

ALABI, O. O.; AYINDE, K.; OYEJOLA, B. A. Empirical investigation of effect of multicollinearity on type ii error rates of the ordinary least squares estimators. **Journal of the Nigerian statistical association**, v. 19, p. 50-57, 2007.

ALMEIDA, J. **A agronomia entre a teoria e a ação**. Porto Alegre: UFRGS-LUME, Textos para discussão. 2004. Disponível em: <http://www.ufrgs.br/pgdr/arquivos/423.pdf>. Acesso em: 12 fev 2015.

ANDRIOLO, J. L. **Fisiologia das culturas protegidas**. Santa Maria: UFSM, 1999. 142p.

BARROSO, L. P.; ARTES, R. **Análise multivariada**. Lavras: UFLA, 2003.

BELSLEY, D. A.; KUH, E.; WELSCH, R. E. **Regression diagnostics**: Identifying influential data and sources of collinearity. New York. 1980.

BELSLEY, D. A. Assessing the presence of harmful collinearity and other forms of weak data though a test for signal-to-noise. **Journal of econometrics**. v. 20, p. 211-253, 1982.

BARTLETT, M. S. Properties of sufficiency and statistical tests. **Proceedings of the Royal Statistical Society**, v. 160, p. 268–282, 1937.

BRUNES, R. R. **Relações entre a qualidade fisiológica de sementes de pimentão e a variabilidade na produção de frutos**. Tese doutorado. UFSM, 2013.

CERMEÑO, Z. S. **Estufas – instalações e manejo**. Lisboa: Litexa, 1990. 355p.

CLEMENTS, M. P.; HENDRY, D. F. Macro-economic forecasting and modelling. **The economic journal**, v. 105, n. 431, p. 1001-1013, 1995.

CORRAR, L. J.; PAULO, E.; DIAS FILHO, J. M. **Análise multivariada**: para os cursos de administração, ciências contábeis e economia. São Paulo: Atlas, 2007.

CORTINA, J. M. Interaction, nonlinearity, and multicollinearity: implications for multiple regression. **Journal of management**. v. 19, n. 4. p. 915-922, 1993.

CRUZ, C. D.; CARNEIRO, P. C. S. **Modelos biométricos aplicados ao melhoramento genético**. 2. ed. Viçosa: UFV, 2006.

CRUZ, C. D.; REGAZZI, A. J. Modelos biométricos aplicados ao melhoramento genético. 2. ed. Viçosa: UFV, 1997.

DORMANN, C. F. et al. Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. **Ecography**, v. 36, p. 027-046, 2003.

FARRAR, D. E., GLAUBER, R. R. Multicollinearity in regression analysis: the problem revisited. **Review of economics and statistics**. v. 49, p. 92–107, 1967.

FÁVERO, L.P. et al. **Análise de dados:** modelagem multivariada para tomada de decisões. Rio de Janeiro: Elsevier, 2009.

FERREIRA, D. F. **Estatística multivariada**. 1. ed. Lavras: Ed. da UFLA, 2008.

FIELD, A. **Descobrindo a estatística utilizando o SPSS**. 2. Ed. 2009. 687 p. Disponível em: <https://books.google.com.br/books>. Acesso em: 29 jan. 2015.

FIGUEIREDO FILHO, D. et al. **O que fazer e o que não fazer com a regressão**: pressupostos e aplicações do modelo linear de Mínimos Quadrados Ordinários (MQO). **Revista política hoje**, v. 20, n. 1, 2011.

FILGUEIRA, F. A. R. **Novo Manual de Olericultura: Agrotecnologia moderna na produção de hortaliças**. Viçosa: UFV. 2002. 402p.

FREUND, R. J.; WILSON, W. J. **Regression analysis**: Statistical modeling of a response variable. San Diego: Academic, 1998.

FRISCH, R. **Statistical confluence analysis by means of complete regression systems**. Institute of economics, Oslo University. n. 5, 1934.

GARSON, D. **Statnotes**: Topics in multivariate analysis. 2011. Disponível em: <http://faculty.chass.ncsu.edu/garson/PA765/statnote.htm>. Acesso em: 24 ago. 2015.

GOLDBERGER, A. S. **A course in econometrics.** Harvard University Press, 1991.

GUJARATI, D. N.; PORTER, D. C. **Econometria básica**. 5. Ed. Mcgraw Hill, 2011.

GUNASEKARA, F. I.; CARTER, K.; BLAKELY, T. Glossary for econometrics and epidemiology. **Journal of epidemiology and community health**, v. 62, n. 10, p. 858-861, 2008.

GREWAL, R.; COTE, J. A.; BAUMGARTNER, H. Multicollinearity and measurement error in structural equation models: Implications for theory testing. **Marketing science**, v. 23, n. 4, p. 519-529, 2004.

HAIR, J. F. et al. **Análise multivariada de dados**. 6. ed. Porto Alegre: Bookman, 2009.

HAITOVSKY, Y. Multicollinearity in regression analysis: Comment. **The review of economics and statistics**. v. 51, n. 4, p. 486-489, 1969.

IBGE - Instituto Brasileiro de Geografia e Estatística. **Levantamento sistemático da produção agrícola**, 2016.

JOHNSON, R. A.; WICHERN, D. W. **Applied multivariate statistical analysis**, 6. Ed. Prentice-Hall. 2007.

KENDALL, M. **Multivariate analysis**. High Wycombe, Charles Griffin, 1980.

KENDALL, M. G.; BUCKLAND, W. R. **A dictionary of statistical terms**. New York: Hafner, 1971.

KENNEDY, P. **A Guide to econometrics**. Boston: MIT Press. 2009.

KMENTA, J. **Elements of econometrics**, 2. ed., Macmillan, New York, 1986.

LANGASKENS, Y. **Introduction à l'économétrie**. Librairie Droz, Genéve, Paris. 1975.

LIEBIG, J. V. **Organic Chemistry in its application to agriculture and physiology**. London: Printed for Taylor and Walton, 1840.

LÚCIO, A. D. et al. Excesso de zeros nas variáveis observadas: estudo de caso em experimento com brócolis. **Revista Bragantia**, v. 69, n. 4, p. 1035 – 1046, 2010.

MARQUARDT, D. W. Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. **Technometrics**, v. 12, p. 591–256, 1970.

MASON, C. H.; PERREAULT, W. D. Jr. Collinearity, power, and interpretation of multiple regression. **Journal of marketing research**, v. 28, n. 3, p. 268-280, 1991.

MINGOTI, S. A. **Análise de dados através de métodos de estatística multivariada:** uma abordagem aplicada. Belo Horizonte: Ed. UFMG, 2005.

MONTGOMERY, D. C.; PECK, E. A.; VINIGIN, G. G. **Introduction to linear regression analysis**. John Wiley & Sons, New York, 5. ed. 2012, p. 642.

MOORE, J. S.; REICHERT, A. K.; CHO, C. C. Analysing the temporal stability of appraisal model coefficients: An application of ridge regression techniques. **Journal of the American real estate and urban economics association**, v. 12, n.1, p.50-71, 1984.

NETER, J.; WASSERMAN, W. **Applied linear statistical models**. Illinois: Richard D. Irwin, 1974.

KUTNER, M. H.; NACHTSHEIM, C. J.; NETER, J. Applied linear regression models. 5. ed. McGraw-Hill Higher Education, 2003.

PASQUALI, L. **Análise fatorial para pesquisadores**. Petrópolis: Vozes, 2004.

PEREIRA, G. A.; MILANI, L. L.; CIRILLO, M. A. Uso de alguns estimadores ridge na análise estatística de experimentos em entomologia. **Revista Ceres**, Viçosa, v. 61, n. 3, p. 338-342, 2014.

PEREIRA, G. A. **Estimadores ridge generalizados adaptados em modelos de equações estruturais**: Estudo de simulação e aplicação no perfil de consumidores de café. 2014.

80 p. Tese (doutorado em estatística e experimentação agropecuária)-UFLA, Lavras, 2014.

PIMENTEL, E. C. G. et al. Estimativas de efeitos genéticos em bezerros cruzados por diferentes modelos e técnicas de estimação. **Revista brasileira de zootecnia**, v. 35, p. 1020-1027, 2006.

R CORE TEAM. **R**: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2015. Disponível em: <http://www.R-project.org/>. Acesso em: 26 fev. 2016.

ROYSTON, J. P. Some techniques for assessing multivariate normality based on the Shapiro-Wilk W. **Journal of the royal statistical society**: série C – apllied statistics, London, v. 32, n. 2, p. 121-133, 1983.

ROYSTON, J. P. A toolkit for testing for non-normality in complete and censored samples. **Journal of the royal statistical society**: série D, London, v. 42, n. 1, p. 37-43, 1993.

SCHNEIDER, P. R.; SCHNEIDER, P. S. P. SOUZA, C. A. M. **Análise de regressão aplicada à Engenharia Florestal**. 2. ed. Santa Maria: UFSM, CEPEF, 2009.

SHAPIRO, S. S.; FRANCIA, R. S. An approximate analysis of variance test for normality. **Journal of the American statistical association**, New York, v. 67, n. 337, p. 215-216, Mar. 1972.

SILVA, R. B. V. **Extensão do teste de normalidade de Shapiro-Francia para o caso multivariado**. Tese (Doutorado), UFLA, Lavras, MG, 2009. 59 p.

SOCIEDADE BRASILEIRA DE CIÊNCIA DO SOLO. **Manual de adubação e de calagem para os Estados do Rio Grande do Sul e de Santa Catarina**. Comissão de Química e Fertilidade do Solo. 10. ed., Porto Alegre, 2004.

STEVENS, J. **Applied multivariate statistics for the social sciences**. 3. ed, Mahwah, NJ: Lawrence Erlbaum Associates, 1996.

TABACHNICK, B. G.; FIDELL, L. S. **Using multivariate statistics**, 6. ed., Boston: Pearson, 2013.

TOEBE, M.; CARGNELUTTI FILHO, A. Não normalidade multivariada e multicolinearidade na análise de trilha em milho. **Pesquisa Agropecuária Brasileira**, Brasília, v. 48, n. 5, p. 466-477, 2013.

VASCONCELLOS, M. A. S.; ALVES, D. **Manual de econometria**. São Paulo: Atlas, 2000.

VOLPATO, G. L. **Ciência**: da filosofia à publicação. 6. ed. São Paulo: Cultura Acadêmica, 2013.

ANEXO A –PROGRAMAÇÃO EM LINGUAGEM R.

```
# 1. Rotina para RHO > 0,8 (Alta correlação)
rho1<-0.8 #correlação mínima
rho<-0.9
p<-c(2) #(2,3,4,5,6,7,8,9,10,12) #número de variáveis (k)
n<-c(25,30,40,50,100,150,200,500,1000)# tamanhos de amostras (j)
simula <- 1000  # contador é "i"
said<-array(0,c(simula,13,length(n)))
for (j in 1:length(n)){
        saida <- as.data.frame(matrix(0,simula,13))
        names(saida)<-
c("nobs","pvariaveis,"pvalorFrancia","det","nc","nvif10","fg","fgx2tab","fgpval","hai","haix2
tab","haipval")
        i<-0 #contador das simulações
        while (i <= simula){
            variav<-c(sample(c(1:12),p,rep=F)) # sorteio das variáveis
            y<-dados[,variav]
            mu<-as.vector(apply(y,2,mean)) #Cálculo das médias
            Sigma<-sigma(y,rho) # Matriz Sigma
            x<-Matcor(n[j],mu,Sigma) # Gera n amostras multivariadas
            m<-matrix(cor(x),length(x[1,]),length(x[1,])) # matriz de correlação X'X
            mvfran<-Mult.SF(x)# Teste de normalidade
            menor<-min(m) #Menor correlação
            if (menor > rho1 & as.numeric(mvfran[5] > 0.05)){
                determ<-det(m) # determinante da matriz de correlação = |X'X|
                NC<-nc(m)
                VIF<-vif(m)
                FG<-fg(n[j],m)
                HAI<-hai(n[j],m)
                saida$nobs[i] <- n[j]
                saida$pvariaveis[i] <- dim(x)[2]
                saida$pvalorFrancia[i] <- as.numeric(mvfran[5])
                saida$det[i]<-determ
                saida$nc[i]<-NC[[1]]
```

```
                                saida$nvif10[i]<-VIF[[2]]
                                saida$fg[i]<-FG[[1]]
                                saida$fgx2tab[i]<-FG[[2]]
                                saida$fgpval[i]<-FG[[3]]
                                saida$hai[i]<-HAI[[1]]
                                saida$haix2tab[i]<-HAI[[2]]
                                saida$haipval[i]<-HAI[[3]]
                                i<-i+1
                        }
                }
        said[,,j]<-data.matrix(saida)
}
```

Funções complementares

```
# Teste Número de Condição (Montgomery et al., 2012)
nc<-function(m){ # m = matriz de correlação entre as variáveis explanatórias.
        nc<-max(eigen(m)$values)/min(eigen(m)$values)
        ncfrac<-0
        ncmod<-0
        ncsev<-0
        for (i in 1:length(nc)){
                if (nc[i]<100){
                        ncfrac<-ncfrac+1}
                else if (nc[i]>1000){
                        ncsev<-ncsev+1}
                        else ncmod<-ncmod+1
        }
        return(list(nc,ncfrac,ncmod,ncsev))
}

# Teste  FIV - Fator de inflação da variância (MARQUARDT, 1970)
vif<-function(m){
        A<-diag(eigen(m)$values)
```

```
        A1<-solve(A)
        V<-eigen(m)$vectors
        Vt<-t(V)
        R1<-V%*%A1%*%Vt
        vif<-as.vector(diag(R1))
        nvif<-0
        for (i in 1:length(vif)){
                if (vif[i]>10){
                        nvif<-nvif+1}
        }
        return(list(vif,nvif))
}

# Teste FG - Índice de Farrar e Glauber (FARRAR e GLAUBER, 1967)
fg<-function(n,m){
        fg<-(-((n-1)-
1/6*(2*length(m[1,])+5))*log(det(eigen(m)$vectors%*%diag(eigen(m)$values)%*%t(eigen(
m)$vectors))))
        quitab<-qchisq(.95,(1/2)*length(m[1,])*(length(m[1,])-1))
        pvalorq<-pchisq(fg,(((1/2)*length(m[1,]))*(length(m[1,])-1)),lower.tail=F)
        return(list(fg,quitab,p.value=pvalorq))
}

# Teste de Haitovsky (1969)
hai<-function(n,m){
        hai<-(1+((2*length(m[1,])+5)/6)-n)*log(1-
det(eigen(m)$vectors%*%diag(eigen(m)$values)%*%t(eigen(m)$vectors)))
        quitabhai<-qchisq(.95,(1/2)*length(m[1,])*(length(m[1,])-1))
        pvalorhai<-pchisq(hai,(((1/2)*length(m[1,]))*(length(m[1,])-1)),lower.tail=F)
        return(list(hai,quitabhai,pvalorhai))
}

# Função para gerar a matriz Sigma
sigma<-function(y,rho){
```

```
        sigma<-matrix(0,length(y[1,]),length(y[1,]))
        var<-mean(diag(var(y)))
        cov<-rho*sqrt(var)*sqrt(var)
        for(i in 1:length(y[1,])){
                for(j in 1:length(y[1,])){
                        if (i==j){
                                sigma[i,j]<-var
                        } else sigma[i,j]<-cov
                }
        }
sigma
}

# Função para gerar a matriz com amostras correlacionadas
Matcor<-function(n,mu,Sigma){
        simulacao<-mvrnorm(n,mu,Sigma)
simulacao
}

#Função para identificar a maior correlação
maxrho<-function(m){
        maior<-0
        for (i in 1:p){
        for (j in 1:p){
                if (i != j ){
                        if (m[i,j] > maior){
                                maior<-m[i,j]
                        }
                }
        }
        }
return(maior)
}
```