UNIVERSIDADE ESTADUAL DO OESTE DO PARANÁ *CAMPUS* CASCAVEL

CENTRO DE CIÊNCIAS EXATAS E TECNOLÓGICAS

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA AGRÍCOLA - PGEAGRI

MÉTODO BOOTSTRAP NA AGRICULTURA DE PRECISÃO

GUSTAVO HENRIQUE DALPOSSO

Cascavel – Paraná

Fevereiro – 2017

GUSTAVO HENRIQUE DALPOSSO

MÉTODO BOOTSTRAP NA AGRICULTURA DE PRECISÃO

Tese apresentada ao Programa de Pós-Graduação em Engenharia Agrícola da Universidade Estadual do Oeste do Paraná, em cumprimento parcial aos requisitos para obtenção do título de Doutor em Engenharia Agrícola, área de concentração Sistemas Biológicos e Agroindustriais

Orientador: Prof. Dr. Miguel Angel Uribe Opazo

Cascavel – Paraná

Fevereiro – 2017

Dados Internacionais de Catalogação-na-Publicação (CIP)

D157m

Dalposso, Gustavo Henrique
Método Bootstrap na agricultura de precisão./Gustavo Henrique Dalposso. Cascavel, 2017.
90 f.

Orientador: Prof. Dr. Miguel Angel Uribe Opazo

Revisão português, inglês e normas: Ana Maria Martins Alves Vasconcelos

Tese (Doutorado) – Universidade Estadual do Oeste do Paraná, Campus de Cascavel, 2017
Programa de Pós-Graduação em Engenharia Agrícola

1. Agricultura – Processamento de dados. I. Uribe Opazo, Miguel Angel. II. Vasconcelos, Ana Maria Martins Alves, Rev. III. Universidade Estadual do Oeste do Paraná. IV. Título.

CDD 20.ed. 630.285
CIP-NBR 12899

Ficha catalográfica elaborada por Helena Soterio Bejio – CRB 9ª/965

Revisão de português, inglês e normas realizada por Ana Martins Alves Vasconcelos em 05 de abril de 2017.

GUSTAVO HENRIQUE DALPOSSO

Método Bootstrap na Agricultura de Precisão

Tese apresentada ao Programa de Pós-Graduação em Engenharia Agrícola em cumprimento parcial aos requisitos para obtenção do título de Doutor em Engenharia Agrícola, área de concentração Sistemas Biológicos e Agroindustriais, linha de pesquisa Geoprocessamento, Estatística Espacial e Agricultura de Precisão, APROVADO(A) pela seguinte banca examinadora:

Orientador(a) - Miguel Angel Uribe Opazo

Universidade Estadual do Oeste do Paraná - Campus de Cascavel (UNIOESTE)

Jerry Adriani Johann

Universidade Estadual do Oeste do Paraná - Campus de Cascavel (UNIOESTE)

Luciana Pagliosa Carvalho Guedes

Universidade Estadual do Oeste do Paraná - Campus de Cascavel (UNIOESTE)

Diogo Francisco Rossoni

Universidade Estadual de Maringá (UEM)

Fernanda De Bastiani

Universidade Federal de Pernambuco (UFPE)

Cascavel, 15 de fevereiro de 2017

BIOGRAFIA

Nome: Gustavo Henrique Dalposso

Ano de nascimento: 1979

Naturalidade: Cascavel – PR

Licenciado em Matemática pela Universidade Estadual do Oeste do Paraná (UNIOESTE) no ano de 2007.

Mestre em Engenharia Agrícola pela Universidade Estadual do Oeste do Paraná (UNIOESTE) no ano de 2010.

Professor da Universidade Tecnológica Federal do Paraná (UTFPR) Campus Toledo, lotado na Coordenação do Curso Superior de Licenciatura em Matemática (COMAT) de 2012 até a presente data.

“Considero feliz aquele que quando se fala de êxito busca a resposta em seu trabalho”.

Ralph Waldo Emerson

Ao meu avô, Nadir André Picolli (*in memorian*).

Agradecimentos

À Universidade Tecnológica Federal do Paraná (UTFPR) - Campus Toledo, pela oportunidade oferecida em fazer o doutorado com afastamento integral.

À Universidade Estadual do Oeste do Paraná (UNIOESTE) – Campus Cascavel, pela oportunidade oferecida em fazer o doutorado.

Ao Programa de Pós-Graduação em Engenharia Agrícola (PGEAGRI) da UNIOESTE, pela oportunidade oferecida na realização deste trabalho.

À Fundação Araucária, pelo apoio financeiro.

Aos professores do Programa de Pós-Graduação em Engenharia Agrícola da UNIOESTE, por contribuírem para minha formação.

Ao meu orientador, Professor Dr. Miguel Angel Uribe Opazo, pela orientação ao longo do desenvolvimento deste trabalho.

Ao professor Dr. Jerry Adriani Johann, pelas contribuições ao trabalho e por sua disponibilidade.

Ao professor Dr. Robert A. LaBudde, pelos ensinamentos oferecidos durante o curso online – Métodos Bootstrap – do Institute for Statistics Education.

Aos colegas do LEE – Laboratório de Estatística Espacial.

Aos colegas do LEA – Laboratório de Estatística Aplicada.

À minha esposa, Solange, eterna companheira e apoiadora e aos meus amados filhos, Henrique e Guilherme.

A todas as pessoas que de alguma forma contribuíram para a realização deste trabalho.

Muito Obrigado.

RESUMO

MÉTODO BOOTSTRAP NA AGRICULTURA DE PRECISÃO

Um problema que ocorre nos estudos vinculados à agricultura de precisão diz respeito aos métodos estatísticos utilizados nas análises inferenciais, pois eles requerem pressupostos que muitas vezes não podem ser assumidos. Uma alternativa aos métodos tradicionais é a utilização do método bootstrap, que utiliza reamostragens com reposição do conjunto de dados originais para realizar inferências. A metodologia bootstrap pode ser aplicada a dados amostrais independentes e também em casos de dependência, como na estatística espacial. No entanto, para se utilizar o método bootstrap em dados espaciais, são necessárias adaptações no processo de reamostragem. Este trabalho teve como objetivo utilizar o método bootstrap em estudos vinculados à agricultura de precisão, cujo resultado é a elaboração de três artigos. No primeiro artigo utilizou-se um conjunto de dados de produtividade de soja e atributos do solo formado com poucas amostras para determinar um modelo de regressão linear múltipla. Foram utilizados métodos bootstrap para a seleção de variáveis, identificação de pontos influentes e determinação de intervalos de confiança dos parâmetros do modelo. Os resultados mostraram que os métodos bootstrap permitiram selecionar os atributos que foram significativos na construção do modelo, construir os intervalos de confiança dos parâmetros e identificar os pontos que tiveram grande influência sobre os parâmetros estimados. No segundo artigo estudou-se a dependência espacial de dados de produtividade de soja e atributos do solo utilizando o método bootstrap na análise geoestatística. Utilizou-se o método bootstrap espacial para quantificar as incertezas associadas à caracterização das estruturas de dependência espacial, aos estimadores dos parâmetros dos modelos ajustados, aos valores preditos por krigagem e ao pressuposto de normalidade multivariada dos dados. Os resultados obtidos possibilitaram quantificar as incertezas em todas as fases da análise geoestatística. No terceiro artigo utilizou-se uma regressão espacial linear para modelar a produtividade de soja em função de atributos do solo. Foram utilizados métodos bootstrap espaciais para determinar estimadores pontuais e por intervalo associados aos parâmetros do modelo. Realizaram-se testes de hipóteses sobre os parâmetros do modelo e foram eleborados gráficos de probabilidade para identificar a normalidade dos dados. Os métodos permitiram quantificar as incertezas associadas à estrutura de dependência espacial, avaliar a significância individual dos parâmetros associados à média do modelo espacial linear e verificar a suposição de normalidade multivariada dos dados. Conclui-se, portanto, que o método bootstrap é uma eficaz alternativa para realizar inferências em estudos vinculados à agricultura de precisão.

Palavras-chave: geoestatística; inferência estatística; produtividade de soja; reamostragem.

ABSTRACT

BOOTSTRAP METHOD IN PRECISION FARMING

One issue in precision agriculture studies concerns about the statistical methods applied in inferential analysis, since they have required assumptions that, sometimes, cannot be assumed. A possibility to traditional methods is to use the bootstrap method, which consists in resampling and replacing the original data set to carry out inferences. The bootstrap methodology can be applied to independent sample data as well as in cases of dependence, such as in spatial statistics. However, adjustments are required during the resampling process in order to use the bootstrap method in spatial data. Thus, this trial aimed at applying the bootstrap method in precision agriculture studies, whose result was the preparation of three scientific papers. Soybean yield and soil attributes datasets formed with few samples were used in the first paper to determine a multiple linear regression model. Bootstrap methods were chosen to select variables, identify influential points and determine confidence intervals of the model parameters. The results showed that the bootstrap methods allowed selecting significant attributes to design a model, to build confidence intervals of the studied parameters and finally to indentify the influential points on the estimated parameters. Besides, spatial dependence of soybean yield data and soil attributes were studied in the second paper by bootstrap method in geostatistical analysis. The spatial bootstrap method was used to quantify the uncertainties associated with the spatial dependence structure, the fitted model parameter estimators, kriging predicted values and multivariate normality assumption of data. Thus, it was possible to quantify the uncertainties in all phases of geostatistical analysis. A spatial linear model was used to analyze soybean yield considering the soil attributes in the third paper. Spatial bootstrap methods were used to determine point and interval estimators associated with the studied model parameters. Hypothesis tests were carried out on the model parameters and probability plots were developed to identify data normality. These methods allowed to quantify the uncertainties associated to the structure of spatial dependence, as well as to evaluate the individual significance of the parameters associated with the average of the spatial linear model and to verify data multivariate normality assumption. Finally, it is concluded that bootstrap method is an effective alternative to make statistical inferences in precision agriculture studies.

Keywords: geostatistic; statistical inference; soybean yield; resampling.

SUMÁRIO

LISTA DE FIGURAS

Artigo 1

Artigo 2

Artigo 3

LISTA DE TABELAS

Artigo 1

Artigo 2

Artigo 3

1 INTRODUÇÃO/JUSTIFICATIVA

A expressão bootstrap está relacionada ao texto "*pulling oneself up by one´s bootstrap*", uma frase usada pela primeira vez no livro "As viagens singulares, campanhas e aventuras do barão de Munchausen", de Rudolph Erich Raspe em 1786 (CHERNICK e LABUDDE, 2011). O termo faz alusão às histórias de que o Barão de Munchausen era capaz de se erguer do pântano puxando as alças das próprias botas o que exemplifica a ação de sair de uma situação difícil a pirtir dos próprios esforços. Em estatística, bootstrap refere-se a fazer inferências a cerca de parâmetros desconhecidos utilizando reamostragens com reposição do conjunto amostral. Cada reamostragem permite calcular uma nova estatística e o conjunto de todas estas estimativas permite elaborar uma distribuição empírica, a qual é utilizada na inferência estatística.

O bootstrap é uma metodologia pertencente a uma classe de métodos, conhecidos como procedimentos de reamostragem. Alguns procedimentos de reamostragem semelhantes ao bootstrap foram introduzidos há algum tempo na literatura. Por exemplo, os métodos de permutação apresentados por Fisher (1935) e o método jackknife apresentado por Quenouille (1949). A utilização de computadores para realizar simulações também remonta ao passado com o surgimento dos computadores no início dos anos 1940 (CHERNICK, 2008). No entanto, um ano especial para o bootstrap foi 1979, devido a um artigo de Bradley Efron publicado no periódico *Annals of Statistics* (EFRON,1979), em que o autor tinha como objetivo explicar o método jackknife em termos de um método mais primitivo, chamado de bootstrap, argumentando que este é mais amplamente aplicável e confiável que o método jackknife.

Como a metodologia bootstrap envolve n amostragens com reposição para uma amostra de tamanho n, existem n^n possíveis amostras bootstrap. Assim, a enumeração completa de todas as amostras bootstrap torna-se inviável exceto em amostras muito pequenas. Desta forma, a amostragem aleatória de um conjunto de possíveis amostras bootstrap torna-se um caminho viável para aproximar a distribuição das amostras bootstrap (CHERNICK e LABUDDE, 2011). Neste sentido, bootstrap refere-se a um caso de simulação de Monte Carlo que trata a amostra original como a pseudo-população ou estimativa da população. A consideração de um método de Monte Carlo foi certamente uma ideia importante expressa por Efron, porém esta ideia já havia sido aplicada. Devido ao estudo apresentado em Simon (1969), o autor reivindicou autoria ao bootstrap, tendo em vista sua recomendação de uma abordagem de Monte Carlo como forma de ensinar probabilidade e estatística de forma mais intuitiva. Porém, Efron (1979) foi o primeiro a apresentar um concorrente aos métodos jackknife e delta (CRAMER, 1946) para determinar uma estimativa do erro padrão do estimador.

Em seus trabalhos seguintes, Efron evidenciou ampla aplicabilidade do bootstrap, utilizando-o em intervalos de confiança, testes de hipótese e demais problemas complexos (EFRON e GONG, 1983; DIACONIS e EFRON, 1983; EFRON e TIBSHIRANI, 1986; EFRON, 1982). Inicialmente, houve uma grande dose de dúvida em relação à metodologia bootstrap, porém a comunidade científica começou a tomar conhecimento do método e reconhecer sua importância, o que fez crescer exponencialmente a quantidade de pesquisas relacionadas ao bootstrap.

Existe uma atração em se aplicar a metodologia bootstrap em uma ampla variedade de situações. Porém, em casos em que as amostras são dependentes, como ocorre por exemplo em séries temporais e dados espaciais, a aplicação direta da metodologia pode reproduzir uma estimativa incorreta do erro padrão, em virtude da quebra da estrutura de autocorrelação dos dados (PLANT, 2012). Para se utilizar eficazmente o bootstrap para estimar o erro padrão em dados correlacionados, a reamostragem deve ser adaptada para gerar uma nova amostra que preserve a estrutura de autocorrelação dos dados. Existem diversas abordagens que podem ser utilizadas: o bootstrap em blocos (HALL, 1985), o bootstrap espacial (SOLOW, 1985), o bootstrap espacial paramétrico (TANG et al., 2006) entre outros.

Na agricultura de precisão, alguns dos objetivos pretendidos referem-se à redução dos custos de produção, ao aumento das produtividades e à redução dos impactos ambientais decorrentes das atividades agrícolas. Para alcançar estes objetivos, constantemente são desenvolvidas pesquisas que buscam compreender as relações existentes entre as diversas variáveis associadas à agricultura e à produtividade das lavouras com a elaboração de mapas temáticos que possibilitem um manejo localizado da lavoura.

Neste sentido, justifica-se a importância da utilização de métodos bootstrap nos estudos relacionados à agricultura de precisão, pois muitas vezes a obtenção de dados agrícolas é uma tarefa laboriosa e onerosa. Como consequência, os conjuntos de dados são montados com poucas amostras, por conseguinte inviabilizam a realização de inferências tendo em vista que os métodos tradicionais requerem uma elevada quantidade de amostras. Outra questão que justifica a utilização do bootstrap diz respeito aos modelos espaciais utilizados na elaboração de mapas temáticos. Como estes modelos são elaborados com uma única amostra, é prudente quantificar as incertezas associadas aos resultados obtidos, procedimento este que pode ser realizado com a adoção de métodos bootstrap que levem em consideração a estrutura espacial das variáveis analisadas.

Desta forma, a presente tese foi elaborada com a seguinte estrutura: a primeira parte aborda a introdução, justificativa e objetivos; na segunda parte apresenta-se uma revisão da literatura utilizada para o desenvolvimento do trabalho. Os resultados são apresentados na forma de três artigos científicos resultantes do projeto de pesquisa desenvolvido. Cada

artigo pode ser lido de forma independente, os quais têm em comum a utilização de métodos de reamostragem bootstrap. O Artigo 1 trata da utilização de métodos bootstrap na modelagem da produtividade de soja utilizando amostras pequenas. O Artigo 2 utiliza o método bootstrap espacial para quantificar incertezas no estudo da dependência espacial de dados de produtividade de soja e atributos do solo. O Artigo 3 apresenta um estudo dos métodos bootstrap espacial e bootstrap espacial paramétrico em uma análise geoestatística da produtividade de soja consideranto atributos do solo como covariáveis. Por último, apresenta-se um capítulo com considerações finais. Destaca-se que o Artigo 1 foi publicado na revista Spanish Journal of Agricultural Research (SJAR), e sua versão *online* pode ser acessada em (DOI: 10.5424/sjar/2016143-8635).

2 OBJETIVOS

2.1 Objetivo geral

O objetivo geral deste trabalho foi utilizar métodos bootstrap em estudos vinculados à agricultura de precisão.

2.1 Objetivos específicos

Especificamente, os objetivos deste trabalho foram:

- Utilizar métodos bootstrap para seleção de variáveis explicativas, identificação de pontos influentes e determinação de intervalos de confiança na modelagem da produtividade de soja considerando atributos químicos e físicos do solo como variáveis explicativas.

- Utilizar o método bootstrap espacial para quantificar as incertezas associadas à estrutura de dependência espacial, aos parâmetros dos modelo, aos valores krigados e ao pressuposto de normalidade multivariada na análise geoestatística da produtividade de soja e dos atributos do solo.

- Utilizar os métodos bootstrap espacial e bootstrap espacial paramétrico para quantificar as incertezas associadas aos parâmetros de um modelo espacial linear Gaussiano da produtividade de soja tendo atributos do solo como covariáveis.

3 REVISÃO BIBLIOGRÁFICA

3.1 O método de reamostragem bootstrap

Seja θ um parâmetro de uma função de distribuição de probabilidade F com base em uma amostra aleatória $\mathbf{x} = (x_1, \ldots, x_n)^T$ obtida de F. Um parâmetro θ é uma função da distribuição de probabilidade F, ou seja, $\theta = s(F)$. Os dados amostrais permitem calcular a estatística $\hat{\theta} = s(\hat{F})$, sendo $\hat{F}$ uma função de distribuição empírica de probabilidades. O procedimento de estimar um parâmetro de F utilizando a correspondente estatística de $\hat{F}$ é conhecido como princípio “plug-in” (EFRON e TIBSHIRANI, 1993). O bootstrap utiliza o princípio plug-in para calcular o erro padrão estimado de uma estatística $\hat{\theta}$.

O método bootstrap depende de $\mathbf{x}^* = (x_1{}^*, \ldots, x_n{}^*)^T$, uma amostra aleatória de tamanho n extraída com reposição de $\mathbf{x}$, conhecida como amostra bootstrap. Correspondendo a uma amostra bootstrap $\mathbf{x}^*$, tem-se a réplica bootstrap de $\hat{\theta}$, denotada por $\hat{\theta}^* = s(\mathbf{x}^*)$. O algoritmo bootstrap funciona quando se obtêm B amostras bootstrap independentes $\mathbf{x}^{*1}, \ldots, \mathbf{x}^{*B}$ e calculam-se as réplicas bootstrap correspondentes, $\hat{\theta}^*(b) = s(\mathbf{x}^{*b}),\ b = 1, \ldots, B$. As réplicas bootstrap calculadas formam uma distribuição empírica utilizada nas inferências estatísticas. Referências sobre o desenvolvimento histórico do bootstrap podem ser encontradas em Chernick (2008) e Chernick e LaBudde (2010).

3.2 Intervalos de confiança bootstrap

Um intervalo de confiança é estimado de um parâmetro de interesse de uma população. Logo, ao invés de se estimar o parâmetro por um único valor, é dado um intervalo de estimativas prováveis. O quanto estas estimativas são prováveis será determinado pelo coeficiente de confiança $(1 - \alpha)$ para $\alpha \in (0,1)$. Intervalos de confiança são usados para indicar a confiabilidade de uma estimativa. Por exemplo, um intervalo de confiança pode ser usado para descrever o quanto os resultados de uma pesquisa são confiáveis. Sendo todas as estimativas iguais, uma pesquisa que resulte em um intervalo de confiança pequeno é mais confiável do que uma que resulte em um intervalo de confiança maior. Embora existam metodologias que permitam elaborar intervalos de confiança, geralmente elas são dependentes de pressupostos sobre a distribuição do estimador. Neste sentido, o método bootstrap constitui uma alternativa eficiente para a teoria usual. Pois esse, além de ser livre de complexidades algébricas, possibilita a obtenção de intervalos de confiança sem a necessidade de pressupostos sobre a distribuição do estimador (CHERNICK e LABUDDE, 2010).

3.2.1 Intervalo de confiança percentil

Segundo Efron (1982), a construção dos intervalos de confiança percentil ou percentil de Efron é a maneira mais simples de construir um intervalo de confiança para um parâmetro com base em estimativas bootstrap. Suponha que $\hat{\theta}_i^*$ seja a i-ésima estimativa bootstrap com base na i-ésima amostra bootstrap. Ao serem ordenadas crescentemente as estimativas $\hat{\theta}_i^*$, $i = 1, \ldots, B$, é de se esperar que um intervalo que contenha $(1-\alpha)\%$ das estimativas seja um intervalo de confiança de $(1-\alpha)\%$ para θ. A maneira mais adequada para escolher este intervalo de confiança é escolher $(1-\alpha)\%$ das estimativas centrais, excluindo $(\alpha/2)\%$ das menores estimativas e $(\alpha/2)\%$ das maiores estimativas (CHERNICK, 2008).

3.2.2 Intervalo de confiança bootstrap normal

O método da aproximação normal é útil quando é possível assumir que a estatística de interesse é normalmente distribuída, mas não há um método analítico para a estimativa do desvio padrão da distribuição amostral. O intervalo de confiança é calculado pela Equação (1) e de acordo com Wasserman (2006), ele não é preciso, a menos que a distribuição de $\hat{\theta}$ seja próxima da distribuição normal.

$$\left(\hat{\theta}_l, \hat{\theta}_u\right) = \left(\hat{\theta} - z_{\alpha/2} \cdot \widehat{se}_B\,,\ \hat{\theta} + z_{\alpha/2} \cdot \widehat{se}_B\right) \tag{1}$$

em que, $\hat{\theta}_l$ é o limite inferior do intervalo, $\hat{\theta}_u$ é o limite superior do intervalo, $z_{\alpha/2}$ é o valor positivo do escore z que está na fronteira vertical de uma área de $\alpha/2$ na cauda direita da distribuição normal padrão e $\widehat{se}_B$ é a estimativa do erro padrão de $\hat{\theta}$.

3.2.3 Intervalo de confiança bootstrap BC

Uma das dificuldades encontradas na utilização do método bootstrap percentil deve-se ao fato deste assumir um não enviesamento da distribuição de $\hat{\theta}^*$, isto é, que $\hat{\theta}^*$ seja um estimador não viesado de $\hat{\theta}$ e $\hat{\theta}$ um estimador não viesado de θ (MOONEY e DUVAL, 1993). Efron e Tibshirani (1986) resolvem este problema fazendo uma correção de viés para o método percentil, chamando este novo método de intervalo de confiança com viés corrigido. O método BC (*bias corrected*) ajusta a distribuição bootstrap de $\hat{\theta}$ utilizando um valor chamado de constante do viés corrigido, apresentado na Equação (2):

$$\hat{z}_0 = \Phi^{-1}\left[\#\left\{\hat{\theta}^*(b) < \hat{\theta}\right\}/B\right] \tag{2}$$

em que, $\Phi^{-1}(\cdot)$ indica a inversa da função de distribuição normal acumulada e # indica o número de réplicas bootstrap inferiores a $\hat{\theta}$. Shasha e Wilson (2011) apresentam um roteiro para obter o intervalo de confiança com viés corrigido, cujos limites são representados pelos percentis da distribuição bootstrap, indicados nas Equações (3) e (4):

$$\hat{\theta}_l = \left[0{,}5 + \Phi\left(Z_{\alpha/2} + 2\hat{z}_0\right)\right]^o \quad (3)$$

$$\hat{\theta}_u = \left[0{,}5 + \Phi\left(Z_{1-\alpha/2} + 2\hat{z}_0\right)\right]^o \quad (4)$$

em que, $Z_{\alpha/2}$ e $Z_{1-\alpha/2}$ são os escores da distribuição normal padronizada e $\Phi(\cdot)$ é a função de distribuição normal reduzida.

3.2.4 Outros intervalos de confiança bootstrap

Embora os intervalos de confiança percentil de Efron e bootstrap normal sejam os mais populares, existem diversos métodos bootstrap destinados à obtenção de intervalos de confiança. O intervalo de confiança BCa (*bias-corrected and accelerated*) (EFRON e TIBSHIRANI, 1986) incorpora um parâmetro que Efron chama de “constante de aceleração”, que melhora o desempenho do intervalo de confiança percentil. Detalhes sobre este intervalo de confiança podem ser vistos em Davison e Hinkley (1997), em que os autores chamam o intervalo BCa de “método percentil ajustado”. Outros métodos são os intervalos de confiança *bootstrap-t, bootstrap* paramétrico*, smoothed bootstrap, tilted bootstrap* e *iterated bootstrap*, que podem ser vistos no livro Good (2006).

3.3 O método bootstrap em regressão

Quando se utiliza a notação matricial, o modelo estatístico de uma regressão linear múltipla com k variáveis independentes fica representado pela Equação (5):

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad (5)$$

em que $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ é o vetor $n \times 1$ contendo os valores da variável dependente, $\mathbf{X} = [\mathbf{1}\ \mathbf{X}_1 \ldots \mathbf{X}_k]$ é a matriz $n \times (k+1)$ das observações das k variáveis explicativas e $\mathbf{1}$ é o vetor $n \times 1$ de uns, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{k+1})^T$ é o vetor $(k+1) \times 1$ dos parâmetros desconhecidos a ser estimado e $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$ é o vetor $n \times 1$ dos erros aleatórios.

Assumem-se os seguintes presupostos ao se estabelecer o modelo de regressão linear múltipla: I) a variável dependente é função linear das variáveis independentes; II) os valores das variáveis independentes são fixos; III) $\mathrm{E}(\varepsilon) = 0$; IV) homocedasticidade dos erros; V) os erros são independentes e VI) os erros têm distribuição normal. As pressuposições I, II e III são necessárias para demonstrar que os estimadores de mínimos quadrados são não tendenciosos e as cinco primeiras pressuposições permitem demonstrar que tais estimadores são lineares não tendenciosos de variância mínima. A pressuposição VI é necessária para realizar testes de hipóteses e para construir intervalos de confiança para os parâmetros. Escolhido o nível de confiança, o intervalo de confiança para β_i é determinado por:

$$\hat{\beta}_i - t_0 s(\hat{\beta}_i) < \beta_i < \hat{\beta}_i + t_0 s(\hat{\beta}_i) \quad (6)$$

em que, $\hat{\beta}_i$ é a estimativa obtida, t_0 é valor crítico da distribuição t de Student e $s(\hat{\beta}_i)$ é a estimativa do desvio padrão de $\hat{\beta}_i$ (HOFFMANN, 2006).

Quando as premissas da modelagem são violadas, as estimativas podem ficar viesadas por não serem robustas. Se a distribuição dos erros é de cauda pesada ou existem *outliers* nos dados, as estimativas de mínimos quadrados atribuem muito peso para estes valores atípicos, por conseguinte há distorção das estimativas. Quando a distribuição dos erros é desconhecida e não normal, o método bootstrap possibilita obter tais estimativas, independente do método utilizado para estimar os parâmetros (CHERNICK e LABUDDE, 2011).

3.3.1 Bootstrap dos resíduos

Há duas abordagens básicas para utilizar bootstrap em regressão e ambas podem ser aplicadas tanto para regressão linear quanto para regressão não linear. A primeira abordagem é conhecida como bootstrap dos resíduos, e o algoritmo utilizado para gerar os conjuntos de dados simulados e as correspondentes estimativas dos parâmetros é o seguinte:

Algoritmo 1: Bootstrap dos resíduos (SHERMAN, 2010).

a) Após determinado $\widehat{\boldsymbol{\beta}}$, calcule os resíduos $\boldsymbol{\varepsilon}_i = \mathbf{y}_i - \mathbf{x}_i^T \widehat{\boldsymbol{\beta}}$, $i = 1, \ldots, n$.
b) Obtenha uma reamostra com reposição $\boldsymbol{\varepsilon}_i^*$, $i = 1, \ldots, n$ de $\boldsymbol{\varepsilon}_i$, $i = 1, \ldots, n$.
c) Calcule $\mathbf{y}_i^* = \mathbf{x}_i^T \widehat{\boldsymbol{\beta}} + \boldsymbol{\varepsilon}_i^*$.

d) Calcule $\widehat{\boldsymbol{\beta}}^{*(1)}$ a partir de $[\mathbf{y}^{*}, \mathbf{X}]$ da mesma maneira que $\widehat{\boldsymbol{\beta}}$ é calculado a partir dos dados originais $[\mathbf{y}, \mathbf{X}]$

e) Obtenha a distribuição bootstrap calculando $\widehat{\boldsymbol{\beta}}^{*(\mathrm{b})}$, $\mathrm{b} = 1, \ldots, \mathrm{B}$.

3.3.2 Bootstrap dos pares

Outro procedimento de reamostragem bootstrap, normalmente chamado de bootstrap dos pares ou bootstrap dos vetores, é frequentemente usado em situações em que há algumas dúvidas sobre a adequação da função de regressão que está sendo considerada ou quando a variância do erro não é constante e/ou quando os regressores não são variáveis fixas (MONTGOMERY *et al.*, 2012). No método bootstrap dos pares, os próprios dados originais (que são vetores) devem ser reamostrados, procedimento este apresentado no Algoritmo 2.

Algoritmo 2: Bootstrap dos pares (DAVISON e HINKLEY, 1997).

Para $\mathrm{r} = 1, \ldots, \mathrm{B}$,

a) Obtenha $\mathrm{i}_1^{*}, \ldots, \mathrm{i}_\mathrm{n}^{*}$ uma amostra aleatória com reposição extraída de $\{1, \ldots, \mathrm{n}\}$.

b) Para $\mathrm{j} = 1, \ldots, \mathrm{n}$, considere $\mathrm{x}_\mathrm{j}^{*} = \mathrm{x}_{\mathrm{i}_\mathrm{j}^{*}}$ e $\mathrm{y}_\mathrm{j}^{*} = \mathrm{y}_{\mathrm{i}_\mathrm{j}^{*}}$; então

c) Ajuste o modelo em $(\mathrm{x}_1^{*}, \mathrm{y}_1^{*}), \ldots, (\mathrm{x}_\mathrm{n}^{*}, \mathrm{y}_\mathrm{n}^{*})$ e obtenha $\hat{\beta}^{*(\mathrm{r})}$.

No modelo linear, quando os resíduos são normalmente distribuídos, o método bootstrap dos resíduos produz resultados acurados. No entanto, o método bootstrap dos pares, na prática, é mais robusto para desvios da normalidade e/ou má especificação do modelo (CHERNICK e LABUDDE, 2011).

3.4 Seleção de modelos utilizando bootstrap

Um dos interesses, ao se realizar uma análise de regressão, é selecionar o modelo mais parcimonioso, ou seja, o modelo que melhor se ajusta aos dados, com um número menor de parâmetros. Dentre os diversos métodos utilizados para seleção de modelos, o critério de informação de Akaike (AIC) (AKAIKE, 1973) é o mais comumente aplicado (KAMO et al., 2013), porém, em pequenas amostras, o AIC é viesado e tende a selecionar modelos super-parametrizados, ou seja, com elevada quantidade de parâmetros (HURVICH e TSAI, 1989). Neste contexto, Austin e Tu (2004) propuseram um método de seleção de modelo, apresentado no Algoritmo 3, que combina reamostragem bootstrap com métodos automatizados de seleção de variáveis.

Algoritmo 3: Método de seleção de modelos utilizando bootstrap.

a) Considere a matriz $[\mathbf{Y}, \mathbf{X}]$ formada com os dados originais.
b) Obtenha B reamostras da matriz anterior utilizando o método bootstrap dos pares.
c) Para cada uma, ajuste um modelo e use o método *backward* (YAN e SU, 2009) via AIC.
d) Para cada variável, determine a frequência com que ela foi selecionada nos B modelos e as porcentagens de vezes em que os coeficientes das variáveis selecionadas apresentaram sinais positivos e negativos.
e) Utilize os resultados do passo anterior para determinar modelos. Deve-se descartar variáveis com baixas porcentagens. Selecione o melhor modelo dentre os candidatos.

Conforme explicam Austin e Tu (2004), o método proposto utiliza a reamostragem bootstrap, não para avaliar a distribuição de um teste estatístico, mas para avaliar a distribuição de uma variável indicadora que denota a inclusão de uma variável de previsão específica em um modelo obtido utilizando eliminação *backward*.

3.5 Diagnósticos de influência global utilizando bootstrap na variável resposta

A detecção de pontos influentes na amostra é de grande importância porque eles exercem grande influência sobre a estimação de parâmetros na análise de regressão (TÜRKAN e TOKTAMIŞ, 2012). Numerosos métodos de diagnósticos são apresentados na literatura, em que a distância de Cook (COOK, 1977) é a mais utilizada para o estudo de influência global na variável resposta (MATSON *et al.*, 1994). A distância de Cook, representada por $\mathrm{D_i}$ na Equação (7), mede o afastamento dos vetores estimados com e sem o ponto considerado influente, isto é,

$$\mathrm{D_i} = \left[\sum_{\mathrm{j=1}}^{\mathrm{n}} \left(\widehat{\mathrm{Y}}_{\mathrm{j}} - \widehat{\mathrm{Y}}_{\mathrm{j(i)}}\right)^2\right] / (\mathrm{p\ MSE}), \tag{7}$$

em que $\widehat{\mathrm{Y}}_{\mathrm{j}}$ é o valor estimado obtido do modelo de regressão completo para a observação j, $\widehat{\mathrm{Y}}_{\mathrm{j(i)}}$ é o valor estimado da observação j de um modelo de regressão em que a observação i foi omitida, MSE é o erro quadrático médio do modelo de regressão e p é o número de parâmetros ajustados no modelo.

Há mais de uma recomendação a respeito de quais pontos de corte que devem ser usados para se a detecção dos pontos influentes. Martin e Roberts (2010) apontam como pontos influentes aqueles em que $\mathrm{D_i} > 1$ ou comparar $\mathrm{D_i}$ com a mediana da distribuição F de Snedecor com graus de liberdade p e $\mathrm{n} - \mathrm{p}$ e Bolboaca e Jantschi (2013) indicam que pontos influentes são aqueles tal que $\mathrm{D_i} > 4/\mathrm{n}$.

Conforme explicam Beyaztas e Alin (2012), quando a distribuição dos erros é normal e o tamanho amostral n é grande, os pontos de corte apresentados funcionam bem. No entanto, no caso da distribuição dos erros não ser normal e em amostras pequenas, eles podem não ser adequados para a detecção das reais observações influentes. Para superar estes problemas, Martin e Roberts (2010) propuseram um método bootstrap baseado na técnica *Jackknife-after-Bootstrap* (JaB) B desenvolvida por Efron (1992) para obter uma aproximação das distribuições amostrais das medidas de influência de interesse. Devido ao processo de construção das reamostragens bootstrap, um ponto pode aparecer várias vezes nas reamostragens. Assim, se o conjunto de dados original contém um ponto influente, a observação pode aparecer nas amostras bootstrap e fornecer resultados insatisfatórios logo, e, a fim de que se obtenham resultados adequados, os pontos de corte devem ser determinados a partir das amostras bootstrap que não contêm o ponto em questão. Efron (1992) mostrou que uma amostra de tamanho n, obtida de $z_1, \dots, z_{i-1}, z_{i+1}, \dots, z_n$, tem a mesma distribuição que uma amostra obtida de $z_1, \dots, z_n$ em que nenhum dos valores é igual a z_i, e, baseado neste lema, Martin e Roberts (2010) propuseram o Algoritmo 4.

Algoritmo 4: Determinação do ponto de corte para D_i utilizando JaB.

a) Ajuste o modelo proposto ao conjunto de dados original e calcule D_i, $i = 1, \dots, n$;
b) Construa B amostras bootstrap utilizando o método bootstrap dos pares;
c) (Passo JaB) Para cada amostra x_i do conjunto de dados original, considere o grupo de amostras bootstrap que não contêm a amostra x_i (aproximadamente B/e grupos) e, para cada amostra deste grupo, calcule os n valores da distância de Cook. Agrupe todos os $n.(B/e)$ valores em um único vetor[1];
d) Os quantis 2,5% e 97,5% da distribuição gerada pelos $n.(B/e)$ valores das distâncias de Cook são usados como pontos de corte e se o valor de D_i estiver fora deste intervalo, x_i é marcado como um ponto influente.

Segundo Martin e Roberts (2010), a lógica desta abordagem é obter uma distribuição bootstrap nula da distância de Cook sob a hipótese de que a i-ésima amostra não é influente. Uma vez que a i-ésima amostra não está presente nas amostras bootstrap, as quais são utilizadas para obter a distribuição bootstrap, ela não pode exercer influência, e com isso, a distribuição gerada é livre da influência deste ponto.

[1] Dado o conjunto amostral $\{y_1, \dots, y_n\}$, a probabilidade de y_j não estar inclusa em uma amostra bootstrap que é de $(1 - n^{-1})^n = e^{-1}$. Assim, em B amostras bootstrap, o número de simulações que não incluem y_j é aproximadamente $B \cdot e^{-1}$, $e \approx 2{,}718$. (DAVISON e HINKLEY, 1997).

3.6 Gráfico Jackknife-after-Bootstrap

A técnica JaB fornece outro recurso que permite estabelecer o efeito de observações individuais sobre a distribuição bootstrap, via elaboração do gráfico *jackknife-after-bootstrap* (EFRON, 1992). Conforme explicam Davison e Hinkley (1997), é possível mensurar o efeito de um ponto amostral nos cálculos e comparar o conjunto de todas as réplicas bootstrap de um parâmetro com o subconjunto de réplicas bootstrap obtidas das amostras bootstrap que não contêm o ponto em questão.

Com base no conjunto de dados original $[\mathrm{Y}, \mathrm{X}]$, considere o conjunto de dados $\left[\mathrm{Y}_{(\mathrm{i})}, \mathrm{X}_{(\mathrm{i})}\right]$ obtido com a exclusão da linha i no conjunto de dados original e calcule a estatística de interesse, denotada por $\mathrm{s}_{(\mathrm{i})}$. A função de influência jackknife para a estatística de interesse é definida por:

$$\mathrm{u_i}\{\mathrm{s}\} = (\mathrm{n}-1)\left(\mathrm{s}_{(\cdot)} - \mathrm{s}_{(\mathrm{i})}\right), \tag{8}$$

em que, $\mathrm{s}_{(\cdot)} = \left[\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{s}_{(\mathrm{i})}\right]/\mathrm{n}$. Intuitivamente, pontos com elevados valores positivos ou negativos de $\mathrm{u_i}\{\mathrm{s}\}$ apresentam elevada influência na estatística calculada. Por fornecer uma interpretação mais clara, comumente se utiliza a função de influência jackknife relativa apresentada na Equação 9, cujo valor dois é definido como ponto de corte (EFRON, 1992). Tais valores são ordenados crescentemente e marcados no eixo das abscissas.

$$\mathrm{u_i^{\uparrow}}\{\mathrm{s}\} = \mathrm{u_i}\{\mathrm{s}\}/\left[\sum_{\mathrm{j}} \mathrm{u_j}\{\mathrm{s}\}^2/(\mathrm{n}-1)\right]^{\frac{1}{2}}. \tag{9}$$

Após o cálculo dos valores de influência jackknife, para cada ponto i do conjunto de dados determinam-se sete pares ordenados, a saber, $\left(\mathrm{u_i^{\uparrow}}\{\mathrm{s}\}, \mathrm{P_k}\right)$, $\mathrm{k} = \{5{,}10{,}16{,}50{,}94{,}90{,}95\}$ em que, $\mathrm{P_k}$ representa o k-ésimo percentil da distribuição bootstrap, formada com as réplicas bootstrap, calculadas a partir das amostras bootstrap que não contêm o ponto i. Para cada percentil, os pares ordenados vizinhos são ligados formando gráficos, os quais são comparados com segmentos de retas pontilhadas perpendiculares ao eixo das ordenadas nos pontos $\mathrm{P_k}$, $\mathrm{k} = \{5{,}10{,}16{,}50{,}94{,}90{,}95\}$, calculados da distribuição bootstrap completa e formada por 3000 réplicas bootstrap. A análise é realizada com destaque para os pontos que ultrapassam o ponto de corte e pela comparação entre as distribuições bootstrap (DAVISON e HINKLEY, 1997).

3.7 Bootstrap com amostras dependentes

Um problema que ocorre na aplicação da metodologia bootstrap é que, ao utilizá-la em dados autocorrelacionados, o processo de reamostragem com reposição rompe a estrutura de correlação dos dados e, por conseguinte, gera uma estimativa incorreta do erro padrão. Deste modo, para utilizar a metodologia bootstrap de modo eficaz, é necessário encontrar alguma forma de gerar amostras que preservem a autocorrelação dos dados (PLANT, 2012).

3.7.1 Bootstrap espacial

O método bootstrap espacial, proposto por Solow (1985), apresentado no Algoritmo 5, permite a obtenção de réplicas bootstrap espacialmente correlacionadas.

Algoritmo 5: Bootstrap espacial.

a) Considerando o conjunto de dados espaciais $\{\mathrm{Z}(\mathbf{s}_1),\ldots,\mathrm{Z}(\mathbf{s}_\mathrm{n})\}$, determine o vetor dos resíduos $\hat{\boldsymbol{\varepsilon}} = (\,\mathrm{Z}(\mathbf{s}_1) - \hat{\mu}, \ldots, \mathrm{Z}(\mathbf{s}_\mathrm{n}) - \hat{\mu})^{\mathrm{T}}$ sendo $\hat{\mu} = \left(1^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}^{-1}1\right)^{-1}1^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}^{-1}\mathrm{Z}$ o estimador de mínimos quadrados de μ e $\widehat{\boldsymbol{\Sigma}}$ a matriz de covariância estimada;
b) Considerando a matriz de covariância estimada $\hat{\Sigma}$, utilize o método de decomposição de Cholesky para obter $\widehat{\boldsymbol{\Sigma}} = \hat{\mathbf{L}}\hat{\mathbf{L}}^{\mathrm{T}}$, em que $\hat{\mathbf{L}}$ é uma matriz triangular inferior de ordem n;
c) Utilizando a matriz $\hat{\mathbf{L}}^{-1}$, determine $\hat{\boldsymbol{\varepsilon}}_{\mathrm{dec}} = \hat{\mathbf{L}}^{-1}\hat{\boldsymbol{\varepsilon}}$, o vetor de resíduos descorrelacionados e centralize seus valores, obtendo $\tilde{\mathbf{e}} = \hat{\boldsymbol{\varepsilon}}_{\mathrm{dec}} - \left(\frac{1}{\mathrm{n}}\right)\sum\hat{\boldsymbol{\varepsilon}}_{\mathrm{dec}}$;
d) Considerando o conjunto dos resíduos descorrelacionados e centralizados $\tilde{\mathrm{e}}_1,\ldots,\tilde{\mathrm{e}}_\mathrm{n}$, realize uma reamostragem com reposição, obtendo o vetor $\boldsymbol{\varepsilon}^*_{\mathrm{SB}} = (\varepsilon^*_1,\ldots,\varepsilon^*_\mathrm{n})^{\mathrm{T}}$;
e) A amostra bootstrap espacial é obtida recorrelacionando os resíduos bootstrap $\mathbf{Z}^* = \hat{\mu} + \hat{\mathbf{L}}\boldsymbol{\varepsilon}^*_{\mathrm{SB}}$.

3.7.2 Bootstrap espacial paramétrico

O método bootstrap espacial paramétrico foi proposto por Tang *et al.* (2006) e consiste de uma modificação do método de Solow que visa a uma melhor cobertura dos intervalos de confiança. O método bootstrap espacial paramétrico é apresentado no Algoritmo 6 e sua criação foi motivada pelos ajustes de semivariogramas Gaussianos em estudos de tomografia computadorizada de regiões profundas do cérebro.

Algoritmo 6: Bootstrap espacial paramétrico.

a) Considerando o conjunto de dados espaciais $\{Z(s_1),\ldots, Z(s_n)\}$, determine o vetor dos resíduos $\hat{\varepsilon} = (Z(s_1) - \hat{\mu}, \ldots, Z(s_n) - \hat{\mu})^T$ sendo $\hat{\mu} = \left(1^T\hat{\Sigma}^{-1}1\right)^{-1}1^T\hat{\Sigma}^{-1}Z$ o estimador de mínimos quadrados de μ e $\hat{\Sigma}$ a matriz de covariância estimada;
b) Considerando a matriz de covariância estimada $\hat{\Sigma}$, utilize o método de decomposição de Cholesky para obter $\hat{\Sigma} = \hat{L}\hat{L}^T$, em que $\hat{L}$ é uma matriz triangular inferior de ordem n;
c) Utilize a distribuição N(0,1) para gerar um vetor de resíduos bootstrap paramétricos $\varepsilon^*_{PSB} = (\varepsilon^*_1, \ldots, \varepsilon^*_n)^T$.
d) A amostra bootstrap espacial é obtida por $Z^* = \hat{\mu} + \hat{L}\varepsilon^*_{PSB}$.

O método bootstrap espacial paramétrico não descorrelaciona o vetor dos resíduos $\hat{\varepsilon}$ como no método bootstrap espacial. Ao invés disto, os resíduos são gerados independentes a partir de uma distribuição normal padrão.

4 REFERÊNCIAS BIBLIOGRÁFICAS

AKAIKE, H. Information theory and an extension of the maximum likelihood principle. In: **Proceedings of the second international symposium on information theory**. PETROV, B.N., CSAKI, F. (eds), p. 267–281. Budapest: Akadémia Kiado, 1973.

AUSTIN, P.; TU, J. Bootstrap methods for developing predictive models. **American Statistician**, v. 58, n. 2, p. 131–137, 2004.

BEYAZTAS, U.; ALIN, A. Jackknife-After-Bootstrap Method for Detection of Influential Observations in Linear Regression Models. **Communication in Statistics-Simulation and Computation**, v. 42, n. 6, p. 1256-1267, 2012.

BOLBOACA, S.; JANTSCHI, L. The effect of leverage and/or influential on structure-activity relationships. **Combinatorial Chemistry & High Throughput Screening**, v. 16, n. 1, p. 288-297, 2013.

CHERNICK, M. R. **Bootstrap methods**: a guide for practitioners and researchers. 2ª ED. New Jersey: John Wiley e Sons, 2008.

CHERNICK, M. R.; LABUDDE, R. A. Revisiting qualms about bootstrap confidence intervals. **American Journal of Mathematical and Management Sciences**, v. 29, p. 437-456, 2010.

CHERNICK, M. R.; LABUDDE, R. A. **An introduction to bootstrap methods with applications to R**. New Jersey: John Wiley e Sons, 2011.

COOK, R. D. Detection of Influential observation in Linear regression. **Technometrics**, v. 19, n. 1, p. 15-18, 1977.

CRAMER, H. **Mathematical Methods of Statistics**. Princeton: Princeton University Press, 1946.

DAVISON, A. C.; HINKLEY, D. V. **Bootstrap methods and their application**. Cambridge: Press Syndicate of the University of Cambridge, 1997.

DIACONIS, P.; EFRON, B. Computer Intensive methods in statistics. **Scientific American**, v. 248, n. 1, p. 116-130, 1983.

EFRON, B. Bootstrap methods: Another look at the jackknife. **Annals of Stattistics**, v. 7, p. 1-26, 1979.

EFRON, B. **The jackknife, the bootstrap and other resampling plans**. Philadelphia: SIAM, 1982.

EFRON, B. Jackknife-after-bootstrap standard errors and influence functions. **Journal of the Royal Statistical Socciety**, v. 54, n. 1, p. 83-127, 1992.

EFRON, B.; GONG, G. A leisurely look at the bootstrap, the jackknife and cross validation. **American Statistician**, v. 37, n. 1, p. 36-48, 1983.

EFRON, B.; TIBSHIRANI, R. Bootstrap methods for standard errors, confidence intervals and other measures of statistical accuracy. **Statistical Science**, v. 1, n. 1, p. 54-75, 1986.

EFRON, B.; TIBSHIRANI, R. **An introduction to the bootstrap**. New York: Chapman e Hall, 1993.

FISHER, R. A. **The design of experiments**. New York: Hafner, 1935.

GOOD, P. I. **Permutation, parametric, and bootstrap tests of hypotheses**. New York: Springer, 2006.

HALL, P. Resampling a coverage process. **Stochastic Processes and their Applications**, v. 20, n. 2, p. 231-246, 1985.

HOFFMANN, R. **Análise de Regressão: uma introdução à econometria**. São Paulo: Hucitec, 2006.

HURVICH, C. M.; TSAI, C. L. Regression and time series model selection in small samples. **Biometrika**, v. 76, n. 2, p. 297-307, 1989.

KAMO, K.; YANAGIHARA, H.; SATOH, K. Bias-corrected AIC for selecting variables in poisson regression models. **Communications in Statistics part A – Theory and Methods**, v. 42, n. 11, p. 1911-1921, 2013.

MARTIN, M. A.; ROBERTS, S. Jackknife-after-bootstrap regression influence diagnostics. **Journal of Nonparametric Statistics**, v. 22, n. 2, p. 257-269, 2010.

MATSON, J. E.; BARRETT, B. E.; MELLICHAMP, J. M. Software development cost estimation using function points. **IEEE Transactions on Software Engineering**, v. 20, n. 4, p. 275-287, 1994.

MONTGOMERY, D. C.; PECK, E. A.; VINING, G. G. **Introduction to Linear Regression Analysis**. New Jersey: John Wiley e Sons, 2012.

MOONEY, C. Z.; DUVAL, R. D. **Bootstrapping**: a nonparametric approach to statistical inference. Newbury Park: Sage,1993.

PLANT, R.E. **Spatial data analysis in ecology and agriculture using R**. Boca Raton: CRC Press, 2012.

QUENOUILLE, M. H. Approximate tests of correlations in times series. **Journal of the Royal Stattistical Socciety**, v. 11, p. 68-84, 1949.

SHASHA, D.; WILSON, M. **Statistic is easy**. San Rafael: Morgan e Claypool Publishers, 2011.

SHERMAN, M. **Spatial Statistics and Spatio-Temporal Data: covariance functions and directional properties**. United Kingdom: John Wiley e Sons, 2010.

SIMON, J. L. **Basic research in social science**. New York: Random House, 1969.

SOLOW, A. Bootstrapping correlated data. **Mathematical Geology**, v. 17, n. 7, p. 769–775, 1985.

TANG, L.; SCHUCANY, W.; WOODWARD, W.; GUNST, R. **A parametric spatial bootstrap. Technical Teport SMU-TR-337**. Dallas: Southern Methodist University, 2006.

TÜRKAN, S.; TOKTAMIŞ, Ö. Detection of influential observations in ridge regression and modified ridge regression. **Journal of Model Assisted Statistics and Applications**, v. 7, n. 2, p. 91-97, 2012.

WASSERMAN, L. **All of nonparametric statistics**. New York: Springer, 2006.

YAN, X.; SU, X. G. **Linear Regression Analysis: Theory and Computing**. Singapore: World Scientific Publishing Co., 2009.

5 ARTIGOS

5.1 ARTIGO 1 : Modelagem da produtividade de soja utilizando métodos bootstrap para pequenas amostras[2]

Resumo: Um dos problemas que ocorrem ao se trabalhar com modelos de regressão diz respeito ao tamanho amostral; pois como os métodos estatísticos utilizados nas análises inferenciais são assintóticos, caso a amostra seja pequena, as análises podem ficar comprometidas, pois as estimações serão viesadas. Uma alternativa é utilizar a metodologia bootstrap, que em sua versão não paramétrica, e não necessita supor nem conhecer a distribuição de probabilidade que gerou a amostra original. Neste trabalho, utiliza-se um conjunto de dados de produtividade de soja e atributos físicos e químicos do solo formado com poucas amostras para determinar um modelo de regressão linear múltipla. Foram utilizados métodos bootstrap para seleção de variáveis, identificação de pontos influentes pela análise de diagnóstico de influência global e determinação de intervalos de confiança dos parâmetros do modelo. Os resultados mostraram que os métodos bootstrap permitiram selecionar os atributos físicos e químicos do solo que foram significativos na construção do modelo de produtividade de soja, construir os intervalos de confiança dos parâmetros e identificar os pontos que tiveram grande influência sobre os parâmetros estimados.

Palavras-chave: diagnóstico de influência global bootstrap; intervalos de confiança bootstrap; regressão linear múltipla; seleção de modelos.

5.1.1 Introdução

A soja (Glycine max (L.) Merrill) é uma das mais importantes culturas agrícolas no que tange à economia (Kulcheski *et al*., 2016). Assim, modelos de regressão são constantemente desenvolvidos com o objetivo de explicar parte da variabilidade da sua produtividade. Na modelagem da produtividade da soja, é comum o uso de variáveis agrometeorológicas (Penalba *et al*., 2007; Tao *et al*., 2008), variáveis agrícolas (Zheng *et al.*, 2009), variáveis de gestão (Lobell *et al.*, 2005), índices de vegetação (Mercante *et al*., 2010) e parâmetros do solo (Garcia-Paredes *et al.*, 2000). Os modelos diferem pela natureza das variáveis explicativas utilizadas na modelagem e uma revisão sobre as classes de modelos pode ser vista em Vera-diaz *et al*. (2008).

Como a determinação dos valores de algumas variáveis muitas vezes é uma tarefa onerosa e laboriosa, em alguns casos as análises são realizadas com amostras pequenas[3].

[2] Este artigo já foi publicado (DOI: 10.5424/sjar/2016143-8635) e segue as normas da Revista Spanish Journal of Agricultural Research (SJAR). As normas podem ser consultadas no Anexo 1 desta tese.

Este fato pode trazer dúvidas em relação às inferências realizadas, pois os métodos tradicionais de inferência são assintóticos e conforme explicam Hao e Naiman (2010), erros-padrão e intervalos de confiança baseados em teoria assintótica podem ser viesados em amostras pequenas.

A seleção de modelos mais parcimoniosos e a determinação de pontos influentes são procedimentos que também podem fornecer resultados enganosos quando se trabalha com um conjunto amostral pequeno. Kamo *et al*. (2013) explicam que o Critério de Informação de Akaike – AIC (Akaike, 1973), utilizado para a seleção de modelos, tem um viés que não pode ser ignorado, principalmente em pequenas amostras, tendo em vista que ele é derivado de propriedades assintóticas.

Em relação às medidas de diagnóstico de influência global, um dos problemas está relacionado aos seus pontos de corte. De acordo com Martin e Roberts (2010), eles são baseados na teoria das grandes amostras, portanto, podem não ser adequados para pequenas amostras.

Uma alternativa aos métodos tradicionais utilizados em análise de regressão é o uso do bootstrap, um método de simulação desenvolvido por Efron (1979) que utiliza reamostragens com reposição do conjunto de dados para realizar inferências estatísticas como testes de hipótese e determinação de intervalos de confiança (Dubreuil *et al.*, 2014). O método bootstrap tem aplicações em análise de regressão (Rahman, 2014), seleção de modelos (Al-Marshadi, 2011) e definição de diagnósticos de influência global (Beyaztas e Alin, 2013).

Ao serem comparados os resultados obtidos de métodos bootstrap com resultados obtidos de métodos assintóticos, Chaves-Neto e Faria (2015) concluíram que o bootstrap teve um bom desempenho em amostras de todos os tamanhos e foi superior ao método assintótico em amostras pequenas. Embora o bootstrap seja uma técnica conhecida e frequentemente utilizada em estudos agrícolas – como pode ser observado nos trabalhos de Sabaghnia *et al*. (2010), García-Gallego *et al*. (2015), Losada *et al*. (2015) e Sutton *et al*. (2016), o desenvolvimento de métodos estatísticos e computacionais impulsionou o estudo de novas técnicas que utilizam o bootstrap.

O objetivo deste trabalho é utilizar métodos bootstrap para selecionar variáveis explicativas, investigar a existência de pontos influentes a partir de estudos de diagnósticos e obter intervalos de confiança dos parâmetros de um modelo de regressão linear múltipla da produtividade de soja considerando atributos físicos e químicos do solo como variáveis explicativas.

---

[3] Não existe uma definição aceita sobre o que constitui uma amostra pequena, pois o tamanho da amostra depende de um número de fatores, incluindo a confiabilidade das estimativas e a variância relativa da variável sob consideração (Levy e Lemeshow, 1980). Aiken *et al*. (1991) consideram $n \leq 60$ como amostra pequena enquanto Irerland (2010) considera como amostra pequena $n \leq 30$.

5.1.2 Material e métodos

5.1.2.1 Área de estudo e dados

Os dados utilizados são do ano agrícola 2013/2014 e provenientes de uma área agrícola comercial de 167,35 hectares, localizada na região Oeste do Paraná, próxima ao município de Cascavel, com coordenadas centrais latitude 24º57'18''S e longitude 53º34'29''W, e altitude média de 714 metros (Figura 1).

O clima da região apresenta-se como temperado mesotérmico e superúmido, tipo climático Cfa (Koeppen) e o solo é classificado como Latossolo Vermelho distroférrico (EMBRAPA, 2013). Considerando-se um conjunto de 30 pontos não correlacionados de produtividade de soja (Prod, t ha^{-1}), foi confirmada a independência pelo teste dos RUNS (Siegel, 1956).

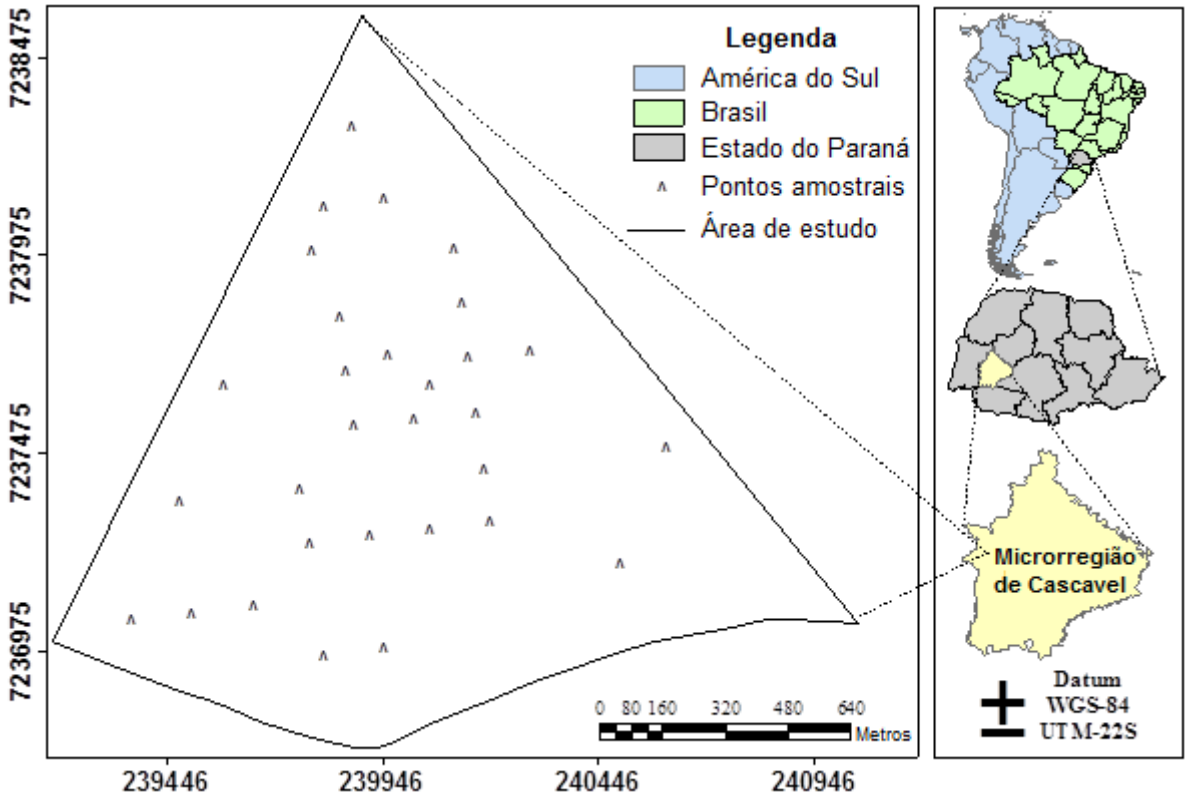

Figura 1 – Mapa de localização da área em estudo.

Os respectivos valores das variáveis explicativas RSP_1, RSP_2 and RSP_3 (resistência do solo à penetração, MPa, de 0 a 0,1 m, 0,1 a 0,2 m e 0,2 a 0,3 m de profundidade, respectivamente), Ca (cálcio, $cmolc/dm^3$), Mg (magnésio, $cmolc/dm^3$), K (potássio, mg/dm^3), P (fósforo, mg/dm^3), Mn (manganês, mg/dm^3), Des_1, Des_2 e Des_3 (densidade do solo, g/cm^3, de 0 a 0,1 m, 0,1 a 0,2 m e 0,2 a 0,3 m de profundidade, respectivamente) foram considerados para cada valor de produtividade.

O uso de atributos físicos e químicos como variáveis explicativas é uma prática comum em estudos de campo, pois as variações nas propriedades do solo são responsáveis pela maioria das variações de rendimento das culturas, conforme destacam Khakural *et al.* (1999).

5.1.2.2 Análise exploratória e modelagem

Foram calculadas estatísticas descritivas das variáveis em estudo e foi realizada uma análise de multicolinearidade das variáveis explicativas. Um modelo de regressão linear múltipla foi construído para descrever a relação entre a produtividade de soja e os atributos do solo, com parâmetros estimados pelo método dos mínimos quadrados ordinários (OLS).

5.1.2.3 Bootstrap dos pares

O método bootstrap dos pares (FREEDMAN, 1981) foi utilizado para determinar as réplicas bootstrap dos parâmetros do modelo, apresentado no seguinte algoritmo:

Algoritmo 1: Bootstrap dos pares.

a) Considere a matriz $[\mathrm{Y}, \mathrm{X}]$ formada com os dados originais; b) Obtenha uma nova matriz $\left[\mathrm{Y}^{*(1)}, \mathrm{X}^{*(1)}\right]$ ao fazer uma reamostragem com reposição das linhas da matriz $[\mathrm{Y}, \mathrm{X}]$; c) Calcule $\hat{\beta}^{*(1)}$ a partir de $\left[\mathrm{Y}^{*(1)}, \mathrm{X}^{*(1)}\right]$ da mesma maneira que $\hat{\beta}$ é calculado a partir dos dados originais $[\mathrm{Y}, \mathrm{X}]$; d) Obtenha a distribuição bootstrap pelo cálculo de $\hat{\beta}^{*(\mathrm{b})}$, $\mathrm{b} = 1, \ldots, \mathrm{B}$.

5.1.2.4 Intervalos de confiança utilizando bootstrap

A determinação dos intervalos bootstrap para os parâmetros do modelo de regressão foi obtida pelos métodos percentil (EFRON,1982) e BC (*bias corrected* – viés corrigido) (EFRON e TIBSHIRANI, 1986). O intervalo percentil de Efron com nível de confiança $(1 - \alpha)\%$ é obtido ao serem ordenadas as estimativas de forma crescente $\hat{\theta}_{\mathrm{i}}^{*}$, $\mathrm{i} = 1, \ldots, \mathrm{B}$, e pela exclusão de $(\alpha/2)\%$ das réplicas situadas nas extremidades.

A técnica utilizada para a construção do intervalo de confiança BC utiliza um valor conhecido como constante de correção de viés para ajustar a distribuição bootstrap de $\hat{\theta}$; e um roteiro para determinar este intervalo pode ser visto em Shasha e Wilson (2011).

5.1.2.5 Seleção de modelos utilizando bootstrap

Para seleção de modelo, utilizou-se o método bootstrap proposto por Austin e Tu (2004), apresentado no Algoritmo 2, que combina reamostragem bootstrap com método automotizado de seleção de variáveis.

Algoritmo 2: Método de seleção de modelos utilizando bootstrap.

a) Considere a matriz $[\mathrm{Y}, \mathrm{X}]$ formada com os dados originais; b) Obtenha B reamostras da matriz anterior utilizando o método bootstrap dos pares; c) Para cada reamostra, ajuste um modelo e aplique o método *backward* via AIC; d) Para cada variável, determine a frequência com que ela foi selecionada nos B modelos e as porcentagens de vezes em que os coeficientes das variáveis selecionadas apresentaram sinais positivos e negativos; e) Utilize os resultados do passo anterior para determinar modelos candidatos e selecione o melhor modelo.

5.1.2.6 Diagnósticos de influência global utilizando bootstrap na variável resposta

Utilizou-se o método proposto por Martin e Roberts (2010) para investigar a existência de pontos influentes, considerando a distância de Cook – Di (Cook, 1977) como medida de influência. O Algoritmo 3 apresenta o método proposto por Martin e Roberts, que é baseado na técnica JaB (*jackknife-after-bootstrap*), desenvolvida por Efron (1992).

Algoritmo 3: Determinação do ponto de corte para $\mathrm{D_i}$ utilizando JaB.

a) Ajuste o modelo proposto ao conjunto de dados original e calcule $\mathrm{D_i},\ \mathrm{i} = 1, \ldots, \mathrm{n}$; b) Construa B amostras bootstrap utilizando o método bootstrap dos pares; c) (Passo JaB) Para cada amostra $\mathrm{x_i}$ do conjunto de dados original, considere o grupo de amostras bootstrap que não contém a amostra $\mathrm{x_i}$ (aproximadamente $\mathrm{B/e}$ grupos) e para cada amostra deste grupo, calcule os n valores da distância de Cook. Agrupe todos os $\mathrm{n.(B/e)}$ valores em um único vetor[4]; d) os quantis 2,5% e 97,5% da distribuição gerada pelos $\mathrm{n.(B/e)}$ valores das distâncias de Cook são usados como pontos de corte e se o valor de $\mathrm{D_i}$ estiver fora deste intervalo, $\mathrm{x_i}$ é marcado como um ponto influente.

[4] Dado o conjunto amostral $\{y_1, \ldots, y_n\}$, a probabilidade de y_j não estar incluso em uma amostra bootstrap é de $(1 - n^{-1})^n = e^{-1}$, assim, em B amostras bootstrap, o número de simulações que não incluem y_j é aproximadamente $B \cdot e^{-1}$ (DAVISON e HINKLEY, 1997).

5.1.2.7 Gráfico Jackknife-after-Bootstrap

A técnica JaB fornece outro recurso que permite estabelecer o efeito de observações individuais sobre a distribuição bootstrap, por elaboração do *jackknife-after-bootstrap plot* (Efron, 1992). Com base no conjunto de dados original $[Y, X]$, considere o conjunto de dados $[Y_{(i)}, X_{(i)}]$ obtido com a exclusão da linha i no conjunto de dados original e calcule a estatística de interesse, denotada por $s_{(i)}$. A função de influência jackknife para a estatística de interesse é definida por:

$$u_i\{s\} = (n-1)(s_{(\cdot)} - s_{(i)}) \quad (1)$$

em que, $s_{(\cdot)} = [\sum_{i=1}^{n} s_{(i)}]/n$. Intuitivamente, pontos com elevados valores positivos ou negativos de $u_i\{s\}$ apresentam elevada influência na estatística calculada. Por fornecer uma interpretação mais clara, comumente se utiliza a função de influência jackknife relativa apresentada na Equação 8 sendo o valor dois definido como ponto de corte (Efron, 1992). Estes valores são ordenados crescentemente e marcados no eixo das abscissas.

$$u_i^{\uparrow}\{s\} = u_i\{s\}/[\sum_j u_j\{s\}^2/(n-1)]^{\frac{1}{2}}. \quad (2)$$

Após o cálculo dos valores de influência jackknife, para cada ponto i do conjunto de dados, determinam-se sete pares ordenados, a saber, $(u_i^{\uparrow}\{s\}, P_k)$, $k = \{5,10,16,50,94,90,95\}$ em que, P_k representa o k-ésimo percentil da distribuição bootstrap formada com as réplicas bootstrap, calculadas das amostras bootstrap que não contêm o ponto i.

Para cada percentil, os pares ordenados vizinhos são ligados para a formação de gráficos, os quais são comparados com segmentos de retas pontilhadas perpendiculares ao eixo das ordenadas nos pontos P_k, $k = \{5,10,16,50,94,90,95\}$, calculados da distribuição bootstrap completa formada por 3000 réplicas bootstrap. A análise é realizada com destaque para os pontos que ultrapassam o ponto de corte e pela comparação das distribuições bootstrap.

5.1.2.8 Recursos computacionais

As análises realizadas neste trabalho foram desenvolvidas no *software* estatístico R (R Core Team, 2014). As réplicas bootstrap utilizadas para determinar as distribuições empíricas das estimativas dos parâmetros dos modelos foram determinadas pela função

lm.boot do pacote simpleboot (Peng, 2008) e os intervalos de confiança foram implementados manualmente.

A determinação das estatísticas relacionadas ao método de seleção de modelos foi obtida pela função boot.stepAIC do pacote bootStepAIC (Rizopoulos, 2009). O algoritmo utilizado para determinação do ponto de corte para a distância de Cook bootstrap foi implementado tendo a distância de Cook calculada pela função *cooks.distance* do pacote stats (R Core Team, 2014) e os gráficos JaB foram implementados pelos autores.

5.1.3 Resultados

As estatísticas descritivas das variáveis explicativas indicaram um comportamento homogêneo das mesmas e verificou-se que não existe multicolinearidade. O modelo de regressão linear múltipla da produtividade da soja estimado por OLS, considerando todas as variáveis explicativas (Equação 3), apresentou coeficiente de determinação ajustado (R^2_{Adj}) igual a 0,41 e raiz quadrada do erro quadrático médio (RMSE) igual a 0,33.

$$\begin{aligned}\text{Prod} = {} & 8{,}858 - 0{,}271\text{RSP}_1 + 0{,}117\text{RSP}_2 - 0{,}003\text{RSP}_3 + 0{,}288\text{Ca} - 0{,}367\text{Mg} \\ & + 1{,}208\text{K} - 0{,}067\text{P} - 0{,}012\text{Mn} - 0{,}629\text{Des}_1 - 2{,}684\text{Des}_2 + 0{,}925\text{Des}_3\end{aligned} \qquad [3]$$

Observa-se que as estimativas dos parâmetros associados às variáveis RSP_1, RSP_3, Des_1 e Des_2 apresentam sinais negativos. Isso mostra que o aumento do valor dessas variáveis implica redução da produtividade da soja (Eq. [3]).

As estimativas dos parâmetros associados às variáveis RSP_2 e Des_3 apresentaram sinais diferentes do esperado, pois indicam uma relação direta de tais variáveis com a produtividade de soja (Eq. [3]). O sinal positivo da estimativa do parâmetro associado à variável K indica que, mantendo as demais variáveis constantes, o aumento de uma unidade na variável K produz aumento, em média, de 1,208 t ha^{-1} na produtividade da soja (Eq. [3]).

Os intervalos bootstrap de 95% de confiança para os parâmetros do modelo de regressão linear múltipla utilizaram as técnicas bootstrap percentil de Efron e bootstrap com viés corrigido (BC bootstrap) (Tabela 1).

Destaca-se que, independente da técnica bootstrap utilizada para determinar os intervalos de confiança, a grande maioria dos intervalos contém o zero. Logo, é possível que, com exceção da variável P, as demais variáveis explicativas podem não ser significativas individualmente.

Tabela 1. Intervalos bootstrap de 95% de confiança para os parâmetros do modelo de regressão linear múltipla elaborado para todas as variáveis explicativas.

Parâmetros[1]	Percentil de Efron			BC[3]		
	$\hat{\theta}_l$	$\hat{\theta}_u$	Amplitude[2]	$\hat{\theta}_l$	$\hat{\theta}_u$	Amplitude
β_{RSP_1}	-0,547	0,130	0,677	-0,530	0,197	0,727
β_{RSP_2}	-0,457	0,750	1,208	-0,423	0,798	1,221
β_{RSP_3}	-1,260	0,923	2,183	-1,336	0,888	2,224
β_{Ca}	-0,214	0,685	0,898	-0,268	0,643	0,911
β_{Mg}	-1,219	0,589	1,808	-1,168	0,654	1,821
β_K	-4,771	6,607	11,378	-5,463	5,928	11,390
β_P	-0,150	-0,007	0,143	-0,160	-0,012	0,147
β_{Mn}	-0,027	0,008	0,035	-0,025	0,009	0,035
β_{Des_1}	-4,412	5,076	9,487	-3,937	6,213	10,150
β_{Des_2}	-6,426	0,585	7,012	-6,505	0,524	7,029
β_{Des_3}	-2,052	2,872	4,925	-2,677	2,498	5,175
Intercepto	2,903	13,952	11,049	2,965	14,279	11,314

[1] β_i : Parâmetro associado à variável i = {RSP_1, RSP_2, RSP_3, Ca, Mg, K, P, Mn, Des_1, Des_2, Des_3}; RSP_1, RSP_2 and RSP_3: resistência do solo à penetração, MPa, de 0 a 0,1 m, 0,1 a 0,2 m e 0,2 a 0,3 m de profundidade, respectivamente; Ca: cálcio, $cmol_c/dm^3$; Mg: magnésio, $cmol_c/dm^3$; K: potássio, mg/dm^3; P: fósforo, mg/dm^3; Mn: manganês, mg/dm^3; Des_1, Des_2 and Des_3: densidade do solo, g/cm^3, de 0 a 0,1 m, 0,1 a 0,2 m e 0,2 a 0,3 m de profundidade, respectivamente; [2]Amplitude: $\hat{\theta}_u - \hat{\theta}_l$; $\hat{\theta}_l$: limite inferior; $\hat{\theta}_u$ **:**limite superior; [3]BC: bias corrected (viés corrigido).

Em Busca de um modelo de regressão linear múltipla mais adequado, foi aplicado o método de seleção de modelos utilizando bootstrap considerando 1000 reamostras (Tabela 2). Observa-se que dos 1000 modelos que foram ajustados às reamostras bootstrap, ao aplicar o método de seleção *backward* com a estatística de Akaike - AIC em cada um deles, em 91% das vezes, a variável preditora P foi selecionada, portanto, o fósforo é importante atributo do solo para a predição da produtividade da soja. Além disso, observa-se que em 100% dos modelos em que o fósforo foi selecionado, o sinal de seu coeficiente foi negativo. Portanto, fica assegurado que, quando as demais variáveis são mantidas constantes, um aumento no nível de fósforo implica redução da produtividade de soja.

Outras variáveis selecionadas na maioria dos modelos foram à Des_2 com uma porcentagem de seleção de 87%, Ca com 81% e a RSP_1 com 79%. Assim, quando se analisam os sinais dos coeficientes associados às variáveis nos modelos em que elas foram selecionadas, destaca-se que, em 94% dos modelos em que a variável Ca foi selecionada, o sinal de seu coeficiente foi positivo. Logo, sugere-se que o aumento dos valores desta variável contribui para o aumento da produtividade de soja. Para os coeficientes associados às variáveis RSP_1 e Des_2, em 98% dos modelos em que elas foram selecionadas, os sinais foram negativos.

Tabela 2. Porcentagem de seleção de variáveis e porcentagem de sinais positivos e negativos dos parâmetros obtidos pelo método backward via critério de informação de Akaike (AIC) em 1000 modelos gerados por bootstrap.

Porcentagem de seleção		Sinais dos parâmetros estimados		
Variáveis [1]	**Pct**	**Parâmetros [2]**	**pct +**	**pct -**
P	91	β_P	0	100
Des_2	87	β_{Des_2}	2	98
Ca	81	β_{Ca}	94	6
RSP_1	79	β_{SRP_1}	2	98
Mn	75	β_{Mn}	6	94
Mg	71	β_{Mg}	9	91
Des_3	61	β_{Des_3}	80	20
Des_1	50	β_{Des_1}	36	64
K	48	β_K	84	16
RSP_3	46	β_{SRP_3}	50	50
RSP_2	42	β_{SRP_2}	78	22

[1] RSP_1, RSP_2 and RSP_3: resistência do solo à penetração, MPa, de 0 a 0,1 m, 0,1 a 0,2 m e 0,2 a 0,3 m de profundidade, respectivamente; Ca: cálcio, $cmol_c/dm^3$; Mg: magnésio, $cmol_c/dm^3$; K: potássio, mg/dm^3; P: fósforo, mg/dm^3; Mn: manganês, mg/dm^3. Des_1, Des_2 and Des_3: densidade do solo, g/cm^3, de 0 to 0,1 m, 0,1 a 0,2 m e 0,2 a 0,3 m de profundidade, respectivamente; [2] β_i : Parâmetro associado à variável i = {P, Des_2, Ca, Mn, Mg, Des_3, Des_1, K, RSP_3, RSP_2}; pct: porcentagem de seleção; pct+: porcentagem de sinais positivos; pct-: porcentagem de sinais negativos.

É evidente que algumas variáveis podem não ser úteis para bem explicar o comportamento da produtividade de soja, pois, por exemplo, dos 1000 modelos obtidos, a variável Des_1 foi selecionada em apenas 500. Além disto, em 180 destes modelos o sinal do coeficiente foi positivo e em 320 o sinal foi negativo. Assim, este conjunto de oscilações é uma garantia de que esta variável não é significativa. Portanto, pode ser excluída sem acarretar prejuízos na modelagem. Caso semelhante ocorre com a variável RSP_3, pois além de ser selecionada em apenas 460 modelos, não é possível inferir sobre qual o sinal adequado de seu coeficiente, tendo em vista que em 230 modelos ele é positivo e em 230 ele é negativo. Diante destas observações, foram definidos quatro modelos para serem analisados, a saber, M_{81}, M_{79}, M_{75} e M_{71} (Tabela 3). Cada um deles foi determinado segundo uma quantidade de variáveis explicativas, escolhidas de acordo com as porcentagens de vezes em que foram selecionadas nos modelos bootstrap.

Os regressores presentes no modelo M_{81} conseguem explicar apenas 37% da variação da produtividade de soja, resultado este inferior ao obtido quando considerado o modelo contendo todas as variáveis explicativas (R^2_{Adj} = 0,41). Os modelos M_{75} (R^2_{Adj} = 0,42) e M_{71} (R^2_{Adj} = 0,49) fornecem maior grau de explicação entre as variáveis explicativas e a produtividade de soja do que o modelo completo. Todavia, o modelo M_{79} (R^2_{Adj} = 0,41) fornece um grau de explicação equivalente, entretanto, esses modelos apresentam maior

RMSE quando comparados com ao modelo completo (RMSE = 0,33), diferença que é mais evidenciada no modelo M_{79} (RMSE = 0,39).

Tabela 3. Estimação dos Parâmetros e estatísticas para os modelos de regressão linear múltipla da produtividade de soja.

Modelos[2]	Parâmetros [1]							Estatísticas	
	Intercepto	β_P	β_{Des_2}	β_{Ca}	β_{SRP_1}	β_{Mn}	β_{Mg}	R^2_{Adj}	RMSE
M_{81}	7,827	-0,079	-1,994	0,099				0,37	0,41
M_{79}	8,482	-0,074	-2,185	0,103	-0,162			0,41	0,39
M_{75}	8,894	-0,067	-2,346	0,146	-0,179	-0,007		0,42	0,37
M_{71}	9,220	-0,076	-2,367	0,356	-0,221	-0,012	-0,479	0,49	0,34

[1] β_i : Parâmetro associado à variável i = {P, Des_2, Ca, RSP_1,Mn, Mg}; P: fósforo, mg/dm^3; Des_2: densidade do solo, g/cm^3, de 0,1 a 0,2 m de profundidade; Ca: cálcio, cmol$_c$/dm^3; RSP_1: resistência do solo à penetração, MPa, de 0 a 0,1 m de profundidade; Mn: manganês, mg/dm^3; Mg: magnésio, cmol$_c$/dm^3; [2] M_i: Modelo contendo as variáveis selecionadas em pelo menos i = {81, 79, 75, 71}% dos modelos bootstrap; R^2_{Adj}: coeficiente de determinação ajustado; RMSE: raiz quadrada do erro quadrático médio.

Como o modelo M_{71} explica 49% da variação da produtividade de soja e o RMSE deste modelo (RMSE = 0,34) está próximo do RMSE do modelo completo (RMSE = 0,33), optou-se por escolher o modelo M_{71} como o melhor modelo e realizar análises utilizando JaB para investigar a existência de pontos influentes. Destaca-se que, ao se considerar o valor 1 como ponto de corte, nenhum ponto é detectado como influente (Figura 2).

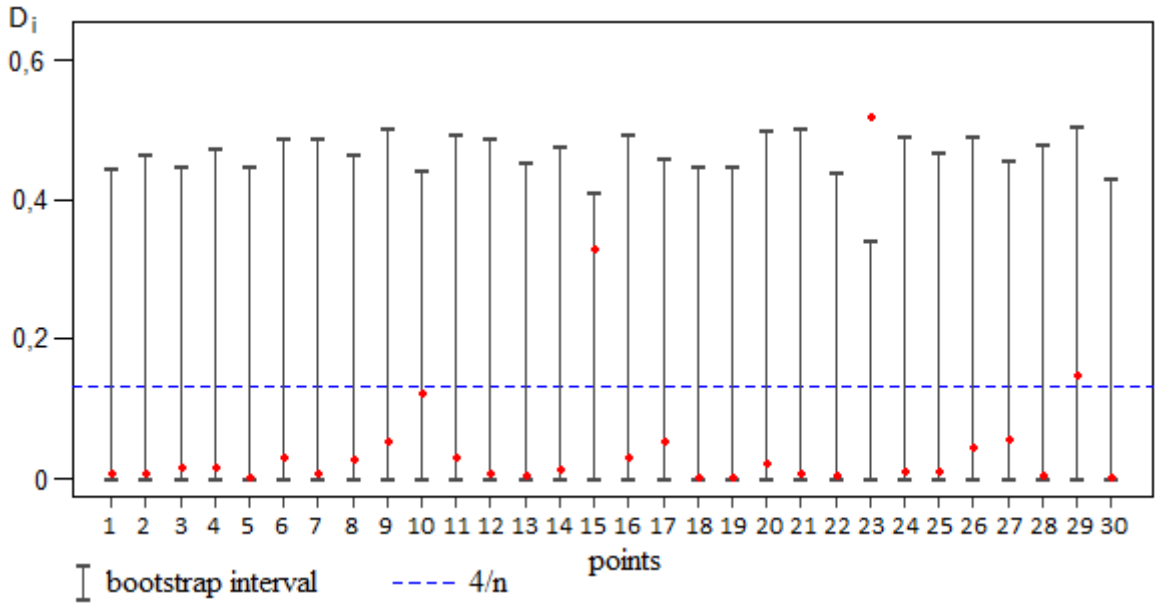

Figura 2. Determinação de pontos influentes de acordo com a distância de Cook (D_i) com a metodologia JaB.

O mesmo ocorre ao se considerar o critério que detecta o ponto i como influente se D_i é superior à mediana da distribuição F de Snedecor com graus de liberdade $p = 6$ e $n -$

$p = 24$, pois para tais valores o ponto de corte é 2,50 e assim, também não são detectados pontos influentes. Considerando-se $4/n \approx 0{,}13$ como ponto de corte, são detectados como influentes os pontos 15, 23 e 29, indicando que tais pontos podem mudar a estimação dos parâmetros no modelo de regressão, logo, é importante investigar o comportamento do modelo sem utilizar os pontos. Destaca-se que apenas o ponto 23 foi detectado como influente pela análise da distância de Cook (D_i) com a metodologia JaB.

Os gráficos JaB foram elaborados para auxiliar a identificação de pontos influentes, que permitem visualizar o quanto determinado ponto afeta a distribuição bootstrap das estimações dos parâmetros do modelo M_{71} (Figura 3).

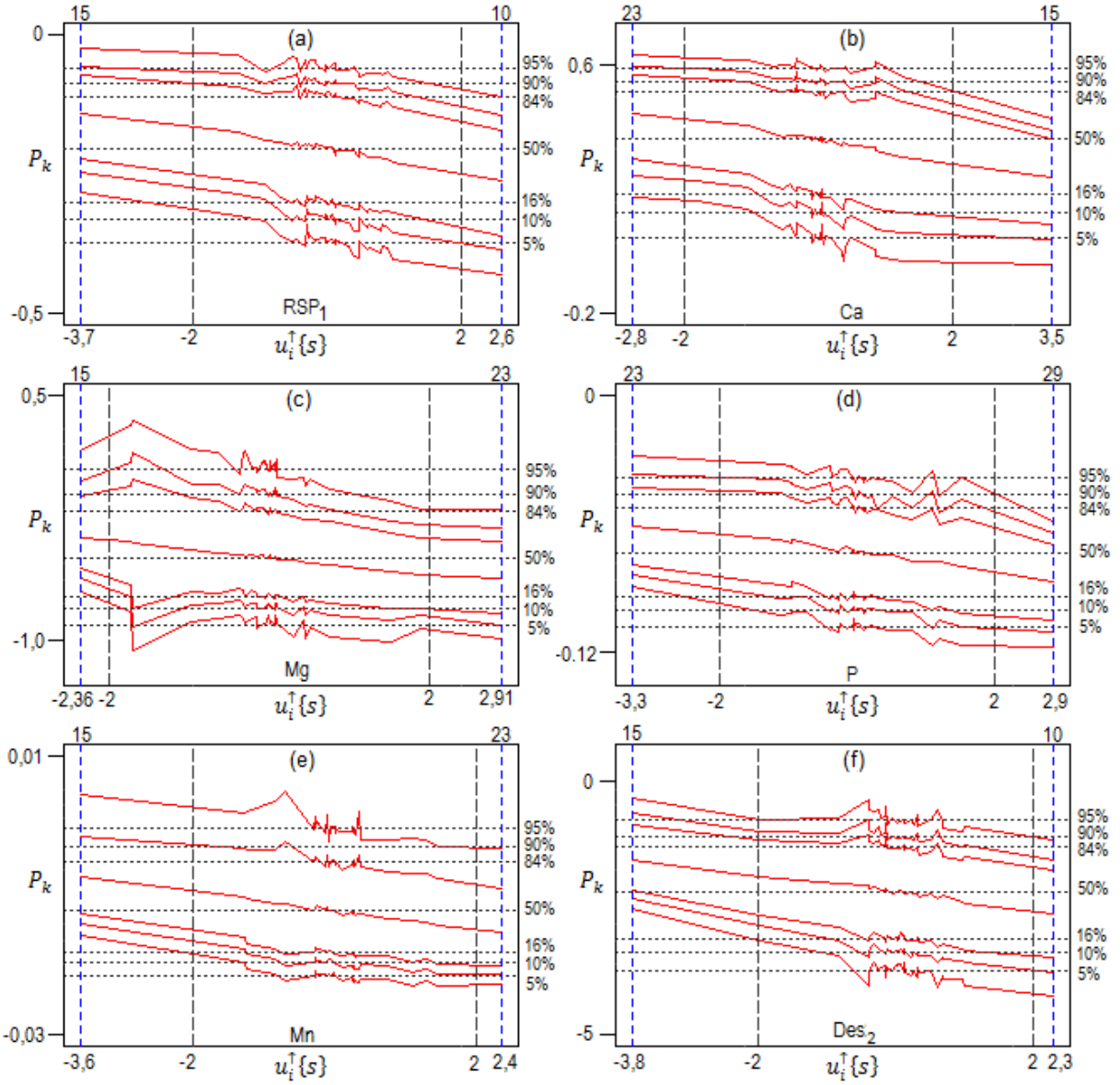

Figura 3. Gráficos JaB para as variáveis explicativas RSP_1 (a), Ca (b), Mg (c), P (d), Mn (e) e Des2 (f) do modelo M_{71}. $u_i^{\uparrow}\{s\}$: Valores da função de influência jackknife relativa em ordem crescente. P_k : *k*-ésimo percentil da distribuição bootstrap, k = {5,10,16,50,84,90,95}.

De acordo com os gráficos na Figura 3, observa-se que os pontos 10, 15, 23 e 29 foram detectados como influentes. Para mensurar o efeito dos pontos influentes na modelagem, foram ajustados dois novos modelos as variáveis P, Des_2, Ca, RSP_1, Mn, Mg. O modelo $M_{71\text{-}\{15,23,29\}}$ foi ajustado ao conjunto de dados sem os pontos (15, 23, 29), já que foram detectados como influentes pelo método da distância de Cook tradicional com ponto de corte $4/n$. Enquanto o $M_{71\text{-}\{10,15,23,29\}}$ foi ajustado ao conjunto de dados sem os pontos (10, 15, 23, 29), pois esses foram considerados influentes pelas análises utilizando JaB (Tabela 4). O modelo $M_{71\text{-}\{15,23,29\}}$, ajustado ao conjunto de dados sem os elementos amostrais 15, 23 e 29, que foram detectados como influentes pelo método tradicional, é mais explicativo do que o modelo M_{71} obtido do conjunto completo de pontos, pois com a retirada de tais , aumentou a porcentagem de variação da produtividade de soja de 49% para 63%, a qual consegue ser explicada pelos regressores.

Tabela 4. Estimação dos parâmetros e estatísticas dos modelos de regressão linear múltipla pela exclusão dos pontos influentes.

Modelos[2]	**Parâmetros** [1]							**Estatísticas**	
	Intercepto	β_P	β_{Des_2}	β_{Ca}	β_{RSP_1}	β_{Mn}	β_{Mg}	R^2_{Adj}	**RMSE**
$M_{71\text{-}\{15,23,29\}}$	7,453	-0,080	-1,335	0,405	-0,131	-0,013	-0,608	0,63	0,24
$M_{71\text{-}\{10,15,23,29\}}$	7,971	-0,080	-1,618	0,414	-0,169	-0,012	-0,635	0,65	0,23

[1] β_i : Parâmetro assiciado à variável i = {P, Des_2, Ca, RSP_1,Mn, Mg}; P: fósforo, mg/dm^3; Des_2: densidade do solo, g/cm^3, de 0,1 a 0,2 m de profundidade; Ca: cálcio, cmol$_c$/dm^3; RSP_1: resistência do solo à penetração, MPa, de 0 a 0,1 m de profundidade; Mn: manganês, mg/dm^3; Mg: magnésio, cmol$_c$/dm^3; [2] $M_{71\text{-}\{15,23,29\}}$: modelo ajustado ao conjunto de dados sem os pontos (15,23,29); $M_{71\text{-}\{10,15,23,29\}}$: modelo ajustado ao conjunto de dados sem os pontos (10,15,23,29); R^2_{Adj}: coeficiente de determinação ajustado; RMSE: raiz quadrada do erro quadrático médio.

Observa-se que o coeficiente de determinação ajustado (0,65) foi maior que o coeficiente de determinação ajustado obtido do modelo $M_{71\text{-}\{15,23,29\}}$, ao se considerar o modelo $M_{71\text{-}\{10,15,23,29\}}$, elaborado sem os pontos 10, 15, 23 e 29, o qual implica um modelo mais explicativo.

Diante destes resultados, optou-se por escolher o $M_{71\text{-}\{10,15,23,29\}}$ como o melhor modelo ajustado à produtividade de soja e determinar intervalos de confiança bootstrap para os parâmetros associados às variáveis explicativas (Tabela 5).

Observa-se que, ao serem comparados os intervalos de confiança dos parâmetros do modelo $M_{71\text{-}\{10,15,23,29\}}$ com os respectivos intervalos obtidos do modelo de regressão linear múltipla, gerado com todas as variáveis explicativas e com todos os pontos amostrais (Tabela 1), independente do método bootstrap utilizado, os intervalos de confiança dos parâmetros do modelo $M_{71\text{-}\{10,15,23,29\}}$ apresentam menor amplitude, indicando que as estimativas deste modelo são mais precisas.

Tabela 5. Intervalos de confiança bootstrap de 95% de confiança para os parâmetros do modelo $M_{71\text{-}\{10,15,23,29\}}$.

Parâmetros[1]	Percentil de Efron			BC[3]		
	$\hat{\theta}_l$	$\hat{\theta}_u$	Amplitude[2]	$\hat{\theta}_l$	$\hat{\theta}_u$	Amplitude
β_P	-0,119	-0,053	0,065	-0,122	-0,055	0,067
β_{Des_2}	-2,461	0,230	2,691	-2,580	-0,073	2,508
β_{Ca}	0,117	0,597	0,480	0,141	0,613	0,472
β_{RSP_1}	-0,326	-0,011	0,316	-0,319	0,016	0,335
β_{Mn}	-0,021	0,000	0,021	-0,021	0,000	0,021
β_{Mg}	-2,461	0,230	2,691	-2,580	-0,073	2,508
Intercepto	5,088	9,601	4,512	5,149	9,727	4,579

[1] β_i : Parâmetro associado à variável i = {P, Des_2, Ca, RSP_1,Mn, Mg}; P: fósforo, mg/dm^3; Des_2: densidade do solo, g/cm^3, de 0,1 a 0,2 m de profundidade; Ca: cálcio, cmol$_c$/dm^3; RSP_1: resistência do solo à penetração, MPa, de 0 a 0,1 m de profundidade; Mn: manganês, mg/dm^3; Mg: magnésio, cmol$_c$/dm^3; [2]Amplitude: $\hat{\theta}_u - \hat{\theta}_l$; $\hat{\theta}_l$: limite inferior; $\hat{\theta}_u$ **:**limite superior. [3]BC: bias corrected (viés corrigido).

5.1.4 Discussão

A média da produtividade de soja na área monitorada (4,305 t ha^{-1}) é considerada elevada quando comparada com outras regiõespois, segundo dados da Conab (2015) no ano agrícola 2013/2014, a média no Brasil foi de 2,854 t ha^{-1} e no Paraná foi de 2,950 t ha^{-1}.

O sinal negativo das estimativas dos parâmetros associados às variáveis RSP_1, RSP_3, Des_1 e Des_2 era esperado, pois a densidade do solo (Des) apresenta uma relação direta com a resistência do solo à penetração (RSP) (BUSSCHER *et al.*, 1997), e como a RSP exerce grande influência sobre o desenvolvimento vegetal, o crescimento das raízes e a produtividade das culturas variam de forma inversamente proporcional ao seu valor (FREDDI *et al.*, 2006). As estimativas dos parâmetros associados às variáveis RSP_2 e Des_3 apresentaram sinais opostos ao esperado, porém, como foi verificado que não existe multicolinearidade, é importante investigar a significância destas variáveis. O sinal positivo da estimação do parâmetro associado à variável K era esperado, pois conforme explica Pettigrew (2008), o potássio é um dos principais nutrientes, considerado essencial para o crescimento das culturas e rendimento da produtividade.

A comparação dos intervalos de confiança pode ser feita em termos de suas amplitudes e, de acordo com Paes (1998), um intervalo de amplitude elevada indica menor precisão da estimação quando comparado com um intervalo de menor amplitude. Desta forma, quando se comparam as duas técnicas de intervalos de confiança bootstrap (Tabela 1), destaca-se que, de maneira geral, os intervalos obtidos pela técnica percentil de Efron apresentaram menor amplitude, logo são os mais precisos. Diante do fato do zero estar contido na grande maioria dos intervalos de confiança (Tabela 1), é prudente investigar

se existem variáveis irrelevantes e/ou pontos influentes no conjunto de dados, pois esstes provocam aumento da variância dos parâmetros (RAO, 1971; MELOUN e MILITKÝ, 2001) e como resultado, os intervalos de confiança tendem a apresentar maior amplitude, por isto perdem a exatidão.

O fato da variável P ser selecionada em uma grande porcentagem de modelos (Tabela 2) e o sinal de seu coeficiente ser negativo em todos eles pode ser explicado pelos elevados valores de fósforo encontrados (em média 12 mg dm^{-3}). E, segundo Popp *et al*. (2002), os rendimentos podem diminuir indiretamente, devido aos desequilíbrios de micronutrientes. A elevada porcentagem de vezes (94%) em que o sinal do coeficiente associado à variável Ca foi positivo também era esperada, pois, conforme relatam Oliveira *et al*. (2009), a deficiência de cálcio está entre os principais fatores que inibem o crescimento radicular, principalmente em Latossolos. E, por sua vez, deixa a planta vulnerável a estresses bióticos, biológicos e nutricionais e consequentemente, leva à redução da produtividade (DOURADO NETO *et al.*, 2014). Como as variáveis RSP_1 e Des_2 são utilizadas para avaliar o estado de compactação do solo, é coerente que o efeito proporcionado por elas seja inverso à produtividade de soja, pois as plantas, em resposta à compactação do solo, apresentam alterações na profundidade, ramificação e distribuição das raízes (ROSOLEM *et al.*, 2002). Há, portanto, o comprometimento da eficiência do uso de nutrientes e de água e a limitação da produtividade da cultura (ALAKUKKU e ELOMEN, 1995).

O método de seleção de modelos a partir do bootstrap foi eficaz na determinação das variáveis significativas, cujo resultado foi um modelo mais parcimonioso e embora o modelo determinado por este método tenha sido o mesmo selecionado pelo método convencional utilizando o critério de Akaike, a aplicação desta metodologia serviu para atestar que o modelo selecionado pelo critério de Akaike não estava super-parametrizado, o que pode ocorrer quando a quantidade de amostras é pequena.

A análise do gráfico da Figura 3(a) referente à distribuição bootstrap das estimativas do parâmetro associado à variável RSP_1, aponta que os pontos 15 e 10 foram detectados como influentes. O ponto 15 apresenta influência negativa (-3,7), cuja remoção diminui a amplitude da distribuição bootstrap. Tal fato ocorre principalmente devido a um deslocamento nos percentis iniciais, quando se consideram tanto a distribuição empírica formada com as 3000 réplicas bootstrap, $\mathrm{P}_5 = -0{,}373$, $\mathrm{P}_{10} = -0{,}330$, $\mathrm{P}_{16} = -0{,}302$ como a distribuição empírica formada somente com as réplicas bootstrap provenientes das amostras bootstrap que não contêm o ponto 15 (1124 amostras), $\mathrm{P}_5 = -0{,}336$, $\mathrm{P}_{10} = -0{,}295$, $\mathrm{P}_{16} = -0{,}270$. A influência exercida pelo ponto 10 é positiva (2,6) e observa-se que, ao se considerar a distribuição bootstrap formada com as amostras bootstrap que não o contêm (1039 amostras), os valores dos percentis considerados diminuem, provocam deslocamento na distribuição e redução da amplitude e passam de 0,865 para 0,727.

O gráfico JaB da Figura 3(b) referente à variável Ca indicou o ponto 23 com influência negativa (-2,8) e o ponto 15 com influência positiva (3,5). Logo, a retirada desses pontos também provoca alterações na distribuição empírica das estimativas bootstrap. Ao desconsiderarem-se as réplicas bootstrap obtidas das amostras bootstrap que contêm o ponto 23, os percentis iniciais aumentam e a amplitude da distribuição passa de 1,214 para 0,999. E, ao serem desconsideradas as réplicas obtidas das amostras que contêm o ponto 15, destaca-se uma redução nos valores dos percentis finais, o que também reduz a amplitude da distribuição empírica. As análises dos demais gráficos (Figuras 3(c) à 3(f)) são semelhantes e indicam que o ponto 15 exerce influência negativa nas distribuições bootstrap dos parâmetros associados às variáveis Mg, Mn e Des_2 . O ponto 23 exerce influência positiva nas distribuições bootstrap dos parâmetros associados às variáveis Mg e Mn e tem influência negativa na distribuição bootstrap associada à variável P. O ponto 29 exerce influência positiva na distribuição bootstrap associada à variável P. Já o ponto 10 exerce influência positiva na distribuição bootstrap associada à variável Des_2.

Quando foram comparados todos os pontos detectados como influentes nos gráficos JaB (Figura 3), verifica-se que o elemento amostral 15 destaca-se por influenciar a maioria das distribuições bootstrap das estimativas dos parâmetros, pois, apenas a distribuição associada à variável P não é influenciada com a exclusão deste elemento. Os elementos amostrais 23 e 10 também destacam-se por exercerem influência em vários intervalos de confiança e o elemento amostral 29 configura-se como o menos influente dos quatro, pois é influenciada ao excluir da distribuição empírica as réplicas bootstrap obtidas das amostras bootstrap que o contêm apenas a distribuição bootstrap das estimativas do parâmetro associado à variável P, tendo sua amplitude reduzida de 0,163 para 0,112.

Em relação às análises de diagnósticos, observou-se que o método de determinação de pontos influentes utilizando a metodologia JaB com a distância de Cook (Figura 2) não identificou alguns pontos que claramente se destacaram como influentes pelos métodos tradicionais e pelos gráficos JaB. Deste modo, os gráficos JaB se mostraram uma ótima alternativa para identificação de pontos influentes, pois além de identificarem os pontos influentes com maior acurácia quando comparados com a análise tradicional, eles fornecem informações sobre as distribuições bootstrap das estimativas dos parâmetros. Portanto, permitem observar o que ocorre com os intervalos de confiança quando as amostras influentes são excluídas.

Em relação à significância das variáveis explicativas presentes no modelo $M_{71-\{10,15,23,29\}}$ observa-se que apenas a variável Mn apresentou intervalos de confiança contendo o zero em ambas as técnicas bootstrap (Tabela 5), dando indícios de que ela pode ser irrelevante. Esta suspeita pode ser descartada, pois temos evidências de que o sinal do parâmetro associado a esta variável é negativo. Basta observar que além do zero ter aparecido no limite superior dos intervalos, no método de seleção de variáveis (Tabela 2), a

variável Mn foi selecionada em 75% dos modelos (750 modelos) e teve sinal negativo em 94% deles (705 modelos). Ao serem comparadas as técnicas de determinação de intervalos de confiança utilizadas, observa-se que eles apresentaram comportamento semelhante, porém, destaca-se o método percentil de Efron por fornecer intervalos de menor amplitude. Cunha e Colosimo (2003) também destacaram o método percentil de Efron ao determinarem intervalos de confiança para modelos de regressão com erros de medida, pois segundo os autores, este método se evidenciou por ter maior simplicidade com igualdade de performance em relação aos demais.

Destaca-se que os métodos bootstrap foram fundamentais para a obtenção de um modelo mais explicativo e mais preciso, pois além do modelo $M_{71-\{10,15,23,29\}}$ fornecer maior porcentagem de explicação da produtividade de soja (65%) do que o modelo modelo inicial apresentado na Equação 3 (41%), ele fornece menor RMSE, indicando ser mais acurado. É importante salientar que o poder explicativo do modelo $M_{71-\{10,15,23,29\}}$ é satisfatório levando-se em consideração que ele é construído apenas com atributos físicos e químicos do solo. A porcentagem de variação da produtividade de soja não explicada por este modelo (35%) se deve às variáveis não consideradas, como as variáveis agrometeorológicas, pois o clima tem significativo impacto no crescimento e desenvolvimento das culturas (HOOGENBOOM, 2000). Salienta-se que a não inclusão de variáveis agrometeorológicas, neste estudo, se deve ao fato de que elas apresentam limitação quanto à representatividade espacial dos resultados por serem obtidas a partir de dados coletados em estações meteorológicas (JUNGES e FONTANA, 2011).

Os resultados mostraram que os métodos bootstrap possibilitaram selecionar os atributos físicos e químicos do solo, significativos na construção do modelo de regressão da produtividade da soja bem como construir os intervalos de confiança dos parâmetros em estudo e identificar os pontos que exerciam grande influência sobre a estimação dos parâmetros do modelo.

5.1.5 Referências

Aiken LS, West SG, 1991. Multiple regression: Testing and interpreting interactions. Sage Publications, Thousand Oaks, CA, USA. 224 pp.

Akaike H, 1973. Information theory and an extension of the maximum likelihood principle. Proc. 2nd Int. Symp. on Information Theory; Petrov BN, Csaki F (eds.). pp: 267–281. Akadémia Kiado, Budapest.

Alakukku L, Elomen P, 1995. Long-term effects of a single compaction by heavy field traffic on yield and nitrogen uptake of annual crops. Soil Till Res 36(3-4): 141-152.

Al-Marshadi AH, 2011. New weighted information criteria to select the true regression model. Aust J Basic Appl Sci 3(3): 317-312.

Austin P, Tu J, 2004. Bootstrap methods for developing predictive models. Am Stat 58(2): 131–137.

Beyaztas U, Alin A, 2013. Jackknife-after-bootstrap method for detection of influential observations in linear regression models. Commun Stat Simulat C 42(6): 1256-1267.

Busscher WJ, Bauer PJ, Camp CR, Sojka RE, 1997. Correction of cone index water content differences in a coastal plain soil. Soil Till Res 43(3-4): 205-217.

Chaves-Neto A, Faria, TMB, 2015. Bootstrap for indentification in Arma(p,q) structures. Ind J Manag Prod 6(1): 169-181.

Conab, 2015. Soja – Brasil: Série histórica de produtividade. http://www.conab.gov.br. [24 March 2015].

Cook RD, 1977. Detection of influential observation in linear regression. Technometrics 19(1): 15-18.

Cunha WJ, Colosimo EA, 2003. Intervalos de confiança bootstrap para modelos de regressão com erros de medida. Rev Mat Estat 21(2): 25-41.

Davison AC, Hinkley DV, 1997. Bootstrap methods and their application. Press syndicate of the University of Cambridge, Cambridge, UK. 582 pp.

Dourado Neto D, Dario GJA, Barbieri APP, Martin TN, 2014. Biostimulant action on agronomic efficiency of corn and common beans. Biosci J 30(1): 371-379.

Dubreuil S, Berveiller M, Petitjean F, Salaün M, 2014. Construction of bootstrap confidence intervals on sensitivity indices computed by polynomial chaos expansion. Reliab Eng Syst Safe 121: 263-275.

Efron B, 1979. Bootstrap methods: Another look at the jackknife. Ann Stat 7(1): 1-26.

Efron B, 1982. The jackknife, the bootstrap and other resampling plans. SIAM, Philadelphia, PA, USA. 93 pp.

Efron B, 1992. Jackknife-after-bootstrap standard errors and influence functions. J R Stat Soc 54: 83-127.

Efron B, Tibsshirani R, 1986. Bootstrap methods for standard errors, confidence intervals and other measures of statistical accuracy. Stat Sci 1(1): 54-75.

Embrapa, 2013. Sistema brasileiro de classificação de solos, 3ª Ed. – Centro Nacional de Pesquisa de Solos, EMBRAPA – SPI, Rio de Janeiro. 412 pp.

Freddi OS, Carvalho MP, Veronesi-Jr V, Carvalho GJ, 2006. Relationship between maize yield and soil mechanical resistance to penetration under conventional tillage. Eng Agric 26(1): 113-121.

Freedman DA, 1981. Bootstrapping regression models. Ann Statist 9(6): 1218-1228.

Freud RJ, Littell RC, 2000. SAS system for regression, SAS Inst., Cary, NC, USA. 264 pp.

García-Gallego JM, Chamorro-Mera A, García-Galán MM, 2015. The region-of-origin effect in the purchase of wine: The moderating role of familiarity. Span J Agric Res 13(3): e0103, 11 pages.

Garcia-Paredes JD, Olson KR, Lang JM, 2000. Predicting corn and soybean productivity for Illinois soils. Agric Syst 64(3): 151-170.

Hao L, Naiman DQ, 2010. Assessing inequality. Sage, Thousand Oaks, CA, USA. 149 pp.

Hoogenboom G, 2000. Contribution of agrometeorology to the simulation of crop production and its applications. Agric For Meteorol 103: 137-157.

Ireland CR, 2010. Experimental statistics for agriculture and horticulture. Cambridge University Press, Cambridge, UK. 384 pp.

Junges AH, Fontana DC, 2011. Agrometeorological-spectral model to estimate wheat yield in the state of Rio Grande do Sul, Brazil. Rev Ceres 58(1): 9-16.

Kamo K, Yanagihara H, Satoh K, 2013. Bias-corrected AIC for selecting variables in poisson regression models. Commun Stat A – Theory 42(11): 1911-1921.

Khakural BR, Robert PC, Huggins DR, 1999. Variability of corn/soybean yield and soil/landscape properties across a southwestern Minnesota landscape. In: Precision Agriculture; Robert PC, Rust RH, Larson WE (eds.). pp: 573-579. Am. Soc. Agron., Madison, WI, USA.

Kulcheski FR, Molina LG, Fonseca GC, Morais GL, Oliveira LFV, Margis R, 2016. Novel and conserved microRNAs in soybean floral whorls. Gene 575(2): 213-223.

Levy P, Lemeshow S, 1980. Sampling for health professionals. LLP, Belmont, CA, USA. 320 pp.

Lobell DB, Ortiz-Monasterio I, Asner GP, Naylor RL, Falcon WP, 2005. Combining field surveys, remote sensing, and regression trees to understand yield variations in an irrigated wheat landscape. Agron J 97: 241-249.

Losada B, Blas C, García-Rebollar P, Cachaldora P, Méndez J, Ibáñez M, 2015. Prediction of apparent metabolisable energy content of cereal grains and by-products for poultry from its chemical composition. Span J Agric Res 13(2): e06SC02, 5 pages.

Martin MA, Roberts S, 2010. Jackknife-after-bootstrap regression influence diagnostics. J Nonparametric Stat 22(2): 257-269.

Meloun M, Militký J, 2001. Detection of single influential points in OLS regression model building. Anal Chim Acta 439(2): 169-191.

Mercante E, Lamparelli RAC, Uribe-Opazo MA, Rocha JV, 2010. Linear regression models to soybean yield estimate in the west region of the state of Paraná, Brazil, using spectral data. Eng Agríc 30(3): 504-517.

Oliveira IP, Costa KAP, Faquin V, Maciel GA, Neves BP, Machado EL, 2009. Effects of calcium sources on Grass growth in monoculture and intercropping. Ciênc Agrotec 33: 592-598.

Paes AT, 1998. Essential items in biostatistics. Arq Bras Cardiol 71(4): 575-580.

Penalba OC, Bettolli ML, Vargas WM, 2007. The impact of climate variability on soybean yields in Argentina. Multivariate regression. Meteorol Appl 14: 3-14.

Peng RD, 2008. Simpleboot: Simple bootstrap routines. R package version 1.1-3.

Pettigrew WT, 2008. Potassium influences on yield and quality production for maize, wheath, soybean and cotton. Physiol Plant 133: 670-681.

Popp JS, Griffin TW, Popp MP, Baker WH, 2002. Profitability of variable rate phosphorus in a two crop rotation. J Ark cad Sci 56: 125-133.

R Core Team, 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Rahman MS, 2014. Coefficient estimation of regression model and hypothesis testing by bootstrap method. Res Rew J Stat 3(2): 1-7.

Rao P, 1971. Some notes on misspecification in multiple regressions. Am Statistician 25(5): 37-39.

Rizopoulos D, 2009. BootStepAIC: Bootstrap stepAIC. R package version 1.2-0.

Rosolem CA, Foloni JSS, Tiritan CS, 2002. Root growth and nutrient accumulation in cover crops as affected by soil compaction. Soil Till Res 65:109-115.

Sabaghnia N, Dehghani H, Alizadeh B, Mohghaddam M, 2010. Interrelationships between seed yield and 20 related traits of 49 canola (Brassica napus L.) genotypes in non-stressed and water-stressed environments. Span J Agric Res 8(2): 356-370.

Shasha D, Wilson M, 2011. Statistic is easy. Morgan & Claypool Publishers, San Rafael, CA, USA. 162 pp.

Siegel S, 1956. Nonparametric statistics for the behavioral sciences. McGraw-Hill, New York. 312 pp.

Sutton NJ, Cho S, Armsworth PR, 2016. A reliance on agricultural land values in conservation planning alters the spatial distribution of priorities and overestimates the acquisition costs of protected areas. Biol Cons 194: 2-10.

Tao F, Yokozawa M, Liu J, Zhang Z, 2008. Climate-crop yield relationships at provincial scales in China and the impacts of recent climate trends. Clim Res 38: 83-94.

Vera-Diaz MC, Kaufmann RK, Nepstad DC, Schlesinger P, 2008. An interdisciplinary model of soybean yield in the Amazon Basin: The climatic, edaphic, and economic determinants. Ecol Econ 65(2): 420-431.

Zheng H, Chen L, Han X, Zhao X, Ma Y, 2009. Classification and regression tree (CART) for analysis of soybean yield variability among fields in northeast China: The importance of phosphorus application rates under drought conditions. Agric Ecosyst Environ 132: 98-105.

5.2 ARTIGO 2: Quantificação da incerteza na modelagem geoestatística da produtividade de soja e atributos do solo utilizando bootstrap espacial

Resumo: Este trabalho teve como objetivo o estudo da dependência espacial de dados de produtividade de soja e atributos do solo de acordo com o método bootstrap na análise geoestatística. Utilizou-se o método bootstrap espacial para a quantificação das incertezas associadas à caracterização das estruturas de dependência espacial, aos estimadores dos parâmetros dos modelos ajustados, aos valores preditos por krigagem e ao pressuposto de normalidade multivariada dos dados. Os resultados obtidos possibilitaram quantificar as incertezas em todas as fases da análise geoestatística. Constatou-se que a inserção do bootstrap espacial na análise geoestatística é uma prática que pode ser adotada na agricultura de precisão, pois o melhor conhecimento dos atributos do solo permite a elaboração de mapas de aplicação de nutrientes mais precisos e proporciona a melhoria das produtividades das lavouras.

Palavras-chave: inferência estatística; quantile-quantile plot; reamostragem; variabilidade espacial.

5.2.1 Introdução

A geoestatística é amplamente utilizada nas ciências do solo, cuja utilização é comum em aplicações associadas à análise espacial de atributos do solo e produtividade das lavouras de soja. Dentre os trabalhos ligados à análise de atributos do solo, destacam-se o estudo da variabilidade espacial de indicadores de acidez (KRUEGER *et al.*, 2016), a utilização de geoestatística bayesiana para modelar o carbono orgânico (XIONG *et al.*, 2015) e a utilização de modelos da família Matérn para modelar o carbono orgânico, pH e a condutividade elétrica (MINASNY e MCBRATNEY, 2005). Em relação às produtividade das lavouras de soja, destaca-se o trabalho de Guedes *et al*. (2016), cujos autores comparam mapas temáticos da produtividade de soja para diferentes grades amostrais e o trabalho de Sobjak *et al*. (2016) em que os autores utilizam a krigagem ordinária para construir um mapa temático da produtividade de soja em um estudo de delineamento de zonas de manejo.

Em geral o objetivo de uma análise geoestatística concentra-se em torno da estimação de parâmetros e da predição de valores em locais não amostrados (DIGGLE e RIBEIRO JR., 2007). Como em geoestatística, geralmente, consideram-se poucos e esparsos elementos, existem incertezas associadas ao semivariograma experimental e aos estimadores dos parâmetros dos modelos ajustados (PARDO-IGÚZQUIZA e OLEA, 2012).

Devido à importância da modelagem da incerteza nas análises geoestatísticas, este tema vem ganhando destaque na literatura (SEBESTYÉN, 2004; SARI *et al.*, 2015). A

incerteza do semivariograma experimental é causada pela necessidade de se calcular a média ponderada dos valores de semivariância para intervalos de distância específica, definidos de maneira empírica, o que pode ocasionar uma suavização da estrutura do semivariograma (WEBSTER e OLIVER, 2007).

As incertezas associadas aos estimadores dos parâmetros do modelo geoestatístico surgem pelo fato desse ser estimado diretamente do semivariograma experimental utilizando métodos de mínimos quadrados ou métodos paramétricos, como os apresentados por Mardia e Marshall (1984), baseados em pressupostos de teoria assintótica. Como os parâmetros do modelo são utilizados na predição de valores em locais não amostrados, é evidente a existência de uma incerteza em relação aos valores preditos. Neste caso, estes serão de interesse apenas se combinados com medidas de sua acurácia (SCHELIN e SJÖSTEDT-DE, 2010). Diante da necessidade de que se quantifiquem as incertezas associadas aos resultados de uma análise geoestatística, uma alternativa aos tradicionais métodos de inferência é o uso do método bootstrap espacial proposto por Solow (1985), uma adaptação do método bootstrap (EFRON, 1979) para dados espacialmente dependentes.

O método bootstrap é bem conhecido e tem sido empregado em estudos relacionados às ciências do solo (ACKERSON *et al.*, 2015 e RODRIGUES JR. *et al.*, 2015; SUTTON *et al.*, 2016). O método bootstrap espacial é menos conhecido e dentre os trabalhos encontrados na literatura destacam-se os que apresentam simulações (TANG *et al.*, 2006; IRANPANAH *et al.*, 2011; GARCÍA-SOIDÁN *et al.*, 2014 e OLEA *et al.*, 2015)

Assim, o objetivo deste trabalho foi utilizar o método bootstrap espacial para quantificar as incertezas associadas à modelagem da dependência espacial da produtividade de soja e de atributos do solo em uma área agrícola.

5.2.2 Material e Métodos

5.2.2.1 Área de estudo e dados

O conjunto de dados foi coletado no ano agrícola 2014/2015 e provém de uma área agrícola de 79,09 hectares, localizada na região Oeste do Paraná, Brasil, próxima ao município de Cascavel, com coordenadas centrais latitude 24º57’25’’S e longitude 53º34’29’’W, e altitude média de 714 m (Figura 1). Segundo a classificação de Köppen, o clima da região da área agrícola é do tipo Cfa (APARECIDO et al., 2016) e o solo é classificado como Latossolo Vermelho distroférrico (EMBRAPA, 2009). Foi realizada uma amostragem sistemática centrada com pares de pontos próximos (*lattice plus close pairs*), composta de 78 elementos amostrais, georreferenciados com um aparelho GPS GEOEXPLORE 3 com precisão de 5 metros. Em cada ponto amostral determinou-se a

produtividade de soja (Prod, t/ha) e os atributos fósforo (P, mg/dm^3), potássio (K, mg/dm^3), matéria orgânica (MO, g/dm^3) e pH do solo (pH). Os pontos P_i, i = 1,...,7, distribuídos entre os elementos amostrais (Fig.1) indicam localizações selecionadas aleatoriamente em que serão estimados os valores dos atributos por interpolação geoestatística e determinados intervalos de confiança por bootstrap espacial.

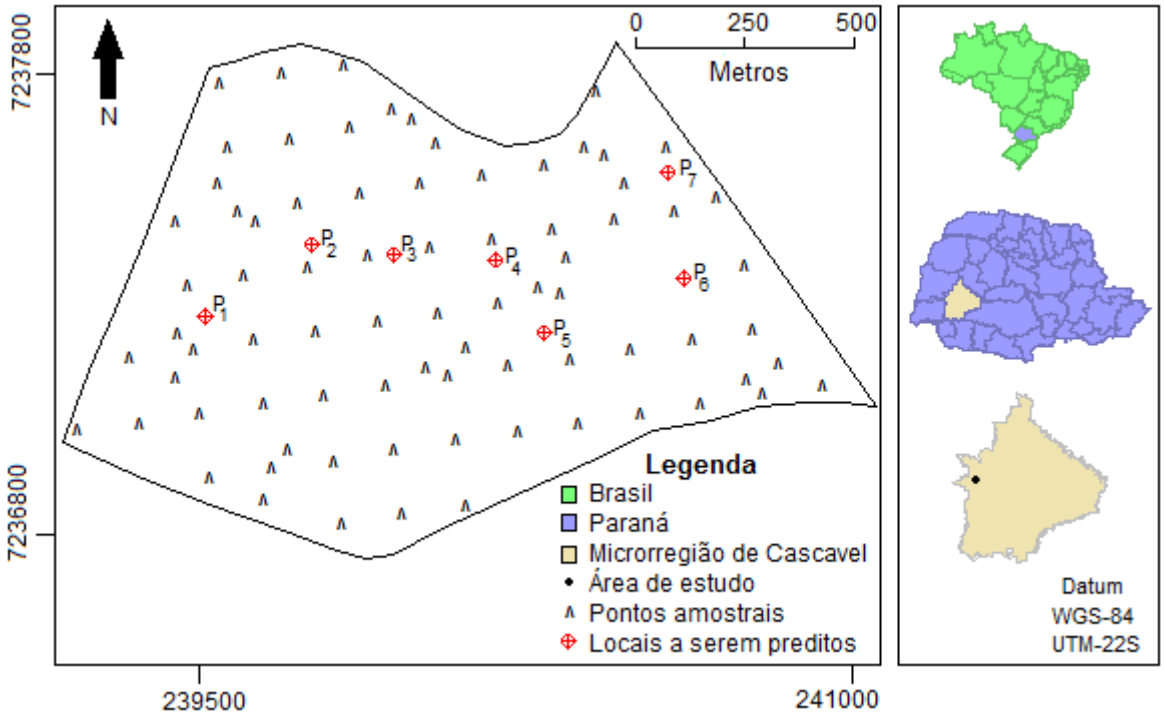

Figura 1 – Mapa da localização da área estudada.

5.2.2.2 Análise geoestatística

O modelo da estrutura de dependência espacial de uma variável regionalizada considerou um processo estocástico $\mathbf{Z} = \{Z(\mathbf{s}), \mathbf{s} \in S\}$, em que $\mathbf{s} = (x, y)^T$ é o vetor que representa uma localização na área em estudo $S \subset \Re^2$, sendo $\Re^2$ o espaço euclidiano bidimensional. Suponha-se que os dados do processo isotóprico e estacionário são registrados em localizações conhecidas e gerados pelo modelo $\mathbf{Z} = \mu 1 + \varepsilon$, em que μ representa um parâmetro desconhecido a ser estimado e ε representa um vetor de erros aleatórios $n \times 1$, com $E(\varepsilon) = 0$ e matriz de covariância Σ representada na forma paramétrica (MARDIA e MARSHAL 1984; URIBE-OPAZO *et al.*, 2012) por $\Sigma = \varphi_1 I_n + \varphi_2 R(\varphi_3)$ em que I_n é a matriz identidade $n \times n$, $\varphi_1 \geq 0$ é o efeito pepita, $\varphi_2 \geq 0$ é a contribuição, $\varphi_3 \geq 0$ é o parâmetro que define o alcance (a) do modelo e $R(\varphi_3) = [(r_{ij})]$ é uma matriz simétrica $n \times n$. Os elementos r_{ij}, $i, j = 1, \ldots, n$ representam a correlação entre os pontos s_i e s_j, sendo $r_{ij} = 1$ se $i = j$; $r_{ij} = 0$ se $i \neq j$ e $\varphi_2 = 0$ e $r_{ij} = \varphi_2^{-1}\sigma_{ij}$ se $i \neq j$ e $\varphi_2^{-1} \neq 0$, sendo $\sigma_{ij} = C(s_i, s_j) =$

$C(h_{ij})$ e $h_{ij} = \|s_i - s_j\|$ a distância euclidiana entre os pontos s_i e s_j (DE BASTIANI *et al*., 2015).

Foram construídos semivariogramas experimentais omnidirecionais para se identificar a estrutura de dependência espacial dos atributos, utilizando o estimador de Matheron; e, para modelar as estruturas de dependência espacial, utilizou-se o modelo da família Matérn (MATÉRN, 1986), apresentado na Equação [1]:

$$C(h_{ij}) = \begin{cases} \varphi_1 + \varphi_2 & , \quad i = j \\ \frac{\varphi_2}{2^{k-1}\Gamma(k)} \left(\frac{h_{ij}}{\varphi_3}\right)^k K_k\left(\frac{h_{ij}}{\varphi_3}\right), & i \neq j \end{cases}, \qquad [1]$$

em que $K_k(.)$ é a função de Bessel do terceiro tipo de ordem $k > 0$ e $\Gamma(.)$ é a função Gama.

Para processos Gaussianos estacionários de segunda ordem e isotrópicos, a função semivariância tem a relação $\gamma(h_{ij}) = C(0) - C(h_{ij})$ para $i, j = 1, \dots, n$ e $h_{ij} \geq 0$. Na Equação [1], o parâmetro de forma k controla o comportamento próximo à origem e à suavização analítica do processo. Neste trabalho, foram considerados os valores fixos $k = \{0{,}5; 1; 1{,}5; 2\}$ e $k \to \infty$. Quando $k = 0{,}5$, o modelo da família Matérn equivale ao modelo exponencial e quando $k \to \infty$, esse equivale ao modelo Gaussiano (DIGGLE e RIBEIRO JR., 2007). As estimações dos parâmetros dos modelos foram realizadas considerando os métodos dos mínimos quadrados ordinários (OLS), mínimos quadrados ponderados (WLS) e máxima verossimilhança (ML). A escolha dos melhores ajustes foi feita considerando as técnicas de validação cruzada (FARACO *et al.*, 2008) e as predições em locais não amostrados foram realizadas considerando a krigagem ordinária.

5.2.2.3 Bootstrap espacial

Seja θ um parâmetro de uma função de distribuição de probabilidade F e $\hat{\theta}$ a estatística estimada com base em uma amostra aleatória $\mathbf{x} = (x_1, \dots, x_n)^T$. O método bootstrap clássico depende de $\mathbf{x}^* = ({x_1}^*, \dots, {x_n}^*)^T$, uma amostra aleatória de tamanho n extraída com reposição de $\mathbf{x}$, conhecida como amostra bootstrap. Correspondendo a uma amostra bootstrap $\mathbf{x}^*$ tem-se a réplica bootstrap de $\hat{\theta}$, denotada por $\hat{\theta}^*$. O algoritmo bootstrap clássico consiste na obtenção de B amostras bootstrap independentes $x^{*1}, \dots, x^{*B}$ e em calcular as réplicas bootstrap correspondentes para construir a distribuição de probabilidade empírica $\hat{F}$, que é utilizada nas inferências estatísticas (EFRON e TIBSHIRANI, 1993).

Como a reamostragem proporcionada pelo bootstrap clássico é válida somente quando as observações são independentes e identicamente distribuídas, utilizou-se, neste trabalho, o método bootstrap espacial, proposto por Solow (1985), apresentado no Algoritmo 1, que permite obter réplicas bootstrap espacialmente correlacionadas.

Algoritmo 1: Bootstrap espacial.

a) Considerando-se o conjunto de dados espaciais $\{Z(s_1),\ldots,Z(s_n)\}$, determine o vetor dos resíduos $\hat{\boldsymbol{\varepsilon}} = (Z(s_1) - \hat{\mu}, \ldots, Z(s_n) - \hat{\mu})^T$ sendo $\hat{\mu} = \left(\mathbf{1}^T\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{1}\right)^{-1}\mathbf{1}^T\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{Z}$ o estimador de mínimos quadrados de μ e $\hat{\boldsymbol{\Sigma}}$ a matriz de covariância estimada. b) Considerando-se a matriz de covariância estimada $\hat{\Sigma}$, utilize o método de decomposição de Cholesky para obter $\hat{\boldsymbol{\Sigma}} = \hat{\mathbf{L}}\hat{\mathbf{L}}^T$, em que $\hat{L}$ é uma matriz triangular inferior de ordem n; c) Utilizando-se a matriz $\hat{L}^{-1}$, determine $\hat{\boldsymbol{\varepsilon}}_{dec} = \hat{\mathbf{L}}^{-1}\hat{\varepsilon}$, que é o vetor de resíduos descorrelacionados e centralize seus valores, em que se obtém $\tilde{\mathbf{e}} = \hat{\boldsymbol{\varepsilon}}_{dec} - \left(\frac{1}{n}\right)\sum\hat{\boldsymbol{\varepsilon}}_{dec}$; d) Considerando-se o conjunto dos resíduos descorrelacionados e centralizados $\tilde{e}_1, \ldots, \tilde{e}_n$, realize uma reamostragem com reposição, em que se obtém o vetor $\boldsymbol{\varepsilon}^*_{SB} = (\varepsilon^*_1, \ldots, \varepsilon^*_n)^T$; e) Cada amostra bootstrap espacial é obtida recorrelacionando-se os resíduos bootstrap $\mathbf{Z}^* = \hat{\mu} + \hat{\mathbf{L}}\boldsymbol{\varepsilon}^*_{SB}$.

5.2.2.4 Quantificação das incertezas na análise geoestatística

Foram utilizados os estimadores de μ e $\boldsymbol{\Sigma}$ de cada variável regionalizada para quantificar as incertezas associadas à análise geoestatística, e o Algoritmo 1 para determinar B = 1000 amostras bootstrap do conjunto de dados. Para cada amostra construiu-se um semivariograma experimental, ajustou-se um modelo e estimaram-se os valores de cada variável regionalizada nos pontos P_i, i = 1,...,7. Este procedimento permitiu construir a distribuição empírica das semivariâncias, dos parâmetros dos modelos e dos valores preditos, e consequentemente, determinar intervalos de confiança utilizando o método percentil de Efron (EFRON, 1982).

Calculou-se a distância de Mahalanobis para as B = 1000 amostras bootstrap para verificação da suposição de normalidade multivariada e, desta forma, estabelecer a proximidade entre cada observação e o centro da distribuição. Como as distâncias de Mahalanobis são independentes e apresentam distribuição assintóticamente qui-quadrado com p graus de liberdade, em que p é o número de variáveis (MARDIA, 1977), utilizou-se a aproximação de Wilson e Hilferty (1931) para transformá-las em escores z. Os valores resultantes foram ordenados e plotados versus os valores esperados das estatísticas de ordem normal, formando gráficos para obter os QQ plots.

A implementação computacional realizada neste trabalho foi desenvolvida no software R (R CORE TEAM, 2016).

5.2.3 Resultados e discussão

5.2.3.1 Estimação de parâmetros

Os valores estimados dos parâmetros dos modelos geoestatísticos da produtividade de soja e dos atributos do solo foram selecionados por validação cruzada e estão apresentados na Tabela 1. Os modelos obtidos pelos métodos OLS e WLS, associados à produtividade de soja (Prod) e aos atributos fósforo (P), potássio (K) e pH, apresentaram efeito pepita $\varphi_1 = 0$. Isso indica que o erro experimental foi o mínimo possível atingido e não existe variação a distâncias menores que as amostradas. Os estimadores do efeito pepita (φ_1) (Tabela 1) dos modelos associados à matéria orgânica (MO) foram diferentes de zero para os métodos de estimação utilizados. Isso indica uma variabilidade não explicada pelos modelos. Destaca-se que, para todas as variáveis, as estimativas da média (μ) apresentaram pouca diferença entre os métodos de estimação.

Tabela 1 -Estimação dos parâmetros dos modelos geoestatísticos.

Métodos	Variáveis	Modelo	Parâmetros estimados				
			$\hat{\mu}$	$\hat{\varphi}_1$	$\hat{\varphi}_2$	$\hat{\varphi}_3$	$\hat{a}$
OLS	Prod	$M_{0.5}$	2,368	0,0000	0,0748	68,3469	204,750
	P	$M_{0.5}$	19,142	0,0000	121,7703	43,7294	131,002
	K	$M_{0.5}$	0,314	0,0000	0,0205	15,1318	45,331
	MO	M_{inf}	50,807	13,9868	26,1408	117,4674	203,314
	pH	$M_{0.5}$	4,837	0,0000	0,1541	71,8227	215,162
WLS	Prod	$M_{0.5}$	2,367	0,0000	0,0749	76,8157	230,119
	P	$M_{0.5}$	19,160	0,0000	121,0163	37,0239	110,914
	K	$M_{1.5}$	0,314	0,0000	0,0206	12,9184	61,283
	MO	$M_{0.5}$	51,084	16,9464	25,4542	150,8846	452,010
	pH	$M_{0.5}$	4,831	0,0000	0,1523	60,2970	180,634
ML	Prod	M_{inf}	2,372	0,0558	0,0179	171,2865	296,466
	P	$M_{0.5}$	19,162	0,0000	121,2936	36,0720	108,062
	K	$M_{2.0}$	0,313	0,0181	0,0020	106,0676	569,410
	MO	$M_{2.0}$	51,375	28,6536	13,2079	156,2700	838,919
	pH	$M_{0.5}$	4,822	0,1237	0,0396	58,3517	174,806

OLS: Mínimos quadrados ordinários, WLS: Mínimos quadrados ponderados, ML: Máxima verossimilhança, Prod: Produtividade de soja (t/ha), P: Fósforo (mg/dm^3), K: Potássio (mg/dm^3), MO: Matéria orgânica (g/dm^3), pH: pH do solo, $\hat{\mu}$: Estimativa da média do modelo, $\hat{\varphi}_1$: Efeito pepita, $\hat{\varphi}_2$:contribuição, $\hat{\varphi}_3$: parâmetro de alcance, $\hat{a}$: alcance (m), M_k: Modelo Matérn com parâmetro de forma k ={0,5; 1; 1,5; 2}, M_{inf}: Modelo Gaussiano.

Na modelagem da estrutura de dependência espacial do fósforo (P) (Tabela 1) ,obteve-se por validação cruzada que o modelo Matérn com parâmetro de forma k = 0,5 ($M_{0.5}$) foi o que melhor se ajustou com todos os métodos utilizados. Vale destacar que os métodos de estimação proporcionaram as estimativas dos parâmetros semelhantes. Fato diferente ocorreu com a modelagem da estrutura de dependência espacial do potássio (K), pois além dos modelos selecionados não apresentarem o mesmo parâmetro de forma, há

uma diferença considerável entre os valores estimados dos parâmetros que definem a estrutura de dependência espacial.

Para o pH, a validação cruzada indicou o modelo exponencial ($M_{0.5}$) como melhor ajuste nos três métodos de ajuste utilizados. Existe uma incerteza associada ao estimador do efeito pepita deste modelo pois, enquanto os modelos ajustados por OLS e WLS não apresentam efeito pepita, o modelo ajustado por ML apresentou um efeito pepita elevado, ou seja, estimativa do efeito pepita maior que $\hat{\varphi}_2$.

5.2.3.2 Intervalos de confiança bootstrap para a semivariâncias

Foram determinados os intervalos bootstrap com 95% de confiança para as semivariâncias da produtividade de soja e dos atributos do solo (Figura 2 - valores alocados acima das semivariâncias indicam a quantidade de pares utilizados no cálculo). Verifica-se uma proximidade entre seus limites quando se comparam os gráficos dos intervalos de confiança bootstrap do semivariograma da produtividade de soja (Figura 2-a). No entanto, a menor amplitude do intervalo obtido pelo modelo gaussiano (M_{inf}) ajustado por ML indica que as réplicas bootstrap, determinadas por este modelo foram mais precisas na caracterização da estrutura de dependência espacial. Os intervalos de confiança bootstrap do semivariograma do fósforo (P) (Figura 2-b) e do potássio (K) (Figura 2-c) foram os que apresentaram maior proximidade entre seus limites pelos métodos de estimação.

Para o P, esta similaridade pode ter ocorrido em função da escolha do mesmo modelo exponencial ($M_{0.5}$) (Tabela 1) pelos métodos de estimação. Portanto, verifica-se um comportamento constante das semivariâncias ao longo das distâncias, indicativo de possível ausência de correlação espacial na escala em que a variável foi mensurada (JOURNEL e HUIJBREGTS, 1978).

Os intervalos de confiança bootstrap para o semivariograma da matéria orgânica (MO) (Figura 2-d) destacam-se por apresentarem uma diferença considerável entre os limites superiores dos intervalos. Isto ocorreu porque as estruturas de dependência espacial da MO modeladas com os três métodos de estimação foram diferentes. Como estas estruturas são utilizadas para gerar as réplicas bootstrap, é coerente que os intervalos apresentem diferenças. Já para o pH do solo (Figura 2e), os intervalos de confiança construídos com réplicas bootstrap, geradas pelos modelos ajustados por OLS e WLS, apresentaram amplitudes semelhantes e superiores às amplitudes dos intervalos construídos com o modelo ajustado por ML. É importante destacar que os intervalos de confiança bootstrap referentes às variáveis P, K e pH apresentam comportamento aparentemente horizontal.

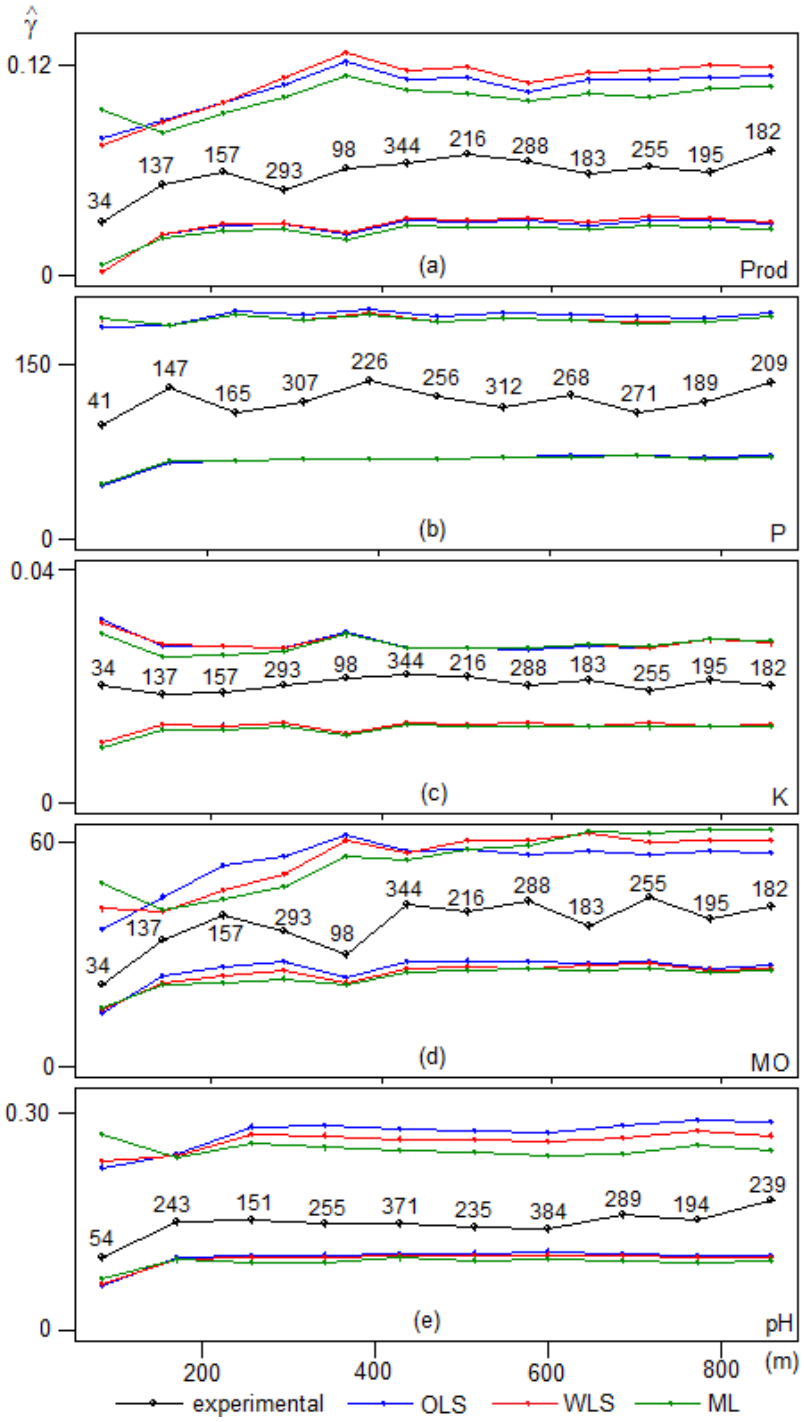

Figura 2 – Intervalos bootstrap com 95% de confiança para os semivariogramas experimentais: (a) Prod: Produtividade de soja (t/ha), (b) P: Fósforo (mg/dm^3), (c) K: Potássio (mg/dm^3), (d) MO: Matéria orgânica (g/dm^3) e (e) pH: pH do solo.

Neste sentido, é prudente quantificar a incerteza associada à modelagem da dependência espacial destas variáveis pois, em situações nas quais o semivariograma experimental tem comportamento horizontal, vários modelos podem ser ajustados (DRUBULE, 1993), e variam a partir de modelos próximos a um efeito pepita puro (fato este observado no ajuste do modelo $M_{2.0}$ obtido por ML aos dados de potássio) até modelos com

alcance próximo à distância mínima entre pontos (fato este observado no ajuste do modelo $M_{0.5}$ obtido por OLS aos dados de potássio).

5.2.3.3 Análise descritiva e intervalos de 95% de confiança para os parâmetros dos modelos

Foram construídas distribuições empíricas e calculadas as estatísticas descritivas e intervalos bootstrap com 95% de confiança para os parâmetros dos modelos relacionados à produtividade de soja (Tabela 2) e atributos do solo (Tabela 3).

Tabela 2 – Estatísticas e intervalos de 95% de confiança percentil de Efron da distribuição bootstrap dos estimadores dos parâmetros dos modelos da estrutura de dependência espacial da produtividade de soja (Prod).

Método/ Modelo	E%	φ	$\hat{\varphi}$	Min	Q_1	Mediana	Média	Q_3	Max	Ep	As	Li	Ls
OLS/ $M_{0.5}$	0,5	φ_1	0,00	0,00	0,00	0,00	0,00	0,00	0,10	0,00	23,60	0,00	0,00
		φ_2	0,07	0,00	0,07	0,08	0,08	0,09	0,14	0,01	0,26	0,05	0,11
		φ_3	68,35	4,78	46,66	65,83	66,27	83,45	392,45	33,50	2,16	7,16	134,60
WLS/ $M_{0.5}$	0,8	φ_1	0,00	0,00	0,00	0,00	0,00	0,00	0,10	0,01	9,65	0,00	0,00
		φ_2	0,07	0,00	0,07	0,08	0,08	0,09	0,16	0,02	-0,41	0,05	0,11
		φ_3	76,82	4,01	55,54	73,43	79,19	99,12	285,17	37,11	1,16	20,00	166,05
ML/ M_{inf}	0,3	φ_1	0,06	0,00	0,02	0,04	0,04	0,06	0,10	0,02	-0,26	0,00	0,08
		φ_2	0,02	0,00	0,02	0,03	0,04	0,06	0,11	0,02	0,51	0,00	0,09
		φ_3	171,29	0,00	52,80	106,00	123,00	167,00	759,00	101,58	2,08	11,73	391,14

φ:parâmetros, φ_1:efeito pepita, φ_2:contribuição, φ_3: parâmetro do alcance, $\hat{\varphi}$:estimadores, Min: mínimo, Q_1: primeiro quartil, Q_3: terceiro quartil, Max: máximo, Ep: erro padrão, As: coeficiente de assimetria, Li: limite inferior do intervalo de confiança, Ls: limite superior do intervalo de confiança, E%: porcentagem de modelos que foram excluídos, OLS: Mínimos quadrados ordinários, WLS: Mínimos quadrados ponderados, ML: Máxima verossimilhança, $M_{0.5}$: Modelo Matérn com parâmetro de forma k =0.5, M_{inf}: Modelo Gaussiano.

Destaca-se que dos B=1000 modelos exponenciais ($M_{0.5}$) ajustados por OLS (Tabela 2), cinco modelos (E = 0,5%) foram excluídos da análise por apresentarem alcance superior à distância máxima entre amostras (1718 m). Desta forma, as distribuições bootstrap dos estimadores dos parâmetros deste modelo foram elaboradas com 995 réplicas bootstrap. O mesmo ocorreu com os modelos exponenciais ($M_{0.5}$) ajustados por WLS com oito modelos excluídos (E = 0,8%). Dos modelos ajustados por ML (Tabela 2), três foram excluídos devido à ocorrência de indeterminações numéricas, ocasionadas em função do modelo gaussiano (M_{inf}) ser suave na origem e isso gera linhas/colunas não distinguíveis numericamente na matriz de correlação.

Quando foram analisadas as estatísticas descritivas das réplicas bootstrap do efeito pepita nos modelos exponenciais ($M_{0.5}$) ajustados por OLS e WLS (Tabela 2), destaca-se que a maioria dos modelos apresentou efeito pepita nulo. Logo, ocasiona-se uma assimetria positiva nas distribuições bootstrap e, consequentemente, a redução dos intervalos de confiança a um único ponto. Diferente do ocorrido com os ajustes por OLS e WLS, tem-se

que o efeito pepita do modelo ajustado por ML não é nulo, pois o estimador bootstrap foi diferente de zero na maioria dos modelos.

Tabela 3 – Estatísticas e intervalos de 95% de confiança percentil de Efron da distribuição bootstrap dos estimadores dos parâmetros dos modelos da estrutura de variabilidade espacial dos atributos do solo.

Atributo	Modelo/ Método	E%	φ	$\hat{\varphi}$	Min	Q_1	Mediana	Média	Q_3	Max	Ep	As	Li	Ls
P	OLS/ $M_{0.5}$	7,7	φ_1	0,00	0,00	0,00	0,00	12,10	2,11	157,00	28,61	2,60	0,00	104,84
			φ_2	121,77	7,56	90,50	114,00	112,00	136,00	235,00	37,87	-0,31	25,02	179,61
			φ_3	43,73	3,54	4,87	51,20	58,40	76,40	512,00	60,56	2,90	3,65	215,10
	WLS/ $M_{0.5}$	8,5	φ_1	0,00	0,00	0,00	0,00	10,90	3,73	154,00	25,03	2,85	0,00	95,21
			φ_2	121,02	10,70	91,70	114,00	111,00	134,00	227,00	34,05	-0,30	33,91	171,93
			φ_3	37,02	0,00	5,27	36,80	45,80	58,00	567,00	55,85	3,75	3,82	193,65
	ML/ $M_{0.5}$	0,1	φ_1	0,00	0,00	0,00	0,00	32,85	69,76	180,10	46,60	1,07	0,00	134,68
			φ_2	121,29	0,00	34,74	98,18	86,34	126,90	215,50	53,08	-0,21	0,02	174,59
			φ_3	36,07	0,00	24,21	38,62	49,99	57,83	537,50	51,98	3,86	0,66	192,90
K	OLS/ $M_{0.5}$	0	φ_1	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	21,76	0,00	0,00
			φ_2	0,02	0,00	0,02	0,02	0,02	0,02	0,03	0,00	-0,54	0,01	0,03
			φ_3	15,13	5,21	6,64	15,00	24,00	38,30	131,00	20,03	0,98	5,66	69,75
	WLS/ $M_{1.5}$	0	φ_1	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	7,50	0,00	0,00
			φ_2	0,02	0,01	0,02	0,02	0,02	0,02	0,03	0,00	0,19	0,02	0,03
			φ_3	12,92	3,60	4,74	15,00	15,10	22,40	58,30	10,36	0,67	3,81	36,42
	ML/ $M_{2.0}$	0,6	φ_1	0,02	0,00	0,00	0,01	0,01	0,02	0,03	0,01	-0,45	0,00	0,02
			φ_2	0,00	0,00	0,00	0,00	0,01	0,01	0,03	0,01	0,74	0,00	0,02
			φ_3	106,07	0,00	14,53	33,87	52,59	74,02	307,40	55,18	1,64	0,00	206,60
M.O	OLS/ M_{inf}	0,7	φ_1	13,99	0,00	4,48	13,70	13,70	21,60	41,00	10,17	0,18	0,00	32,46
			φ_2	26,14	4,58	18,40	28,00	27,20	35,40	60,30	10,69	-0,01	7,93	46,12
			φ_3	117,47	15,90	93,20	121,00	144,00	157,00	980,00	103,70	3,89	47,79	453,70
	WLS/ $M_{0.5}$	18,6	φ_1	16,95	0,00	0,00	6,85	8,91	16,30	37,00	9,22	0,65	0,00	28,58
			φ_2	25,45	0,00	24,10	33,00	32,70	40,80	77,70	12,10	0,15	9,85	58,00
			φ_3	150,88	0,00	68,30	108,00	138,00	170,00	569,00	107,70	1,59	4,20	456,20
	ML/ $M_{2.0}$	3,4	φ_1	28,65	0,00	17,06	24,32	21,84	28,87	46,79	10,27	-0,81	0,00	36,83
			φ_2	13,21	0,00	8,65	14,69	16,74	23,65	53,50	10,49	0,68	1,52	40,93
			φ_3	156,27	0,00	46,72	92,92	102,00	141,50	315,30	67,40	0,76	4,14	262,60
pH	OLS/ $M_{0.5}$	3,5	φ_1	0,00	0,00	0,00	0,00	0,03	0,04	0,24	0,05	1,77	0,00	0,18
			φ_2	0,15	0,00	0,12	0,17	0,16	0,20	0,34	0,07	-0,68	0,00	0,26
			φ_3	71,82	5,20	52,20	76,30	82,20	90,00	571,00	65,92	4,13	6,81	287,50
	WLS/ $M_{0.5}$	5,1	φ_1	0,00	0,00	0,00	0,00	0,04	0,10	0,26	0,07	1,29	0,00	0,19
			φ_2	0,15	0,00	0,10	0,15	0,14	0,19	0,31	0,08	-0,53	0,00	0,25
			φ_3	60,30	0,00	47,60	67,90	70,30	90,00	544,00	46,29	4,29	19,99	145,80
	ML/ $M_{0.5}$	0,5	φ_1	0,12	0,00	0,00	0,07	0,07	0,14	0,24	0,07	0,27	0,00	0,20
			φ_2	0,04	0,00	0,01	0,08	0,09	0,15	0,26	0,07	0,25	0,00	0,22
			φ_3	58,35	0,00	3,85	30,00	44,40	57,70	572,00	57,18	3,60	0,00	200,27

φ:parâmetros, φ_1:efeito pepita, φ_2:contribuição, φ_3: parâmetro do alcance, $\hat{\varphi}$: estimadores, Min: mínimo, Q_1: primeiro quartil, Q_2: segundo quartil (mediana), Q3: terceiro quartil, Max: máximo, Ep: erro padrão, As: coeficiente de assimetria, Li: limite inferior do intervalo de confiança, Ls: limite superior do intervalo de confiança, E%: porcentagem de modelos que foram excluídos, P: Fósforo (mg/dm^3), K: Potássio (mg/dm^3), MO: Matéria orgânica (g/dm^3), pH: pH do solo), M_k: Modelo Matern com parâmetro de forma k ={0.5, 1, 1.5, 2}, M_{inf}: Modelo Gaussiano.

As medidas de tendência central dos estimadores bootstrap do parâmetro de alcance dos modelos de dependência espacial da produtividade de soja ajustados por OLS e WLS (Tabela 2) estão próximas dos valores obtidos pelo ajuste dos modelos ao conjunto de dados original (Tabela 1). Isso indica que os estimadores do parâmetro de alcance destes

modelos estão bem definidos. Como 75% das réplicas bootstrap do estimador do parâmetro de alcance nos modelos de variabilidade espacial da produtividade de soja ajustados por ML foram inferiores ao valor do estimador obtido pela amostra original (171,29 m), a indicação de que este valor pode estar sendo superestimado.

Destaca-se a ocorrência de efeito pepita nulo em 50% dos modelos de dependência espacial do fósforo (Tabela 3) obtidos por bootstrap. As estimativas da contribuição dos modelos de dependência espacial do fósforo são semelhantes, entretanto, pode-se afirmar que o estimador da contribuição do modelo exponencial ($M_{0.5}$) obtido por WLS é o mais acurado, tendo em vista a menor amplitude de seu intervalo de confiança. Como os modelos de dependência espacial do fósforo (P) foram os mesmos ($M_{0.5}$), os parâmetros de alcance destes modelos são comparáveis. Neste sentido, destaca-se o parâmetro de alcance do modelo ajustado por ML, tendo em vista que suas réplicas bootstrap apresentaram um menor erro padrão.

A mediana das réplicas bootstrap dos estimadores dos parâmetros de alcance associados aos modelos de dependência espacial do potássio (K) gerados por OLS (Tabela 3) indica que 50% dos modelos apresentaram alcance prático inferior à distância mínima entre pontos amostrais. Esta quantidade é menos evidenciada nos ajustes realizados por WLS, pois a porcentagem de modelos com alcance prático inferior à distância mínima foi de 29,7%. Destaca-se que o estimador do parâmetro de alcance do modelo de dependência espacial do potássio gerado por WLS é mais acurado do que o estimador do modelo gerado por OLS, pois apresentou um intervalo de confiança de menor amplitude.

O estimador do parâmetro de alcance do modelo de dependência espacial do potássio obtido por ML (Tabela 3) é considerado elevado tendo em vista que ele é superior ao terceiro quartil da distribuição das réplicas bootstrap. Assim como ocorreu no modelo Matérn com parâmetro de forma k = 2 ($M_{2.0}$) do potássio obtido por ML ajustado à amostra original, os modelos obtidos pelas amostras bootstrap também apresentaram valores do efeito pepita aproximadamente iguais ao patamar. E, segundo Jensen et al. (2000), é caracterizado como efeito pepita puro e indica a inexistência de autocorrelação espacial dos dados.

Quando se comparam as porcentagens de exclusão dos modelos de dependência espacial da matéria orgânica (MO) (Tabela 3), destacam-se os modelos exponenciais ($M_{0.5}$) obtidos por WLS, tendo em vista a elevada quantidade de exclusões (E = 18,6%) ocorridas por apresentarem um alcance prático superior à distância máxima entre os pontos.

É notável a incerteza associada ao verdadeiro valor do efeito pepita do modelo exponencial ($M_{0.5}$) da MO obtido por WLS pois, além de 25% dos modelos $M_{0.5}$ ajustados por WLS terem apresentado efeito pepita nulo, o valor obtido da amostra original (16,95) é superior ao valor obtido em 75 % das réplicas bootstrap. No ajuste realizado por OLS, há evidências de que o estimador do efeito pepita do modelo gaussiano (M_{inf}) obtido pela

amostra original (13,99) está bem definido, pois ele é muito próximo da média e da mediana da distribuição bootstrap deste estimador.

A ocorrência de efeito pepita nulo em 50% dos modelos de dependência espacial do pH do solo ajustados por OLS e WLS indica a não existência de variação do pH em distâncias pequenas. O estimador do parâmetro de alcance do modelo de dependência espacial ($M_{0.5}$) do pH obtido por OLS é superior aos estimadores do parâmetro de alcance dos modelos exponenciais ($M_{0.5}$) obtidos por WLS e ML. Entretanto, esse não é o mais acurado, pois seu intervalo de confiança apresenta a maior amplitude quando comparado aos demais. O parâmetro de alcance do modelo de dependência espacial do pH obtido por WLS apresenta o menor erro padrão, logo, é coerente assumir que tal estimador seja o mais acurado.

5.2.3.4 Mapas de variabilidade espacial

Quando se comparam os mapas da produtividade de soja (Figuras 3-a, 3-b e 3-c), verifica-se que enquanto os mapas provenientes dos modelos ajustados por OLS e WLS são similares e apresentam a produtividade de soja distribuída em cinco classes, o mapa proveniente do modelo ajustado por ML apresenta menor variação entre os valores interpolados. Esta diferença é explicada pelo fato do modelo de dependência espacial da produtividade de soja ajustado por ML apresentar fraca dependência espacial segundo a classificação proposta por Cambardella *et al*. (1994). Logo, são produzidos valores interpolados próximos da média dos dados.

Os mapas de contorno do fósforo (Figuras 3-d, 3-e e 3-f) apresentaram comportamento semelhante e destaca-se a formação de algumas regiões circulares centradas nos pontos amostrais. Tais regiões também são evidentes nos mapas do potássio, gerados pelos modelos obtidos por OLS e WLS (Figuras 3-g e 3-h), os quais, segundo Menezes *et al*. (2016), representam o fenômeno conhecido como "*bull eyes effect*". Esse efeito ocorre quando os modelos apresentam um raio de dependência espacial próximo da distância mínima entre pontos. Como o modelo de dependência espacial do potássio obtido por ML apresentou um raio de dependência espacial elevado (569 m), o mapa de contorno obtido com seus estimadores (Figura 3-i) não apresentou as regiões circulares. Entretanto, como o referido modelo apresentou dependência espacial fraca, os valores interpolados apresentaram comportamento homogêneo, cujo resultado foi um mapa suavizado com poucos intervalos.

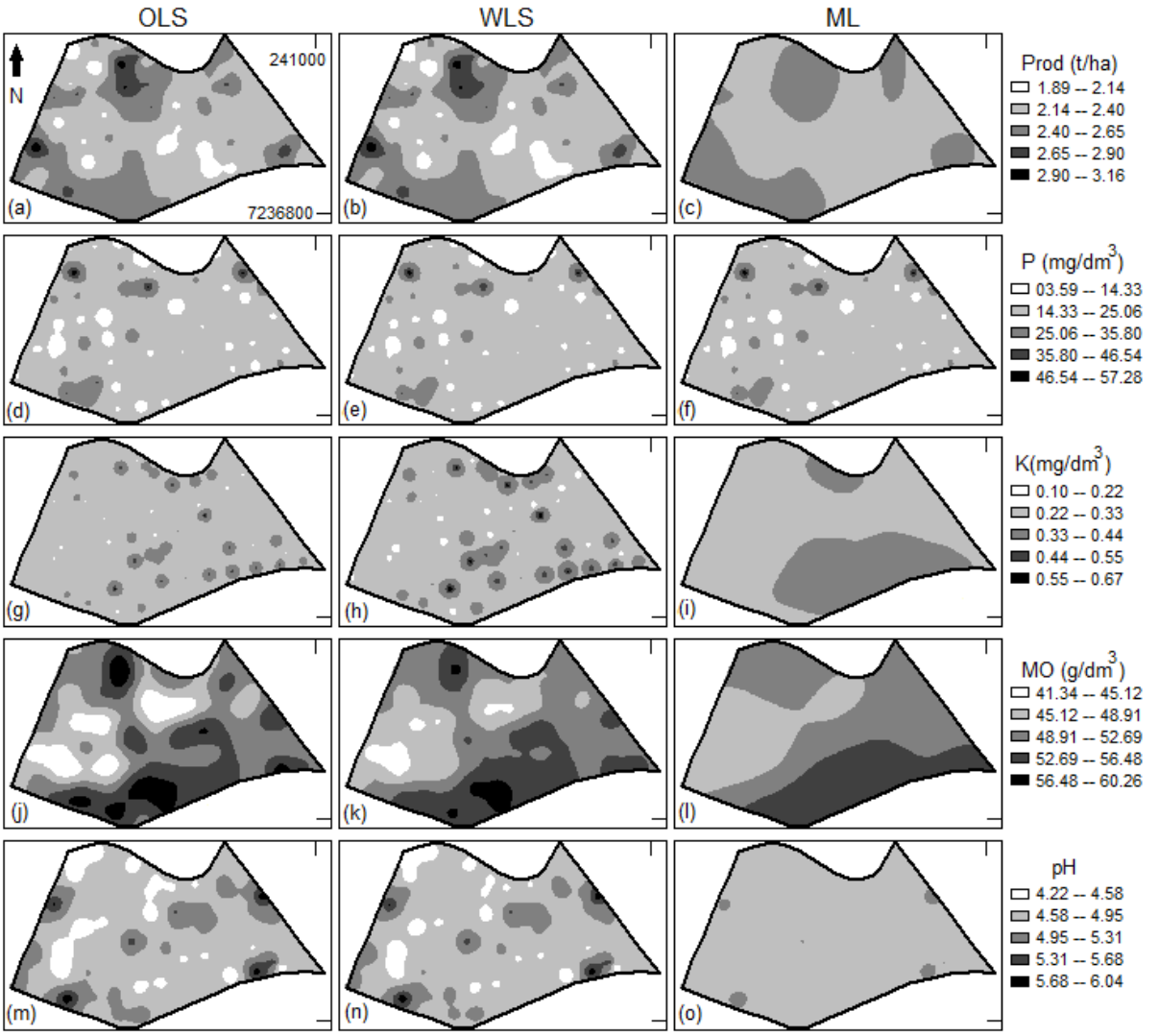

Figura 3 – Mapas de contorno gerados utilizando krigagem ordinária. (a) Prod/OLS; (b) Prod/WLS; (c) Prod/ML; (d) P/OLS; (e) P/WLS; (f) P/ML; (g) K/OLS; (h) K/WLS; (i) K/ML; (j) MO/OLS; (k) MO/WLS; (l) MO/ML; (m) pH/OLS; (n) pH/WLS e (o) pH/ ML.

Era esperada a suavização do mapa da matéria orgânica (MO) (Figura 3(l)), obtido a partir do modelo ajustado por ML, tendo em vista que este modelo apresentou efeito pepita superior à contribuição. E, conforme explicam Molin *et al*. (2015), situações com elevado efeito pepita originam mapas interpolados mais suavizados e mais suscetíveis a erros de estimação.

A diferença entre os graus de dependência espacial também pode ser observada nos mapas do pH do solo (Figuras 3-m, 3-n e 3-o). Os mapas gerados com os modelos ajustados por OLS e WLS são mais heterogêneos do que o mapa gerado pelo modelo ajustado por ML. Isto decorreu porque houve forte dependência espacial para OLS e WLS e fraca dependência espacial para o modelo ajustado por ML, segundo a classificação proposta por Cambardella *et al*. (1994).

5.2.3.5 Intervalos bootstrap de 95% de confiança para os valores preditos

De maneira geral, os valores de produtividade de soja preditos que utilizaram o modelo gaussiano ajustado por ML (Tabela 4) foram superiores aos valores obtidos pelos demais modelos e, além disto, apresentaram menor erro padrão. Vale ressaltar que, na modelagem da produtividade de soja, o modelo gaussiano ajustado por ML (Tabela 4) apresentou fraca dependência espacial, e consequentemente, os valores interpolados ficaram próximos da média da produtividade de soja da amostra original (2,37 t/ha).

Tabela 4 – Intervalos de 95% de confiança bootstrap para os valores preditos em sete locais não amostrados destacados na Figura 1.

		OLS / M05				WLS / M05				ML / GAUSS			
Var	Pontos	$\hat{z}$	Li	Ls	E	$\hat{z}$	Li	Ls	E	$\hat{z}$	Li	Ls	E
prod	P_1	2,23	2,09	2,66	0,14	2,22	2,07	2,71	0,16	2,31	2,14	2,61	0,12
	P_2	2,27	2,08	2,68	0,15	2,27	2,04	2,69	0,17	2,38	2,12	2,63	0,12
	P_3	2,53	2,10	2,67	0,14	2,55	2,08	2,69	0,15	2,47	2,16	2,63	0,11
	P_4	2,19	2,08	2,69	0,15	2,18	2,06	2,72	0,17	2,29	2,12	2,65	0,13
	P_5	2,19	2,13	2,63	0,12	2,17	2,11	2,67	0,14	2,23	2,18	2,60	0,10
	P_6	2,34	2,22	2,54	0,07	2,33	2,19	2,55	0,09	2,34	2,23	2,51	0,07
	P_7	2,38	2,08	2,65	0,14	2,38	2,05	2,69	0,16	2,38	2,13	2,63	0,12
P	P_1	15,45	12,01	28,09	3,79	16,18	13,22	26,17	3,12	16,29	13,54	27,04	3,16
	P_2	14,09	12,44	32,94	4,73	14,89	13,30	30,40	4,03	15,02	13,65	28,31	3,51
	P_3	19,64	12,43	31,54	4,45	19,36	13,42	29,62	3,78	19,33	13,71	27,14	3,28
	P_4	14,21	11,81	30,77	4,61	15,04	12,95	28,77	3,86	15,17	13,63	26,65	3,33
	P_5	17,66	12,74	27,50	3,49	18,09	13,86	25,91	2,85	18,15	14,16	25,07	2,63
	P_6	18,15	15,21	24,10	2,25	18,61	15,92	22,93	1,81	18,66	16,13	22,94	1,74
	P_7	25,31	11,42	29,81	4,37	24,14	12,67	28,72	3,61	23,95	13,78	27,36	3,31
K	P_1	0,31	0,26	0,37	0,03	0,31	0,25	0,38	0,03	0,27	0,22	0,41	0,04
	P_2	0,31	0,26	0,36	0,03	0,29	0,22	0,42	0,05	0,29	0,22	0,42	0,05
	P_3	0,31	0,26	0,38	0,03	0,31	0,23	0,43	0,04	0,31	0,23	0,41	0,05
	P_4	0,31	0,26	0,37	0,03	0,30	0,23	0,43	0,05	0,32	0,22	0,42	0,05
	P_5	0,31	0,26	0,37	0,03	0,31	0,26	0,38	0,03	0,33	0,23	0,40	0,04
	P_6	0,31	0,27	0,36	0,02	0,31	0,28	0,35	0,02	0,31	0,25	0,37	0,03
	P_7	0,31	0,26	0,38	0,03	0,30	0,25	0,40	0,04	0,28	0,22	0,41	0,05
MO	P_1	46,24	43,52	58,77	3,78	45,53	43,91	58,67	3,57	45,37	44,95	58,48	3,44
	P_2	46,44	43,21	59,53	4,09	47,29	43,83	58,89	3,80	48,23	44,61	58,76	3,55
	P_3	48,66	43,15	59,47	3,92	48,52	43,43	58,37	3,67	48,74	44,94	58,46	3,47
	P_4	46,23	43,17	59,73	4,10	47,47	44,10	58,60	3,67	49,44	44,63	58,53	3,51
	P_5	53,00	44,06	58,53	3,67	53,24	44,46	58,25	3,39	52,93	44,85	58,57	3,51
	P_6	50,36	46,46	55,04	2,06	50,88	45,64	56,07	2,57	51,54	45,35	57,15	2,95
	P_7	49,42	43,95	57,97	3,64	49,96	43,16	59,09	3,86	50,41	44,69	57,91	3,37
pH	P_1	4,55	4,45	5,29	0,19	4,59	4,50	5,21	0,17	4,75	4,66	5,02	0,09
	P_2	4,62	4,49	5,38	0,21	4,64	4,52	5,30	0,18	4,77	4,63	5,05	0,10
	P_3	4,70	4,49	5,45	0,22	4,71	4,52	5,40	0,20	4,79	4,65	5,05	0,10
	P_4	5,11	4,46	5,37	0,21	5,08	4,47	5,31	0,19	4,89	4,62	5,07	0,11
	P_5	4,72	4,50	5,25	0,18	4,74	4,54	5,18	0,16	4,81	4,68	5,00	0,08
	P_6	4,84	4,61	5,12	0,12	4,83	4,65	5,04	0,10	4,83	4,71	4,95	0,06
	P_7	4,96	4,47	5,30	0,20	4,94	4,50	5,22	0,18	4,87	4,63	5,06	0,10

Prod: Produtividade de soja (t/ha), P: Fósforo (mg/dm^3), K: Potássio (mg/dm^3), M.O: Matéria orgânica (g/dm^3), pH: pH do solo, $\hat{z}$: estimador, Li: limite inferior do intervalo de confiança, Ls: limite superior do intervalo de confiança, E: erro padrão.

O modelo de dependência espacial do fósforo ajustado por ML destaca-se em relação aos modelos ajustados por OLS e WLS por apresentar valores preditos com menor erro padrão, o que permite conhecer o comportamento do fósforo na região com maior acurácia.

Os intervalos de confiança dos valores preditos de potássio utilizando o modelo exponencial ($M_{0.5}$) obtido por OLS apresentaram as menores amplitudes com valores preditos muito próximos da média dos dados (0,313) na região monitorada. Este fato era esperado, pois o raio de dependência espacial deste modelo (45 m) é inferior à distância mínima entre pontos (49,07 m).

Embora menos evidente, esse comportamento também é observado nos valores preditos de potássio obtidos utilizando os modelos ajustados por WLS e ML e confirmam a observação feita por Drubule (1993) de que em um efeito pepita puro, valores interpolados serão representados pela média.

O grau de dependência espacial dos modelos da matéria orgânica e pH ajustados por ML foi inferior ao dos modelos ajustados por OLS e WLS. Logo, a comparação apenas entre os ajustes realizados por OLS e WLS permite observar que o modelo exponencial ($M_{0.5}$) ajustado por WLS proporcionou os valores interpolados de MO e pH mais acurados, pois a maioria dos pontos preditos por estes modelos apresentaram intervalos de confiança de menor amplitude.

5.2.3.6 Análise dos gráficos QQ plots

De acordo com as interpretações sugeridas por Chambers *et al*. (1983) e Fowlkes (1987), se os quantis das distribuições teóricas estiverem de acordo com os quantis observados (Figura 4), os pontos traçados cairão sobre ou próximos da reta $y = x$, indicando que os dados seguem distribuição normal.

Na Figura 4 verifica-se que, para todos os casos, os pontos encontram-se sobre a linha de referência, ou seja, não existe indício de pontos discrepantes e de anomalias nas caudas das distribuições. Portanto, é coerente assumir que tanto a produtividade de soja quanto os atributos do solo analisados seguem distribuição normal.

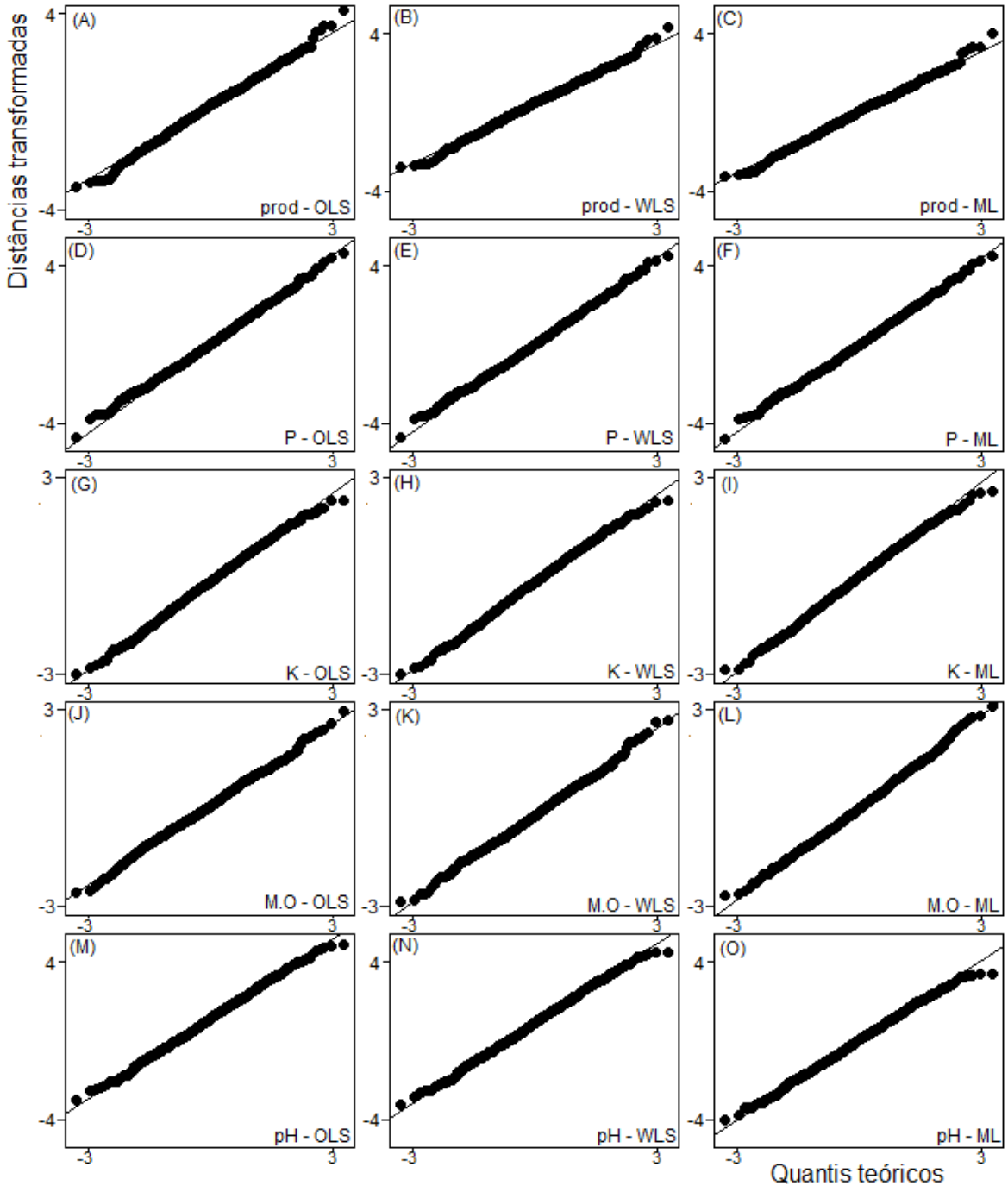

Figura 4 – Gráficos QQ *plots* multivariados, (A) Prod/OLS; (B) Prod/WLS; (C) Prod/ML; (D) P/OLS; (E) P/WLS; (F) P/ML; (G) K/OLS; (H) K/WLS; (I) K/ML; (J) MO/OLS; (K) MO/WLS; (L) MO/ML; (M) pH/OLS; (N) pH/WLS e (O) pH/ML.

5.2.4 Conclusões

Os intervalos de confiança bootstrap para os semivariogramas experimentais permitiram identificar o comportamento das estruturas espaciais das variáveis monitoradas, em especial para as variáveis fósforo (P), potássio (K) e pH, cujas estruturas espaciais apresentaram comportamento constante ao longo das distâncias.

Em relação aos modelos de variabilidade espacial ajustados, destacam-se as estatísticas das réplicas bootstrap das estimativas dos parâmetros, pois elas possibilitaram garantir a não existência de efeito pepita na modelagem do fósforo (P) e também identificar a baixa autocorrelação espacial dos dados de potássio (K).

Os intervalos de confiança bootstrap dos valores krigados permitiram detectar as estimativas de produtividade soja (Prod) e potássio (K) provenientes dos modelos ajustados por ML como as mais acuradas. E também evidenciaram que a utilização de variáveis com baixa autocorrelação espacial ocasiona valores interpolados próximos da média, fato este destacado nos valores estimados de potássio (K).

A utilização do bootstrap espacial para a elaboração dos gráficos QQ *plots* foi fundamental, pois os gráficos possibilitaram a verificação da suposição de normalidade multivariada dos dados.

Como uma análise geoestatística clássica é realizada com um único conjunto amostral, o fato de poder observar o comportamento dos resultados em réplicas deste conjunto possibilitou a realização de inferências de maneira prática. Como consequência, as incertezas associadas à modelagem foram quantificadas e possibilitaram a determinação de modelos geoestatísticos mais acurados e a elaboração de mapas com maior precisão.

Neste sentido, constata-se que a inserção do bootstrap espacial na análise geoestatística é uma prática que pode ser adotada na agricultura de precisão, pois o melhor conhecimento dos atributos do solo permite a elaboração de mapas de aplicação de nutrientes mais precisos, além de proporcionar a melhoria das produtividades das lavouras.

5.2.5 Referências

ACKERSON, J.P., DEMATTÊ, J.A.M., MORGAN, C.L.S. Predicting clay content on field-moist intact tropical soils using a dried, ground VisNIR library with external parameter orthogonalization. **Geoderma**, v. 259–260, p. 196–204, 2015

APARECIDO, L.E.O., ROLIM, G.S., RICHETTI, J., SOUZA, P.S., JOHANN, J.A. Köppen, Thornthwaite and Camargo climate classifications for climatic zoning in the State of Paraná, Brazil. **Ciênc. agrotec.**, v. 40, n. 4, p. 405-417, 2016.

CHAMBERS, J., CLEVELAND, W., KLEINER, B., TUKEY, P. **Graphical Methods for Data Analysis**, Belmont: Wadsworth International Group, 1983.

CAMBARDELLA, C.A., MOORMAN, T.B., NOVAK, J.M., PARKIN, T.B., KARLEN, D.L., TURCO R.F., KONOPKA, A.E. Field scale variability of soil properties in central Iowa soils. **Soil Sci. Soc. Am. J.** v. 58, n. 5, p. 1501-1511, 1994.

DE BASTIANI, F., CYSNEIROS, A.H.M.A., URIBE-OPAZO, M.A., GALEA, M. Influence diagnostics in elliptical spatial linear models. **Test**. v. 24, n. 2, p. 322-340, 2015.

DIGGLE, P.J., RIBEIRO JR., P.J. **Model–based geostatistics**. New York: Springer, 2007.

DUBRULE, O. **Introducing more geology in stochastic reservoir modelling**. In: SOARES, A. (Ed.), Geostatistics Tróia ’92. Kluwer Academic, Dordrecht, p. 351–369, 1993.

EMBRAPA. **Sistema brasileiro de classificação de solos**. Rio de Janeiro: Embrapa, 2009.

EFRON, B. Bootstrap methods: Another look at the jackknife. **Ann. Stat.,** v. 7, n. 1, p. 1-26, 1979.

EFRON, B. **The jackknife, the bootstrap and other resampling plans**. Philadelphia: SIAM, 1982.

EFRON, B., TIBSHIRANI, R.J., 1993. **An Introduction to the Bootstrap**. New York: Chapman e Hall, 1993.

FARACO, M.A., URIBE-OPAZO, M. A., SILVA, E. A. A., JOHANN, J. A., BORSSOI, J. A., 2008. Selection criteria of spatial variability models used in thematical maps of soil physical attributes and soybean yield. **R. Bras. de Ci. Solo**., v. 32, n. 2, p. 463-476, 2008.

FOWLKES, E.B. **A folio of Distributions: A collection of Theoretical Quantile Quantile Plots**. New York: Marcel Dekker, 1987.

GARCÍA-SOIDÁN, P., MENEZES, R., RUBIÑOS, Ó. Bootstrap approaches for spatial data. **Stoch. Environ. Res. Risk. Assess**. v. 28, n. 5, p. 1207-1219, 2014.

GUEDES, L.P.C., RIBEIRO-JUNIOR, P.J., URIBE-OPAZO, M.A., DE BASTIANI, F. Soybean yield maps using regular and optimized sample with different configurations by simulated annealing. **Eng. Agríc.,** v. 36, n. 1, p. 114-125, 2016.

IRANPANAH, N., MANSOURIAN, A., TASHAYO, B., HAGHIGHI, F. Spatial semi-parametric bootstrap method for analysis of kriging predictor of random field. **Procedia Environ. Sci.,**v. 3, p. 81-86, 2011.

JENSEN, J.L., LAKE, L.W., CORBETT, P.W.M., GOGGIN, D.J. **Statistics for petroleum engineers and geoscientists**. Amsterdam: Elselvier, 2000.

JOURNEL, A.G., HUIJBREGTS, C.J. **Mining Geostatistics**. London: Academic Press, 1978.

KRUEGER, J., BÖTTCHER, J.; SCHMUNK, C., BACHMANN, J. Soil water repellency and chemical soil properties in a beech forest soil — Spatial variability and interrelations. **Geoderma**, v. 271, p. 50-62, 2016.

MARDIA, K.V. **Mahalanobis distances and angles**. In: KRISHNAIAH, P.R. (Ed.), Multivariate Analysis. New York: Noth-Holland, p. 176–181, 1977.

MARDIA, K.V., MARSHALL, R.J. Maximum likelihood estimation of models for residual covariance in spatial regression. **Biometrika**, v. 71, n. 1, p. 135-146, 1984.

MATÉRN, B. **Spatial Variation**. New York: Springer, 1986.

MENEZES, M.D., SILVA, S.H.G., MELLO, C.R., OWENS, P.R., CURI, N. Spatial prediction of soil properties in two contrasting physiographic regions in Brazil. **Sci. Agric.**, v. 73, n. 3, p. 274-285, 2016.

MINASNY, B., MCBRATNEY, A.B. The Matérn function as a general model for soil variograms. **Geoderma**, v. 128, p. 192-207, 2005.

MOLIN, J.P., AMARAL, L.R., COLAÇO, A. **Agricultura de precisão**. São Paulo: Oficina de Textos, 2015.

OLEA, R.A., PARDO-IGÚZQUIZA, E., DOWN, P.A. Robust and resistant semivariogram modelling using a generalized bootstrap. **J. S. Afr. I. Min. Metall.**, v. 115, n. 1, p. 37-44, 2015.

PARDO-IGÚZQUIZA, E., OLEA, R.A. Varboot: A spatial bootstrap program for semivariogram uncertainty assessment. **Comput. Geosci.**, v. 41, p. 188-198, 2012.

URIBE-OPAZO, M., BORSSOI, J. A., GALEA, M. Influence diagnostics in gaussian spatial linear models. **J. Appl. Statist.**, v. 39, n. 3, p. 615-630, 2012.

R CORE TEAM. **R: A Language and Environment for Statistical Computing**, Vienna, 2016.

RODRIGUES, Jr., F.A.; BRAMLEY, R.G.V., GOBBETT, D.L. Proximal soil sensing for Precision Agriculture: Simultaneous use of electromagnetic induction and gamma radiometrics in contrasting soils. **Geoderma**, v. 243-244, p. 183-195, 2015.

SARI, K.N., PASARIBU, U.S., NESWAN, O. Estimation of the parameters of isotropic semivariogram model through bootstrap. **App. Math. Sci.**, v. 9, n. 103, p. 5123-5137, 2015.

SEBESTYÉN, Z.F. Uncertainty investigation of the semivariogram by means of the bootstrap method. **Acta Geol. Hung.**, v. 47, n. 1, p. 83-91, 2004.

SCHELIN, L., SJÖSTEDT-DE, S.L. Kriging prediction intervals based on semiparametric bootstrap. **Math. Geosci.**, v. 42, p. 985-1000, 2010.

SOBJAK, R., SOUZA, E.G., BAZZI, C.L., URIBE-OPAZO, M.A., BETZEK, N. M. Redundant variables and the quality of management zones. **Eng. Agríc.**, v. 36, n. 1, p. 78-93, 2016.

SOLOW, A. Bootstrapping correlated data. **Math. Geol.**, v. 17, n. 7, p. 769-775, 1985.

SUTTON, N.J., CHO, S., ARMSWORTH, P.R. A reliance on agricultural land values in conservation planning alters the spatial distribution of priorities and overestimates the acquisition costs of protected areas. **Biol. Cons.**, v. 194, p. 2-10, 2016.

TANG, L., SCHUCANY, W., WOODWARD, W., GUNST, R., 2006. **A Parametric Spatial Bootstrap**. Technical Report SMU-TR-337, Southern Methodist University, Dallas, Texas.

XIONG, X., GRUNWALD, S., MYERS, D.B., KIM, J., HARRIS, W.G., BLIZNYUK, N. Assessing uncertainty in soil organic carbon modeling across a highly heterogeneous landscape. **Geoderma**, v. 251, p. 105-116, 2015.

WEBSTER, R., OLIVER, M.A. **Geostatistics for environmental scientists**. West Sussex: John Wiley e Sons, 2007.

WILSON, E., HILFERTY, M. The distribution of chi-square. **Proc. Nat. Acad. Sci.**, v. 17, p. 684-688, 1931.

5.3 ARTIGO 3 : Inferência em um modelo espacial linear gaussiano da produtividade de soja utilizando métodos bootstrap para dados espaciais

Resumo: Este trabalho tem como objetivo utilizar regressão espacial linear para modelar a produtividade de soja em função dos teores de cálcio, magnésio, potássio, fósforo, manganês e pH do solo. Foram utilizados métodos bootstrap espaciais para determinar estimadores pontuais e por intervalo, associados aos parâmetros do modelo. Realizaram-se testes de hipóteses sobre os parâmetros do modelo e elaboraram-se gráficos de probabilidade *quantile-quantile plots* para identificar a normalidade dos dados. Foram quantificadas as incertezas associadas aos parâmetros da estrutura de dependência espacial e identificaram-se os parâmetros associados às variáveis teor de potássio, teor de fósforo e pH do solo como significativos. Estas variáveis foram utilizadas para a elaboração de um novo modelo, que apresentou raio de dependência espacial próximo da distância mínima entre pontos e proporcionou um mapa de contorno caracterizado por regiões circulares e elevada quantidade de valores próximos à média. A análise dos gráficos *quantile-quantile plots* indicou que os dados de produtividade de soja seguem uma distribuição normal de probabilidade.

Palavras-chave: bootstrap espacial; geoestatística; regressão espacial.

5.3.1 Introdução

A soja é a espécie leguminosa mais importante mundialmente (GALLON *et al*., 2016) e destaca-se no agronegócio brasileiro pelo elevado potencial produtivo nas diferentes regiões (CRUZ *et al*., 2016). A soja é amplamente utilizada para a elaboração de rações animais e óleo e conforme explicam Ávila e Albrecht (2010), ela vem sendo enfatizada como alternativa na prevenção de doenças e na alimentação humana, além de , poder ser transformada em diversos alimentos proteicos.

Dentre os diversos estudos associados à produtividade de soja, destacam-se os métodos geoestatísticos, utilizados para detectar a variabilidade espacial existente nas lavouras (BORSSOI *et al*., 2011; DALCHIAVON *et al*., 2011; KESTRING *et al*., 2015; GUEDES *et al*., 2016). Embora a geoestatística possibilite a detecção da variabilidade espacial da produtividade da soja, como são poucas e esparsas as amostras utilizadas nas análises geralmente (PARDO-IGÚZQUIZA e OLEA, 2012), existem incertezas associadas aos resultados obtidos.

As incertezas ocorrem pelo fato do modelo espacial linear ser estimado por métodos paramétricos, como os apresentados por Mardia e Marshall (1984). Portanto, para quantificar as incertezas associadas aos resultados de uma análise geoestatística, uma alternativa aos métodos tradicionais de inferência é o uso dos métodos de reamostragem

bootstrap espacial (BS) (SOLOW,1985) e bootstrap espacial paramétrico (BSP) (TANG *et al*. 2006), adaptações do método bootstrap (EFRON, 1979) para dados espacialmente dependentes.

O método bootstrap é bem conhecido e tem sido empregado em estudos envolvendo amostras independentes de produtividade de soja (BUENO *et al.*, 2013; DALPOSSO *et al*., 2016; GUPTA e MANJAYA, 2016). Os métodos bootstrap para dados espacialmente dependentes vem ganhando destaque na literatura, devido à importância da modelagem da incerteza nas análises, como pode ser observado nos trabalhos (KANG *et al.* 2008; SCHELIN E SJÖSTEDT-DE LUNA, 2010; OLEA e PARDO-IGÚZQUIZA, 2011 e PARDO-IGÚZQUIZA E OLEA, 2012).

O objetivo deste trabalho foi utilizar os métodos BS e BSP para quantificar as incertezas associadas à modelagem da dependência espacial da produtividade de soja cujas covariáveis são os teores de cálcio, magnésio, potássio, fósforo, manganês e pH do solo.

5.3.2 Material e Métodos

5.3.2.1 Área de estudo e dados

O conjunto de dados foi coletado no ano agrícola 2012/2013 e provém de uma área agrícola de 167,35 hectares, localizada na região Oeste do Paraná, Brasil, próxima ao município de Cascavel, com as seguintes coordenadas centrais: latitude 24º57’25’’S e longitude 53º34’29’’W, e altitude média de 714 m.

O clima da região da área agrícola é segundo a classificação de Köppen do tipo Cfa (APARECIDO *et al.*, 2016) e o solo é classificado como Latossolo Vermelho distroférrico (EMBRAPA, 2009).

Foi realizada uma amostragem sistemática centrada com pares de pontos próximos (*lattice plus close pairs*), composta de 99 elementos amostrais (Figura 1), georreferenciados com um aparelho GPS GEOEXPLORE 3 com precisão de 5 m.

Em cada ponto amostral determinaram-se a produtividade de soja (Prod, t/ha) e os atributos cálcio (Ca, cmolc/dm^3), magnésio (Mg, cmolc/dm^3), potássio (K, mg/dm^3), fósforo (P, mg/dm^3), manganês (Mn, mg/dm^3) e o pH do solo.

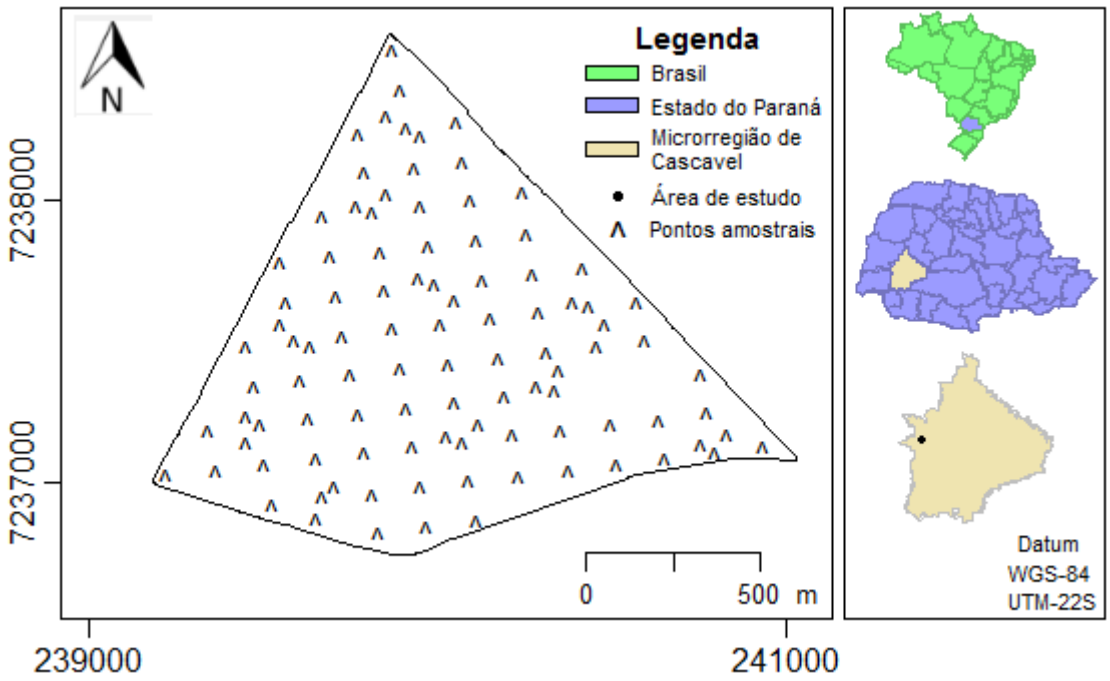

Figura 1 – Mapa de localização da área em estudo.

5.3.2.2 Análise geoestatística

Considere um processo estocástico gaussiano $\{Z(\mathbf{s}), \mathbf{s} \in S\}$, com $S \subset \Re^d$, sendo $\Re^d$ o espaço euclidiano, *d*-dimensional ($d \geq 1$). Suponha-se que os dados deste processo, $Z(s_1), \ldots, Z(s_n)$, sejam registrados em localizações espaciais conhecidas s_i $(i = 1, \ldots, n)$, e gerados pelo seguinte modelo, escrito na forma matricial:

$$\boldsymbol{Z}(\boldsymbol{s}) = \boldsymbol{\mu}(\boldsymbol{s}) + \boldsymbol{\varepsilon}(\boldsymbol{s}), \tag{1}$$

em que, o termo determinístico $\boldsymbol{\mu}(\boldsymbol{s})$ é um vetor $n \times 1$ que representa a média do processo $\boldsymbol{Z}(\boldsymbol{s})$ e o termo estocástico $\boldsymbol{\varepsilon}(\boldsymbol{s})$ é um vetor $n \times 1$ com vetor de média zero $E[\boldsymbol{\varepsilon}(\boldsymbol{s})] = \mathbf{0}$ e matriz de covariância $\boldsymbol{\Sigma} = C[(\sigma_{iu})]$, onde $\sigma_{iu} = cov(\varepsilon(s_i), \varepsilon(s_u))$.

O vetor de médias $\boldsymbol{\mu}(s)$ pode ser escrito como um modelo espacial linear da forma $\boldsymbol{\mu}(\boldsymbol{s}) = \boldsymbol{X\beta}$, onde $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ é um vetor $p \times 1$ de parâmetros desconhecidos e $\boldsymbol{X} = \boldsymbol{X}(\boldsymbol{s})$ é uma matriz $n \times p$, formada com p variáveis explicativas na posição $s \in S$. Segundo Mardia e Marshall (1984), a matriz de covariância pode ser expressa na forma paramétrica $\boldsymbol{\Sigma} = \varphi_1 \mathbf{I}_n + \varphi_2 \mathbf{R}(\varphi_3)$ em que $\mathbf{I}_n$ é a matriz identidade $n \times n$, $\varphi_1 \geq 0$ é o parâmetro denominado efeito pepita; $\varphi_2 \geq 0$ é o parâmetro denominado contribuição; $\varphi_3 \geq 0$ é o parâmetro que define o alcance (a) do modelo e $\mathbf{R}(\varphi_3) = \left[(r_{ij})\right]$ é uma matriz simétrica $n \times n$. Os elementos r_{ij} da matriz $\mathbf{R}$, $i, j = 1, \ldots, n$, representam a correlação entre os pontos s_i e s_j, sendo $r_{ij} = 1$ se $i = j$; $r_{ij} = 0$ se $i \neq j$ e $\varphi_2 = 0$ e $r_{ij} = \varphi_2^{-1}\sigma_{ij}$ se $i \neq j$ e $\varphi_2^{-1} \neq 0$. Em que

$\sigma_{ij} = C(h_{ij})$ uma função que depende de $h_{ij} = \|s_i - s_j\|$, que é a distância euclidiana entre os pontos s_i e s_j (DE BASTIANI *et al.*, 2015).

A identificação da estrutura de dependência espacial dos atributos demandou a construção de semivariogramas experimentais omnidirecionais utilizando o estimador de Matheron, e, para modelar as estruturas de dependência espacial, utilizou-se o modelo da família Matérn (Matérn, 1986), apresentado na Equação (2):

$$C(h_{ij}) = \begin{cases} \varphi_1 + \varphi_2 & , \quad i = j \\ \frac{\varphi_2}{2^{k-1}\Gamma(k)} \left(\frac{h_{ij}}{\varphi_3}\right)^k K_k\left(\frac{h_{ij}}{\varphi_3}\right), & i \neq j \end{cases}, \tag{2}$$

em que $K_k(.)$ é a função de Bessel do terceiro tipo de ordem $k > 0$ e $\Gamma(.)$ é a função Gama. Para processos Gaussianos estacionários de segunda ordem e isotrópicos, a função semivariância tem a relação $\gamma(h_{ij}) = C(0) - C(h_{ij})$ (URIBE-OPAZO *et al.*, 2012) para $i, j = 1, \ldots, n$ e $h_{ij} \geq 0$. Na Equação (2), o parâmetro de forma k controla o comportamento próximo à origem e à suavização analítica do processo e neste trabalho foram considerados os valores fixos $k = \{0{,}5; 1{,}5; 2{,}5; 4{,}5\}$ (DIGGLE e RIBEIRO JR., 2007).

A estimação dos parâmetros dos modelos foi realizada considerando o método da máxima verossimilhança (ML) e os critérios considerados para escolher o modelo geoestatístico para a matriz de covariância foram a validação cruzada (VC), o traço da matriz de covariância (Tr) e o máximo valor da log-verossimilhança (LMV) (DE BASTIANI *et al.*, 2015).

A investigação da significância das variáveis explicativas foi feita a partir do teste da razão de verossimilhança (RAO, 1973) no vetor de parâmetros $\boldsymbol{\beta}$ e para realizar predições em locais não amostrados, utilizou-se a krigagem com deriva externa (WACKERNAGEL, 1995).

5.3.2.3 Bootstrap espacial

Neste trabalho, foram utilizados dois métodos: o bootstrap espacial (BS), proposto por Solow (1985) e apresentado no Algoritmo 1 e o bootstrap espacial paramétrico (BSP), proposto por Tang *et al*. (2006) e apresentado no Algoritmo 2.

Algoritmo 1: Bootstrap espacial (SOLOW, 1985).

a) Considerando-se o conjunto de dados espaciais $\{Z(s_1), \ldots, Z(s_n)\}$, determine o vetor dos resíduos $\hat{\boldsymbol{\varepsilon}}_{(S)} = \mathbf{Z}(s) - \hat{\boldsymbol{\mu}}$ sendo $\hat{\boldsymbol{\mu}} = \left(\mathbf{X}^T \hat{\boldsymbol{\Sigma}}^{-1} \mathbf{X}\right)^{-1} \mathbf{X}^T \hat{\boldsymbol{\Sigma}}^{-1} \mathbf{Z}$ o estimador de $\boldsymbol{\mu}$ e $\hat{\boldsymbol{\Sigma}}$ a matriz de covariância estimada de $\boldsymbol{\Sigma}$. b) Considerando-se a matriz de covariância estimada $\hat{\boldsymbol{\Sigma}}$, utilize o

método de decomposição de Cholesky para obter $\hat{\boldsymbol{\Sigma}} = \hat{\mathbf{L}}\hat{\mathbf{L}}^{\mathrm{T}}$, em que $\hat{\mathbf{L}}$ é uma matriz triangular inferior de ordem n; c) Utilizando-se a matriz inversa $\hat{\mathbf{L}}^{-1}$, determine $\hat{\boldsymbol{\varepsilon}}_{\mathrm{dec}} = \hat{\mathrm{L}}^{-1}\hat{\varepsilon}$, o vetor de resíduos descorrelacionados e centralize seus valores, obtendo $\tilde{\mathbf{e}} = \hat{\boldsymbol{\varepsilon}}_{\mathrm{dec}} - \left(\frac{1}{\mathrm{n}}\right)\sum_{i=1}^{n}\hat{\boldsymbol{\varepsilon}}_{\mathrm{dec}}$; d) Realize *n* reamostragens com reposição do conjunto dos resíduos descorrelacionados e centralizados $\tilde{\mathrm{e}}_1, \ldots, \tilde{\mathrm{e}}_{\mathrm{n}}$, para formar o vetor de resíduos bootstrap $\boldsymbol{\varepsilon}^*_{\mathrm{SB}} = (\varepsilon^*_1, \ldots, \varepsilon^*_{\mathrm{n}})^{\mathrm{T}}$; e) A amostra bootstrap espacial é obtida com ao serem recorrelacionadso os resíduos bootstrap $\mathbf{Z}^* = \hat{\boldsymbol{\mu}} + \hat{\mathbf{L}}\boldsymbol{\varepsilon}^*_{\mathrm{SB}}$.

Algoritmo 2: Bootstrap espacial paramétrico (TANG *et al*., 2006).

a) Considerando-se o conjunto de dados espaciais $\{\mathrm{Z}(\mathrm{s}_1),\ldots,\mathrm{Z}(\mathrm{s}_{\mathrm{n}})\}$, determine o vetor dos resíduos $\hat{\boldsymbol{\varepsilon}}_{(S)} = \mathbf{Z}(\mathrm{s}) - \hat{\boldsymbol{\mu}}$ sendo $\hat{\boldsymbol{\mu}} = \left(\mathbf{X}^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{Z}$ o estimador de mínimos quadrados de μ e $\hat{\Sigma}$ a matriz de covariância estimada, respectivamente; b) Utilize o método de decomposição de Cholesky para obter $\hat{\boldsymbol{\Sigma}} = \hat{\mathbf{L}}\hat{\mathbf{L}}^{\mathrm{T}}$, em que $\hat{\mathbf{L}}$ é uma matriz triangular inferior de ordem n; c) Utilize a distribuição normal padrão N(0,1) para criar um vetor de tamanho *n*, denominado de vetor de resíduos bootstrap paramétricos $\boldsymbol{\varepsilon}^*_{\mathrm{PSB}} = (\varepsilon^*_1, \ldots, \varepsilon^*_{\mathrm{n}})^{\mathrm{T}}$.
d) A amostra bootstrap espacial é obtida pela recorrelação entre os resíduos bootstrap $\mathbf{Z}^* = \hat{\boldsymbol{\mu}} + \hat{\mathbf{L}}\boldsymbol{\varepsilon}^*_{\mathrm{PSB}}$.

Conforme explicam Tang *et al*. (2006), o método BSP não descorrelaciona o vetor dos resíduos $\hat{\boldsymbol{\varepsilon}}$ como no método BS. Ao invés disto, os resíduos são gerados independentes a partir de uma distribuição normal padrão. A fundamentação teórica do método BSP pode ser vista em Sjöstedt-de Luna e Young (2003).

5.3.2.4 Quantificação das incertezas na análise geoestatística

Utilizaram-se os Algoritmos 1 e 2 para determinar B = 1000 amostras bootstrap do conjunto de dados de produtividade de soja. Para cada amostra ajustou-se um modelo, procedimento este que permitiu construir a distribuição empírica dos parâmetros do modelo, e consequentemente, determinar estimadores pontuais e intervalos de confiança dos parâmetros utilizando o método percentil (EFRON, 1982).

Para verificação da suposição de normalidade multivariada, calculou-se a distância de Mahalanobis $S_i = (Z_i - \boldsymbol{X\beta})^T\boldsymbol{\Sigma}^{-1}(Z_i - \boldsymbol{X\beta})$ para as B = 1000 amostras bootstrap, desta forma, estabeleceu-se a proximidade entre cada observação e o centro da distribuição. Como as distâncias de Mahalanobis são independentes e apresentam distribuição assintóticamente qui-quadrado com *p* graus de liberdade, sendo *p* o número de variáveis explicativas (MARDIA, 1977), utilizou-se a aproximação de Wilson e Hilferty (1931) para

transformá-las em escores z. Os valores resultantes foram ordenados e plotados versus os valores esperados das estatísticas de ordem normal.

A implementação computacional realizada neste trabalho foi desenvolvida no software R (R CORE TEAM, 2016).

5.3.3 Resultados e discussão

Segundo dados da CONAB (2016), a média da produtividade de soja da área estudada em 2012/2013 (Tabela 1) foi inferior à média do estado do Paraná (3,38 t/ha) e superior à média nacional (2,94 t/ha). Os teores de cálcio (Ca) variaram de 2,30 cmolc/dm^3 a 13,10 cmolc/dm^3 e os teores de magnésio (Mg) variaram de 0,85 cmolc/dm^3 a 5,75 cmolc/dm^3, sendo classificados como altos teores (TOMÉ JR., 1997).

Tabela 1. Análise descritiva da produtividade de soja e das variáveis explicativas.

Estatísticas	Prod	Ca	Mg	K	P	Mn	pH
Mínimo	2,15	2,30	0,85	0,09	5,80	37,00	3,90
Q_1	2,95	5,49	1,99	0,22	13,40	62,00	4,70
Mediana	3,23	6,37	2,53	0,31	17,40	71,00	5,00
Média	3,25	6,51	2,64	0,33	18,20	74,50	5,00
Q_3	3,53	7,48	3,10	0,40	21,40	84,00	5,25
Máximo	4,51	13,10	5,75	1,11	52,40	140,00	7,10
DP	0,47	1,64	0,90	0,15	7,56	19,36	0,45
CV (%)	14,30	25,20	34,30	47,10	41,50	25,98	9,06

Prod: produtividade de soja (t/ha), Ca: cálcio (cmolc/dm^3), Mg: magnésio (cmolc/dm^3), K: potássio (mg/dm^3), P: fósforo (mg/dm^3), Mn: manganês (mg/dm^3), pH: pH do solo, Q_1: primeiro quartil, Q_3: terceiro quartil, DP: desvio padrão, CV: coeficiente de variação.

E acordo com a análise da classificação do CV, destaca-se elevada dispersão dos teores de potássio (K) e fósforo (P) e homogeneidade dos dados de pH. Para todos os modelos ajustados, os valores estimados por ML de φ_1 foram iguais a zero, enquanto os valores estimados de φ_2 foram semelhantes e os valores estimados de φ_3 indicaram um raio de dependência espacial com variação de 109 a 112 m.

De acordo com os critérios VC, LMV e Tr (Tabela 2), o melhor ajuste é proporcionado pelo modelo Matérn com parâmetro de forma k = 4,5.

Tabela 2. Critério para seleção do modelo de produtividade de soja elaborado com covariáveis considerando a função de covariância Matérn.

k	VC	LMV	Tr
0,5	0,2616456	-54,56541	1,298518
1,5	0,2589309	-54,07624	1,285411
2,5	0,2576171	-53,89274	1,277847
4,5	**0,2563692**	**-53,73963**	**1,270118**

k: parâmetro de forma do modelo Matérn; VC: validação cruzada; LMV: máximo valor do logaritmo da função verossimilhança; TR: traço da matriz de covariância.

O valor associado ao teor de potássio (K) (Tabela 3) destaca-se como o mais elevado e por apresentar sinal positivo, indicando ser a variável que mais contribui para o aumento da média da produtividade de soja (t ha^{-1}). Este resultado corrobora com o estudo de Dos Passos et al. (2015), tendo em vista que o potássio é o segundo nutriente mais absorvido pela planta de soja (DOS PASSOS *et al*., 2015), portanto, é essencial para o crescimento e desenvolvimento das plantas (LI *et al.*, 2015). As estimações dos parâmetros associadas às variáveis teor fósforo (P) e teor de manganês (Mn) indicam que essas variáveis pouco influenciam a média do processo estocástico (Tabela 3).

Tabela 3. Parâmetros estimados para o modelo espacial linear pelo método da máxima verossimilhança considerando a função de covariância Matérn com parâmetro de forma k = 4,5.

$\hat{\beta}_0$	$\hat{\beta}_{Ca}$	$\hat{\beta}_{Mg}$	$\hat{\beta}_K$	$\hat{\beta}_P$	$\hat{\beta}_{Mn}$	$\hat{\beta}_{PH}$	$\hat{\varphi}_1$	$\hat{\varphi}_2$	$\hat{\varphi}_3$
4,54	0,06	0,06	0,85	-0,01	0,00	-0,44	0,00	0,18	14,25
(0,68)	(0,05)	(0,06)	(0,30)	(0,00)	(0,00)	(0,18)	(0,08)	(0,09)	(0,98)

$\hat{\beta}_0$: Intercepto; $\hat{\beta}_i$: Estimador do parâmetro associado a variável i = {Ca, Mg, K, P, Mn, PH}, Ca: teor de cálcio (cmolc/dm^3), Mg: magnésio (cmolc/dm^3), K: potássio (mg/dm^3), P: fósforo (mg/dm^3), Mn: manganês (mg/dm^3), pH: pH do solo; $\hat{\varphi}_1$: estimador do efeito pepita; $\hat{\varphi}_2$:estimador da contribuição; $\hat{\varphi}_3$: estimador do parâmetro que define o alcance. Os valores entre parênteses correspondem ao desvio padrão de cada parâmetro estimado.

O fato de o efeito pepita ser nulo ($\hat{\varphi}_1 = 0$) indica que, em pequenas distâncias, a variabilidade também é pequena. Isso significa que as distâncias consideradas entre as amostras foram adequadas. Para avaliar a confiabilidade das estimativas obtidas, foram determinados os intervalos de confiança bootstrap (Tabela 4).

Destaca-se que dos B=1000 modelos ajustados considerando o método BS, dois modelos (0,2%) foram excluídos e dos B=1000 modelos ajustados considerando o método BSP, apenas um modelo (0,1%) foi excluído. As exclusões foram realizadas pelo motivo de que nestes ajustes o raio de dependência espacial foi superior à distância máxima entre amostras (1766 m). Conforme destacam Dalposso *et al*. (2009), modelos que apresentam

este comportamento devem ser descartados, pois consideram as informações que vão além da área em estudo. E, independente do método bootstrap utilizado, verifica-se que os intervalos de confiança para os parâmetros associados aos teores de cálcio (Ca), magnésio (Mg) e manganês (Mn) contém o zero, assim, tais variáveis podem não ser significativas (Tabela 4).

Os estimadores bootstrap do efeito pepita em ambos os métodos bootstrap destacam a ocorrência de efeito pepita nulo em 50% dos modelos (Tabela 4). A distribuição bootstrap das réplicas bootstrap da contribuição ($\hat{\varphi}_2$) foi assimétrica negativa para ambos os métodos bootstrap considerados, portanto, a ocorrência de valores elevados foi mais frequente. Embora o zero esteja contido nos intervalos de confiança das réplicas bootstrap da contribuição ($\hat{\varphi}_2$), há evidências para se assumir que a contribuição não é nula, pois além do valor obtido na amostra original ser diferente de zero (0,18), 94 % (95,1%) das réplicas bootstrap obtidas pelo método BS (BSP) apresentaram contribuição diferente de zero.

Tabela 4. Estatísticas descritivas e intervalos de 95% de confiança percentil de Efron da distribuição bootstrap dos estimadores dos parâmetros do modelo da estrutura de dependência espacial da produtividade de soja (Prod) considerando as covariáveis Ca, Mg, K, P, Mn e PH.

Métodos	Parâmetros	Min	Q_1	Mediana	Média	Q_3	Max	DP	As	Li	Ls
BS	β_0	2,50	4,03	4,54	4,53	4,99	7,16	0,71	0,15	3,15	5,94
	β_{Ca}	-0,12	0,02	0,05	0,05	0,09	0,27	0,06	0,31	-0,04	0,17
	β_{Mg}	-0,16	0,02	0,07	0,06	0,11	0,28	0,07	-0,11	-0,07	0,20
	β_K	-0,05	0,64	0,84	0,84	1,05	1,83	0,30	0,01	0,25	1,43
	β_P	-0,04	-0,02	-0,01	-0,01	-0,01	0,01	0,01	-0,04	-0,03	-0,00
	β_{Mn}	-0,00	0,00	0,00	0,00	0,01	0,01	0,00	-0,02	-0,00	0,01
	β_{PH}	-1,14	-0,56	-0,43	-0,43	-0,31	0,15	0,19	-0,15	-0,80	-0,09
	$\hat{\varphi}_1$	0,00	0,00	0,00	0,05	0,10	0,24	0,06	0,90	0,00	0,18
	$\hat{\varphi}_2$	0,00	0,05	0,14	0,12	0,17	0,30	0,07	-0,47	0,00	0,21
	$\hat{\varphi}_3$	0,00	13,00	16,40	22,30	21,40	283,00	24,80	4,52	11,40	26,20
BSP	β_0	2,25	4,13	4,55	4,53	4,98	6,37	0,68	-0,09	3,12	5,86
	β_{Ca}	-0,14	0,02	0,06	0,06	0,10	0,23	0,06	-0,18	-0,05	0,17
	β_{Mg}	-0,18	0,02	0,06	0,06	0,10	0,26	0,06	0,01	-0,06	0,19
	β_K	-0,08	0,66	0,86	0,86	1,06	2,06	0,31	-0,00	0,23	1,49
	β_P	-0,04	-0,02	-0,01	-0,01	-0,01	0,01	0,01	0,04	-0,03	-0,00
	β_{Mn}	-0,00	0,00	0,00	0,00	0,01	0,01	0,00	-0,04	-0,00	0,01
	β_{PH}	-0,97	-0,56	-0,44	-0,44	-0,33	0,13	0,18	0,09	-0,79	-0,07
	$\hat{\varphi}_1$	0,00	0,00	0,00	0,05	0,10	0,22	0,06	0,94	0,00	0,18
	$\hat{\varphi}_2$	0,00	0,06	0,14	0,12	0,17	0,28	0,07	-0,50	0,00	0,22
	$\hat{\varphi}_3$	0,00	13,20	16,70	23,10	21,70	232,20	24,90	3,99	11,70	27,70

BS: bootstrap espacial, BSP: bootstrap espacial paramétrico, β_0: Intercepto do modelo, β_i: Parêmetro associado a variável i = {Ca, Mg, K, P, Mn, PH}, Ca: cálcio (cmolc/dm^3), Mg: magnésio (cmolc/dm^3), K: potássio (mg/dm^3), P: fósforo (mg/dm^3), Mn: manganês (mg/dm^3), pH: pH do solo; $\hat{\varphi}_1$:efeito pepita, $\hat{\varphi}_2$: contribuição, $\hat{\varphi}_3$: parâmetro de alcance, Min: mínimo, Q_1: primeiro quartil, Q_3: terceiro quartil, Max: máximo, DP: desvio padrão, As: coeficiente de assimetria, Li: limite inferior do intervalo de confiança, Ls: limite superior do intervalo de confiança.

As réplicas bootstrap do parâmetro que define o alcance ($\hat{\varphi}_3$), obtidas pelos métodos BS e BSP, apresentaram uma distribuição assimétrica positiva, fato este que também pode ser observado no trabalho de Olea e Pardo-Igúzquiza (2011). Isto ocorre porque alguns ajustes geram raios de dependência espacial elevados, que deslocam a cauda da distribuição.

Na comparação dos desvios padrões dos parâmetros do modelo (Tabela 3), com os respectivos desvios padrões, obtidos por métodos bootstrap (Tabela 4), destaca-se que, com exceção do desvio padrão do parâmetro de alcance ($\hat{\varphi}_3$), os demais apresentaram valores semelhantes.

Assim, para investigar a significância dos parâmetros estimados, inicialmente testaram-se as hipóteses $\mathcal{H}_0: \beta_{\mathrm{i}} = 0,\ i = \{Ca, Mg, K, P, Mn, PH\}$ contra as respectivas hipóteses alternativas bilaterais $\mathcal{H}_1: \beta_{\mathrm{i}} \neq 0$ (Tabela 5).

Tabela 5. Valor do teste de razão de verossimilhança (LR) e p-valor para as hipóteses $\mathcal{H}_0: \beta_{\mathrm{i}} = 0,\ \ i = \{Ca, Mg, K, P, Mn, PH\}$ e $\mathcal{H}_0: \beta_{\mathrm{Ca}} = \beta_{\mathrm{Mg}} = \beta_{\mathrm{K}} = \beta_{\mathrm{P}} = \beta_{\mathrm{Mn}} = \beta_{\mathrm{PH}} = 0$ no modelo espacial linear.

Hipótese nula	LR	p-valor
$\beta_{\mathrm{Ca}} = 0$	1,18	0,2033
$\beta_{\mathrm{Mg}} = 0$	0,86	0,2782
$\beta_{\mathrm{K}} = 0$	7,68	0,0030 *
$\beta_{\mathrm{P}} = 0$	4,40	0,0211 *
$\beta_{\mathrm{Mn}} = 0$	2,76	0,0604
$\beta_{\mathrm{PH}} = 0$	5,66	0,0098 *
$\beta_{\mathrm{Ca}} = \beta_{\mathrm{Mg}} = \beta_{\mathrm{K}} = \beta_{\mathrm{P}} = \beta_{\mathrm{Mn}} = \beta_{\mathrm{PH}} = 0$	18,94	0,0017 *

LR: Teste da razão de verossimilhança, p-valor: nível descritivo do teste ao 5% de significância

Destaca-se que as hipóteses $\mathcal{H}_0: \beta_{\mathrm{Ca}} = 0$, $\mathcal{H}_0: \beta_{\mathrm{Mg}} = 0$ e $\mathcal{H}_0: \beta_{\mathrm{Mn}} = 0$ não são rejeitadas ao nível de significância de 5%. O mesmo resultado foi observado nos intervalos de confiança das réplicas bootstrap (Tabela 4), logo, os métodos bootstrap espaciais figuram como alternativa para testar o efeito individual das variáveis no modelo.

Em virtude de não serem significativos os parâmetros associados aos teores de cálcio (Ca, cmolc dm^{-3}), magnésio (Mg, cmolc/dm^3) e manganês (Mn, mg/dm^3), optou-se pelo ajuste de modelos espaciais lineares considerando o potássio (K, mg/dm^3), o fósforo (P, mg/dm^3) e o pH do solo, que apresentaram parâmetros significativos.

De acordo com os critérios VC, LMV e Tr, o melhor ajuste foi proporcionado pelo modelo com parâmetro de forma k = 4,5 (Tabela 6). Diante destas observações, evidencia-se que os métodos BS e BSP permitem a obtenção de um modelo com menor quantidade de parâmetros sem a perda da capacidade de explicação da produtividade de soja.

Tabela 6. Parâmetros estimados para o modelo espacial linear pelo método da máxima verossimilhança considerando a função de covariância Matérn com parâmetro de forma k = 4,5 e as variáveis K, P e pH.

$\hat{\beta}_0$	$\hat{\beta}_K$	$\hat{\beta}_P$	$\hat{\beta}_{PH}$	$\hat{\varphi}_1$	$\hat{\varphi}_2$	$\hat{\varphi}_3$
3,99	0,84	-0,01	-0,16	0,00	0,19	13,14
(0,49)	(0,30)	(0,00)	(0,10)	(0,11)	(0,12)	(0,93)

$\hat{\beta}_0$: Intercepto; $\hat{\beta}_i$: Estimador do parâmetro associado a variável i = {K, P, PH}, K: teor de potássio (mg/dm^3), P: teor de fósforo (mg/dm^3), pH: pH do solo; $\hat{\varphi}_1$: estimador do efeito pepita; $\hat{\varphi}_2$:estimador da contribuição; $\hat{\varphi}_3$: estimador do parâmetro que define o alcance. Os valores entre parênteses correspondem ao desvio padrão de cada parâmetro estimado.

De acordo com o mapa de contorno da produtividade gerado utilizando krigagem com deriva externa (Figura 2) considerando o modelo espacial linear apresentado na Tabela 6, destaca-se que a maioria da área (84%) ficou com valores de produtividade, cuja variação foi de 3,175 t/ha a 3,592 t/ha.

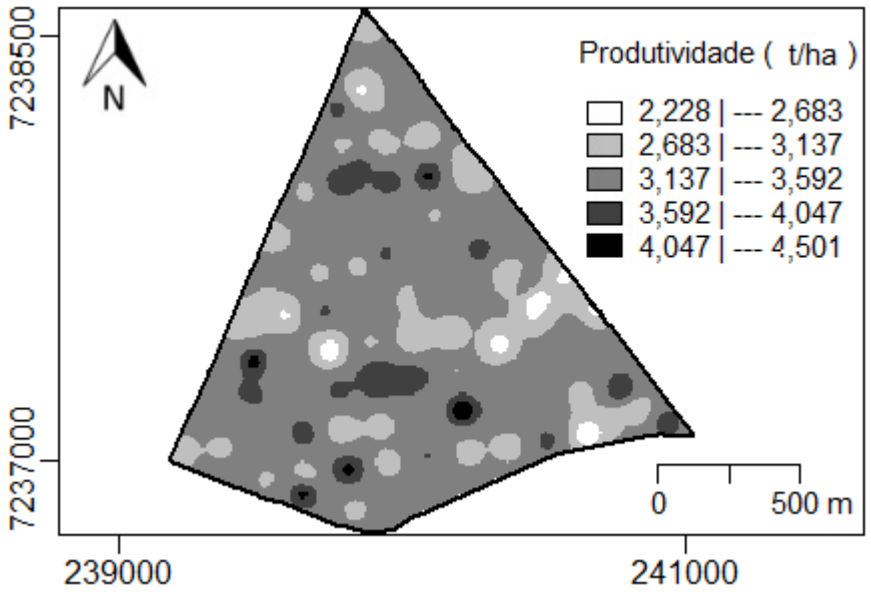

Figura 2 – Mapa de produtividade de soja gerado utilizando krigagem com deriva externa.

Outra característica observada no mapa de produtividade de soja (Figura 2) é a presença de regiões circulares centradas nos pontos amostrais. Tais regiões representam o fenômeno conhecido como "*bull eyes effect*" que, conforme explicam Menezes *et al*. (2016), ocorre quando o modelo apresenta raio de dependência espacial próximo à distância mínima entre pontos amostrais.

A análise dos gráficos QQ plots (Figura 3) mostra que os pontos se encontram sobre a linha de referência. Desta forma, é coerente assumir que a produtividade de soja segue uma distribuição normal de probabilidade.

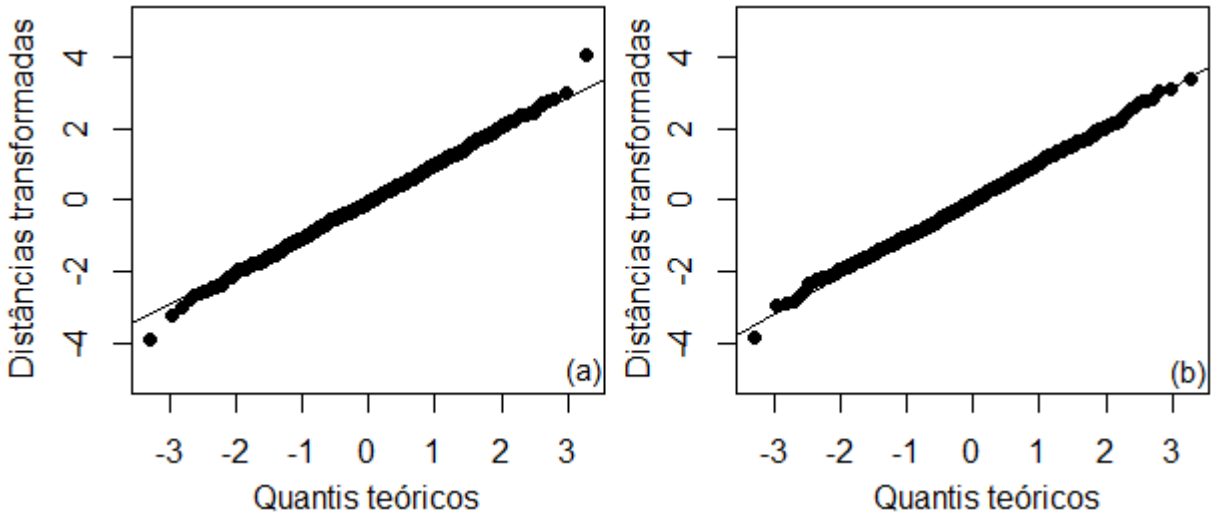

Figura 3 – Gráficos QQ plots multivariados da produtividade de soja (t ha^{-1}) (a) Considerando réplicas bootstrap geradas pelo método BS e (b) Considerando réplicas bootstrap geradas pelo método BSP.

5.3.4 Conclusões

Os métodos BS e BSP permitiram quantificar as incertezas associadas à estrutura de dependência espacial e avaliar a significância individual dos parâmetros associados à média do modelo espacial linear. Assim, é possível a determinação de um modelo com menor quantidade de parâmetros sem a perda da capacidade de explicação da produtividade de soja.

Evidencia-se que o fato do modelo espacial linear formado com as variáveis potássio (K, mg dm^{-3}), fósforo (P, mg dm^{-3}) e pH do solo ter apresentado raio de dependência espacial próximo à distância mínima entre pontos proporcionou um mapa caracterizado por regiões circulares e com elevada quantidade de valores próximos à média.

Por fim, a utilização dos métodos bootstrap na elaboração dos gráficos *quantile-quantile plots* permitiu verificar a suposição de normalidade multivariada dos dados.

5.3.5 Referências Bibliográficas

APARECIDO, L.E.; ROLIM, G.S.; RICHETTI, J.; SOUZA, P.S.; JOHANN, J.A. Köppe, Thornthwaite and Camargo climate classifications for climatic zoning in the State of Paraná, Brazil. **Ciência e Agrotecnologia**, v. 40, n. 4, p. 405-417, 2016.

ÁVILA, M. R.; ALBRECHT, L. P. Isoflavonas e a qualidade das sementes de soja. **Informativo Abrates**, v. 20, n. 1, p. 15-29, 2010.

BORSSOI, J.A.; URIBE-OPAZO, M.A.; GALEA, M. Técnicas de diagnóstico de influência local na análise da produtividade da soja. **Engenharia Agrícola**, v. 31, n. 2, p. 376-387, 2011.

BUENO, R.D.; BORGES, L.L.; ARRUDA, M.A.; BHERING, L.L.; BARROS, E.G.; MOREIRA, M.A. Genetic parameters and genotype x environment interaction for productivity, oil and protein content in soybean. **African Journal of Agricultural Research**, v. 8, n. 38, p. 4853-4859, 2013.

CONAB, 2016. **Soja – Brasil**: série histórica de produtividade. http://www.conab.gov.br. [17 Nov. 2016].

CRUZ, S. C. S.; SENA-JUNIOR, D. G.; SANTOS, D. M. A.; LUNEZZO, L. O.; MACHADO, C. G. Cultivo de soja sob diferentes densidades de semeadura e arranjos espaciais. **Revista de Agricultura Neotropical**, v. 3, n. 1, p. 1-6, 2016.

DALCHIAVON, F.C.; CARVALHO, M.P.; NOGUEIRA, D.C.; ROMANO, D.; ABRANTES, F.L.; ASSIS, J.T.; OLIVEIRA, M.S. Produtividade da soja e resistência mecânica à penetração do solo sob sistema plantio direto no cerrado brasileiro. **Pesquisa Agropecuária Tropical**, v. 41, n. 1, p. 8-19, 2011.

DALPOSSO, G.H.; URIBE-OPAZO, M.A.; BORSSOI, J.A.; JOHANN, J.A.; MERCANTE, E. Previsão da produção de Trigo utilizando métodos geoestatísticos. In: **Avances en Ingenieria Rural 2007-2009**. Rosario: Editorial de la Universidad Nacional de Rosario, 2009. v. 1, n. 1, p. 78-86.

DALPOSSO, G.H.; URIBE-OPAZO, M.A.; JOHANN, J. A. Soybean yield modeling using bootstrap methods for small samples. **Spanish Journal of Agricultural Research**, v. 14, n. 3, p. e0207, 2016.

DE BASTIANI, F.; MARIZ DE AQUINO CYSNEIROS, A.H.; URIBE-OPAZO, M.A.; GALEA. M. Influence diagnostics in elliptical spatial linear models. **Test**, v. 24, n. 2, p. 322-340, 2015.

DIGGLE, P.J.; RIBEIRO JR., P.J. **Model–based geostatistics**. New York: Springer, New York, 2007.

DOS PASSOS, A.M.A.; REZENDE, P.M.; CARVALHO, E.R.; ÁVILA, F.W. Biochar, farmyard manure and poultry litter on chemical attributes of a Distrophic Cambissol and soybean crop. **Revista Brasileira de Ciências Agrárias**, v. 10, n. 3, p. 382-388, 2015.

EFRON, B. Bootstrap methods: Another look at the jackknife. **Annals of Stattistics**, v. 7, n. 1, p. 1-26, 1979.

EFRON, B. **The jackknife, the bootstrap and other resampling plans**. Philadelphia: SIAM, 1982.

EMBRAPA. **Sistema brasileiro de classificação de solos**. Rio de Janeiro: Embrapa, 2009.

GALLON, M.; BUZZELLO, G.L.; TREZZI, M.M.; DIESEL, F.; SILVA, H.L. Ação de herbicidas inibidores da PROTOX sobre o desenvolvimento, acamamento e produtividade da soja. **Revista Brasileira de Herbicidas**, v. 15, n. 3, p. 232-240, 2016.

GUEDES, L.P.C.; RIBEIRO JR., P.J.; URIBE-OPAZO, M.A.; DE BASTIANI, F. Mapas da produtividade da soja usando configurações amostrais regulares e otimizadas pela têmpera simulada. **Engenharia Agrícola**, v. 36, n. 1, p. 114-125, 2016.

GUPTA, S.K.; MANJAYA, J.G. Assessment of genetic variation at soybean mosaic virus resistance loci in Indian Soybean (*Glycine max L. Merill*) genotypes using SSR markers. **Eletronic Journal of Plant Breeding**, v. 7, n. 2, p. 392-400, 2016.

KANG, C.; CHOI, S.; YOO, S.H. A spatial bootstrap method for kriging variance. **Journal of the Korean Data Analysis Society**, v. 10, n. 3, p. 1247-1254, 2008.

KESTRING, F.B.F.; GUEDES, L.P.C.; DE BASTIANI, F.; URIBE-OPAZO, M.A. Comparação de mapas temáticos de diferentes grades amostrais para a produtividade da soja. **Engenharia Agrícola**, v. 35, n. 4, p. 733-743, 2015.

LI, L.; XU, L.; WANG, X.; PAN, G.; LU, L. De novo characterization of the alligator weed (*Alternanthera philoxeroides*) transcriptome illuminates gene expression under potassium deprivation. **Journal of Genetics**, v. 94, n. 1, p. 95-104, 2015.

MARDIA, K.V. Mahalanobis distances and angles, in: Krishnaiah, P.R. (Ed.), **Multivariate Analysis**. Noth-Holland, New York, p. 176–181, 1977.

MARDIA, K.V.; MARSHALL, R.J. Maximum likelihood estimation of models for residual covariance in spatial regression, **Biometrika**, v. 71, n. 1, p. 135-146, 1984.

MENEZES, M.D., SILVA, S.H.G., MELLO, C.R., OWENS, P.R., CURI, N., 2016. Spatial prediction of soil properties in two contrasting physiographic regions in Brazil. **Scientia Agricola**, v. 73, n. 3, p. 274-285, 2016.

OLEA, R.A.; PARDO-IGÚZQUIZA, E. Generalized bootstrap method for assessment of uncertainty in semivariogram inference. **Mathematical Geosciences**, v. 43, n. 2, p. 203-228, 2011.

PARDO-IGÚZQUIZA, E.; OLEA, R.A. VARBOOT: A spatial bootstrap program for semivariogram uncertainty assessment. **Computers e Geosciences**, v. 41, n. 1, p. 188-198, 2012.

R CORE TEAM, R: A Language and Environment for Statistical Computing. **R Foundation for Statistical Computing**, Vienna, 2016.

RAO, C.R. **Linear statistical inference and its applications**. New York: John Wiley e Sons, 1973.

SOLOW, A. Bootstrapping correlated data. **Math. Geol.**, v. 17, n. 7, p. 769–775, 1985.

SCHELIN, L.; SJÖSTEDT-DE LUNA, S. Kriging prediction intervals based on semiparametric bootstrap, **Mathematical Geosciences**, v. 42, n. 1, p. 985-1000, 2010.

SJÖSTEDT-DE LUNA, S.; YOUNG, A. The bootstrap and kriging prediction intervals. **Scandinavian Journal of Statistics**, v. 30, n. 1, p. 175-192, 2003.

TANG, L.; SCHUCANY, W.; WOODWARD, W.; GUNST, R. **A parametric spatial bootstrap**. Technical Report SMU-TR-337. Dallas: Southern Methodist University, 2006.

TOMÉ JR., J. B. **Manual para interpretação de análise de solo**. Guaíba: Agropecuária, 1997.

URIBE-OPAZO, M., BORSSOI, J. A., GALEA, M. Influence diagnostics in gaussian spatial linear models. **Journal of Applied Statistcs**, v. 3, n. 39, p. 615-630, 2012.

WACKERNAGEL, H., 1995. **Multivariate Geostatistics**. Berlin: Springer-Verlag, 1995.

WILSON, E., HILFERTY, M. The distribution of chi-square. **Procedings of the National Academy of Sciences**, v. 17, n. 12, p. 684-688, 1931.

6 Considerações Finais

Considerando que os resultados da tese foram apresentados em forma de artigos e que as conclusões foram apresentadas em cada um deles, estão relacionadas abaixo as principais contribuições verificadas no desenvolvimento deste trabalho:

- O trabalho constribuiu com a agricultura de precisão uma vez que propôs uma metodologia eficaz para a realização de inferências estatísticas em estudos relacionados a esta área de pesquisa. Em um primeiro momento, foram utilizados métodos bootstrap considerando dados independentes (Artigo 1). Posteriormente, utilizou-se o método bootstrap com dados espaciais (Artigos 2 e 3). Em ambos os casos foi possível a realização de inferências utilizando a reamostragem bootstrap.

- Foram utilizadas técnicas que permitiram obter um modelo de regressão linear múltipla, formado com uma quantidade mínima de parâmetros e com maior capacidade de explicação do comportamento da produtividade de soja.

- Constatou-se que a inserção do bootstrap espacial na análise geoestatística é uma prática que pode ser adotada na agricultura de precisão, pois o melhor conhecimento dos atributos do solo permite a elaboração de mapas de aplicação de nutrientes mais precisos, proporcionando a melhoria das produtividades das lavouras.

Além das contribuições apresentadas acima, surgiram assuntos de interesse a serem abordados em futuros trabalhos ao longo do desenvolvimento da tese:

- Utilizar o método bootstrap dos resíduos em estudos de regressão linear múltipla.
- Utilizar os métodos bootstrap-t e Bca (bias-corrected and accelerated) para determinar intervalos de confiança.
- Desenvolver trabalhos utilizando bootstrap em blocos tanto para dados temporais quanto para dados espaciais.
- Elaborar um trabalho visando ao estudo dos erros em modelos de regressão linear utilizando intervalos de confiança de Atkison.
- Pesquisar sobre a metodologia bootstrap bagging.
- Utilização de outros métodos de reamostragem (Jackknife, Delta, .632+ bootstrap).

7 Anexos

7.1 Anexo A – Normas da revista SJAR

O Jornal Espanhol de Pesquisa Agrícola (SJAR) é uma revista internacional trimestral que aceita artigos de pesquisa, revisões e comunicações curtas de conteúdo relacionado à agricultura. Os artigos de pesquisa e comunicações curtas devem relatar trabalhos originais não publicados anteriormente em qualquer idioma e não submetidos para publicação em outro lugar. O texto do artigo deve conter as seguintes seções: Resumo, Introdução, Material e Métodos, Resultados, Discussão e Referências. Abaixo são apresentados alguns exemplos para elaboração das referências pela revista.

Journal article

Romero-del-Castillo R, Costell E, Plans M, Simó J, Casañas F, 2012. A standardized method of preparing common beans (Phaseolus vulgaris L.) for sensory analysis. J Sens Stud 27: 188-195.

Vasileiadis VP, Froud-Williams RJ, Loddo D, Eleftherohorinos IG, 2016. Emergence dynamics of barnyardgrass and jimsonweed from two depths when switching from conventional to reduced and no-till conditions. Span J Agric Res 14 (1): e1002.

Books

Milthorpe FL, Moorby J, 1999. An introduction to crop physiology. CAB Intnal, Wallingford, UK. 244 pp.

Madsen E (ed), 2007. Effect of CO2 concentration on morphological, histological and cytological and physiological processes in tomato plants. State Seed Testing Station, Denmark. 246 pp.

MARM, 2008. Anuario de estadística agroalimentaria. Ministerio de Medio Ambiente y Medio Rural y Marino, Gobierno de España.

Chapters of books

Pla I, 1996. Soil salinization and land desertification. In: Soil degradation and desertification in Mediterranean environments; Rubio JL, Calvo A (eds.). pp: 105-129. Elsevier, Amsterdam.

Doctoral or master thesis

Flores M, 2000. Las técnicas biomoleculares en el diagnóstico y tipificación de los patógenos vegetales. Doctoral thesis. Univ. Politécnica, Valencia, Spain.

Fernández JL, 2010b. Estudio agroecológico del cultivo del maíz y sus potencialidades en la sustentabilidad de pequeñas fincas campesinas. Master's thesis. Univ. Int. de Andalucía, Cádiz, Spain. 143 pp.

Conferenceproceedings
Sanz-Romero P, Gonzalez-Mesa JC, Calvo-Gutierrez F, 2000. Nonpoint sources of water contamination and their impacts on sustainability. Proc V Int Conf on Tomato Breeding and Genetics, Kaunas (Lithuania), Sept 13-16. pp: 187-192.

Work documents
Miravete EJ, 1999. Aplicación de los modelos de elección discreta al análisis de la adopción de innovaciones tecnológicas. Instituto Valenciano de Investigaciones Económicas. Valencia, Spain. EC Work Document 99-04.

Cathagne A, Guyomard H, Levert F, 2006. Milk quotas in the European Union: distribution of marginal costs and quota rents. European Dairy Industry Model. Working paper 01/2006.

Legal documents
BOE, 2000. Royal decree 995/2000, of 20 June, that established water quality objectives for several pollutants. Boletín Oficial del Estado (Spain) No. 147, 20/06/00.

EC, 2004. Council Directive 2004/68/EC laying down animal health rules for the importation into and transit through the Community of certain live ungulate animals, amending Directives 90/426/EEC and 92/65/EEC. 26 April 2004 [LEX-FAOC065206].

Maiores informações podem ser consultadas no endereço eletrônica da revista: http://revistas.inia.es/index.php/sjar/index