ADRIANA MARIA ROCHA TRANCOSO SANTOS

OUTLIERS EM VARIÁVEIS GEOESPACIAIS: PROPOSIÇÕES UTILIZANDO GEOESTATÍSTICA

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do programa de pós-graduação em Engenharia Civil, para obtenção do título de Doctor Scientiae.

VIÇOSA
MINAS GERAIS – BRASIL
2016

Ficha catalográfica preparada pela Biblioteca Central da Universidade Federal de Viçosa - Câmpus Viçosa

T
S237o
2016

Santos, Adriana Maria Rocha Trancoso, 1973-
Outliers em variáveis geoespaciais : proposições utilizando geoestatística / Adriana Maria Rocha Trancoso Santos. – Viçosa, MG, 2016.
xvii, 63f. : il. (algumas color.) ; 29 cm.

Orientador: Nilcilene das Graças Medeiros.
Tese (doutorado) - Universidade Federal de Viçosa.
Inclui bibliografia.

1. Sistemas de informação geográfica. 2. Estatística.
I. Universidade Federal de Viçosa. Departamento de Engenharia Civil. Programa de Pós-graduação em Engenharia Civil.
II. Título.

CDD 22 ed. 910.285

ADRIANA MARIA ROCHA TRANCOSO SANTOS

OUTLIERS EM VARIÁVEIS GEOESPACIAIS: PROPOSIÇÕES UTILIZANDO GEOESTATÍSTICA

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do programa de pós-graduação em Engenharia Civil, para obtenção do título de Doctor Scientiae.

APROVADA: 16 de dezembro de 2016.

João Marcos Louzada	Afonso de Paula dos Santos
Gérson Rodrigues dos Santos	Eduardo Antônio Gomes Marques (Coorientador)

Nilcilene das Graças Medeiros
(Orientadora)

Ao meu GRANDIOSO DEUS,
Criador e Mantenedor
de todas as coisas,
DEDICO.

Aos meus amados, esposo e filha,
companheiros de todas as horas,
OFEREÇO.

“Nós somos do tamanho dos
sonhos de DEUS!”
Autor Desconhecido

AGRADECIMENTOS

À Deus, agradeço e divido esta alegria. “Cuidei de você no seu dia mais feliz! Cuidei de você em cada conquista que fez! Filha, eu cuido de você!”

Aos meus pais Alcides e Margarida por serem meus exemplos, por me ensinarem valores eternos, por me mostrarem que com ética, honestidade, trabalho, humildade e solidariedade podemos quebrar muros e construir pontes. O que sabemos ninguém nos “rouba”, o que norteia tudo é a educação, ela transforma pessoas, muda o mundo e somos agentes nesta mudança.

A minha filha amada e querida, presente de DEUS! Obrigada filha por ser quem você é!

Ao meu esposo Gérson, que além de ser um exemplo como Profissional é meu professor preferido! Com seu amor, dedicação me fez gostar de “sua ciência”, a Geoestatística. É o meu amado e amigo de todos os momentos, que me mostra sempre que existe, sim, alguém com quem posso contar. Obrigada por seu amor, dedicação, altruísmo, carinho, paciência, incentivo, cumplicidade e “broncas”. Muito obrigada! Sem você, este sonho, não seria real.

Aos meus familiares pelo incentivo, principalmente os meus irmãos. Amo vocês.

À Universidade Federal Viçosa, como um todo, por realizar o meu grande sonho de estudar numa FEDERAL e por toda a infraestrutura disponibilizada. UFV, você sempre será a melhor. Estudar, Saber, Agir e Vencer.

A todos os servidores do Departamento de Biologia Geral pelo apoio, principalmente ao Gustavo, Leandro, Betinha, Joãozinho e Letícia por “segurarem as pontas” enquanto eu precisava sair pra estudar. Obrigada!

Ao Professor João Marcos de Araújo que, com seu jeito único, me deu aulas de ética, amizade e nunca me impediu de crescer.

À professora Maria Goretti que também não mediu esforços para possibilitar o meu crescimento. Obrigada!

À todos os professores do Departamento de Biologia Geral que em cada Colegiado aprovava meu crescimento.

À professora Nilcilene Medeiros pela confiança, orientação, ensinamentos, atenção e dedicação a este trabalho. Sou sua fã!

Ao professor Eduardo Marques pela Co-orientação e pelo aprendizado.

Aos coautores professor Paulo César Emiliano, professor João Marcos Louzada e M.Sc. Ricardo Soares Ramos pelas grandes contribuições neste trabalho.

Ao professor Afonso Santos pela educação, humildade, conhecimento, disposição e sabedoria compartilhadas nessa minha caminhada.

Aos professores Maria Lúcia Calijuri, Leonardo, Eduardo Marques, Nilcilene Medeiros, José Ivo Junior, Gérson Santos e Afonso Santos pelas excelentes aulas.

Aos meus colegas de Doutorado e de disciplinas pelas noites de estudos e pela cumplicidade, principalmente Kamilla Andrade e Letícia D'Agosto, obrigada por serem além de colegas.

Aos colegas Alessandra e Francisco pelos materiais cedidos e troca de informações.

À equipe administrativa do Departamento de Engenharia Civil pela prestatividade e competência, principalmente à Cilene e Cristina.

À Lidiane Rosa pela disposição de sempre ajudar.

À equipe da PPG e PGP da Universidade Federal de Viçosa por serem tão prestativos quando precisei, principalmente Gil e Wilson.

À Dra. Amy Kaleita pela orientação no período que estudei na Iowa State University. Thank you so much, my advisor!

Aos amigos que fiz em Ames, IA, durante o intercâmbio científico, principalmente aos amáveis e queridos JP Foote, Nancy Foote, Emilly, Jo Ellen, Farshad, Yun Gao, and 7th-Day Adventist Church members.

À Dra. Laurimar por ter cedido o conjunto de dados desse trabalho. Obrigada!

À Lígia de Oliveira Serrano Pruski pela paciência em nos conduzir nos primeiros dias em Ames e pela amizade e dicas tão preciosas.

À família Tormena pela amizade, cumplicidade e belos e inesquecíveis momentos, que mesmo longe de casa e sem o baldinho da Gigi não deixamos apagar a estrela interior!

À família Precci pela amizade e, os inesquecíveis momentos em Ames.

Aos "amigos para sempre" Lincoln, Sandra, Felipe e Sérgio, pelo amor dedicado. We are missing you, our friends!

À querida prima Dainha, pelos momentos inesquecíveis!

À família da nossa amiga-irmã Andreia Vieira, Dr. Peter e lindos filhos.

Aos amigos inesquecíveis Joey e Chris Church. Thank you, so much, guys. You are awesome friends!

À Capes pelo investimento.

A todos os amigos da Pós-Graduação em Informações Espaciais, pela amizade, horas de estudos, discussões e conversas, principalmente ao Paulo, Sílvia e Edilson.
A todos que, de alguma forma, contribuíram para realização deste trabalho.

SUMÁRIO

LISTA DE FIGURAS

INTRODUÇÃO

CAPÍTULO 1

CAPÍTULO 2

CAPÍTULO 3

LISTA DE TABELAS

CAPÍTULO 1

CAPÍTULO 2

CAPÍTULO 3

RESUMO

SANTOS, Adriana Maria Rocha Trancoso, D.Sc., Universidade Federal de Viçosa, dezembro de 2016. **Outliers em variáveis geoespaciais: proprosições utilizando Geoestatística.** Orientadora: Nilcilene das Graças Medeiros. Coorientador: Eduardo Antonio Gomes Marques.

As observações que se afastam estatisticamente das demais em um conjunto de dados comumente são denominadas de outliers. Tal comportamento faculta o surgimento de hipóteses como por exemplo, a de que os dados pertencem à outra população. Contudo, independentemente das hipóteses que podem surgir, é importante considerar frequentemente a adequabilidade das metodologias existentes aos diversos tipos de variáveis envolvidas em investigações científicas. Na literatura especializada, é comum encontrar na metodologia o uso do Box Plot como principal mecanismo de detecção, e a exclusão dos dados "discrepantes", detectados por este mecanismo, do conjunto de dados em estudo. Como o Box Plot é um mecanismo que não leva em consideração a posição geográfica dos dados, tem-se como hipótese a não aplicabilidade deste em dados geoespaciais contínuos. Assim, apresenta-se neste trabalho um estudo sobre a importância da proposição de métodos de detecção de outliers que incorporam a localização dos dados, bem como a comparação de seu desempenho com o Box Plot. No primeiro capítulo foi proposto um novo método de detecção de outliers para dados geoespaciais contínuos, em que um conjunto de dados reais, sabidamente com outliers, foi analisado tanto pelo Box Plot quanto pelo método em proposição. No segundo capítulo foi proposto um novo método de detecção de outliers para dados geoespaciais contínuos, cujas variáveis são não-negativas. Um conjunto de dados reais foi analisado usando o Box Plot e usando o novo método proposto. Finalmente, no terceiro capítulo foi proposto um mecanismo metodológico para a decisão de exclusão dos dados com alta probabilidade de discrepância. Neste capítulo foram utilizados quatro conjuntos de dados, sendo três simulados computacionalmente e um conjunto de dados reais. Visando robustecer teoricamente toda a proposição do trabalho, adotou-se como princípios norteadores uma combinação de teoremas da Estatística Clássica e da aplicação da Geoestatística, como principal metodologia de apoio. A Geoestatística foi adotada por incorporar a localização geográfica dos dados no processo analítico, estar baseada em suas características estatisticamente ótimas, ou seja, uma metodologia criada para ser sem tendência e com

variância mínima na predição de valores não observados, além de levar em consideração na modelagem e predição a estrutura de dependência espacial das amostras, o que é inerente aos dados geoespaciais.

ABSTRACT

SANTOS, Adriana Maria Rocha Trancoso, D.Sc., Universidade Federal de Viçosa, December, 2016. **Outliers in geospatial variables: propositions using Geostatistics.** Advisor: Nilcilene das Graças Medeiros. Co-advisor: Eduardo Antonio Gomes Marques.

The observations that differ statistically from the others in a data set commonly are named outliers. Such behavior empowers the emergence of hypothesis such as, the data belong to another population. However, independently from the hypothesis that may arise, it is important to consider frequently the suitability of the existent methodologies to the many types of involved variables in scientific investigations. In the specialized literacy, it is common to find in the suggested methodology the use of the Box Plot as a main mechanism of detection, and the exclusion of "discrepant" data of the data set studied, detected by this mechanism. Since the Box Plot is a mechanism that does not take into consideration the geographic position of the data, there is the hypothesis of the non-suitability of such mechanism in continuous geospatial data. Thus, it is presented in this work a study about the importance of a proposition of methods of outliers detection that incorporate the localization of the data, comparing them to the Box Plot. In the first chapter it was proposed a new method of outliers detection for continuous geospatial data, in which the real data set, with known outliers, was analyzed through the Box Plot and the proposition method. In the second chapter it was proposed a new method of outliers detection for continuous geospatial data, which variables are nonnegatives. A real data set, was analyzed using the Box Plot and using the new proposed method. Finally, in the third chapter it was proposed a methodological mechanism for the decision of exclusion of the data with high probability of discrepancy. In this chapter there were utilized four data sets, being one a real data set and three simulated computationally. Aiming to theoretically strengthen in all of the work's proposition, it was adopted as guiding principles a combination of theorems of Classic Statistics and of the application of Geostatistics, as main support methodology. The Geostatistics was adopted for incorporating a geographic localization of the data in the analytical process, being based in its statistically great characteristics, meaning that, a created methodology to be without trend and with minimum variance in the prediction of non observed values, besides taking

into consideration in the modeling and prediction the structure of the spatial dependence of the samples, with is inherent to the geospatial data.

INTRODUÇÃO GERAL

Os dados provenientes de variáveis geoespaciais contínuas necessitam de procedimentos analíticos sofisticados, pois do contrário não seria possível caracterizar, modelar e inferir com precisão tais dados. Essas ações são primordiais no gerenciamento eficaz consequente (QIAO et al., 2013).

Qualquer variável pode conter discrepâncias em pequena, média e grande escala. Assim, torna-se necessária a constante condução de estudos que avaliem estes comportamentos, tidos como perturbadores do esperado probabilisticamente para os dados. As causas podem ser diversas, tais como, mudança do local de observação, erros instrumentais, erros dos observadores, problemas na mecanização de monitoramento, entre outros (MORETTIN e TOLÓI, 2002).

Metodologias de detecção de outliers têm sido um tema recorrente de pesquisadores de todas as épocas. Muitos trabalhos surgiram com o objetivo de contribuírem com este assunto, como é o caso de Anscombe (1960), Grubbs (1969), Beckman e Cook (1983), Rousseuw e Zomeren (1990), Muñoz-Garcia et al. (1990), Barnett e Lewis (1994), entre outros pioneiros.

Alguns destes autores afirmam que a preocupação com dados discrepantes é tão antiga quanto às primeiras tentativas de análises de um conjunto de dados, como é o caso dos comentários de Bernoulli, no século XVIII, sobre a existência de tais dados.

Metodologias mais recentes têm sido criadas para atender às demandas das diversas áreas do conhecimento científico, como é o caso de Hongxing et al. (2001) para dados espaciais distribuídos de forma irregular em malhas amostrais, Barua e Alhajj (2007) para processamento de imagens, Qiao et al. (2013) para dados provenientes de satélites e Appice et al. (2014) para fluxo de dados geofísicos.

Ferreira (2009) afirma que provavelmente todo pesquisador já tenha se deparado com um conjunto de dados em que algumas observações se afastam demasiadamente das restantes, sugerindo, inclusive, como pertencentes de outra população.

Pode-se perceber então que um outlier é caracterizado pela sua relação com as demais observações que fazem parte de um conjunto amostral. O seu distanciamento em relação às outras observações sempre foi o primeiro alvo de investigação.

Um dos métodos mais utilizados para a detecção de outliers é o Método de Box Plot, criado por Tukey (1977).

Devido sua simplicidade de construção, utilização e interpretação, muitos são os trabalhos que apresentam essa metodologia de detecção.

Basicamente, é um método que pode ser aplicado a variáveis quantitativas contínuas, em que a mediana, primeiro quartil, terceiro quartil, valor mínimo e valor máximo são utilizados para a construção da "caixinha" de decisão.

As regras para a criação do Box Plot se resumem a, conforme apresentado na Figura 1: (1) obter a AIQ - amplitude interquartílica, ou seja, a distância entre o Q1 - primeiro quartil e o Q3 - terceiro quartil; (2) obter o fechamento do Box, tomando o primeiro e o terceiro quartil, ou seja, Q1 e Q3; (3) obter os limites aceitáveis de 1,5 AIQ acima do Q3 e abaixo do Q1, ou seja, [Q1 – 1,5 x AIQ; Q3 + 1,5 x AIQ]; (4) a decisão sobre o que é ou não outlier é tomada ao se observar os valores que estão acima e abaixo dos limites aceitáveis (TUKEY, 1977).

Apesar da simplicidade do método, não há base estatística para a decisão de Tukey para considerar 1,5 ou mais AIQ acima e abaixo do quartis.

Uma das vantagens do Método de Box Plot é a sua robustez em relação à quebra da pressuposição de normalidade dos dados ou de qualquer outra distribuição.

Uma das desvantagens de método é a grande perda de informações para a criação do mesmo, ou seja, os dados são utilizados apenas na ordenação e contagem da amostra, pois a escolha dos quartis depende desses fatores.

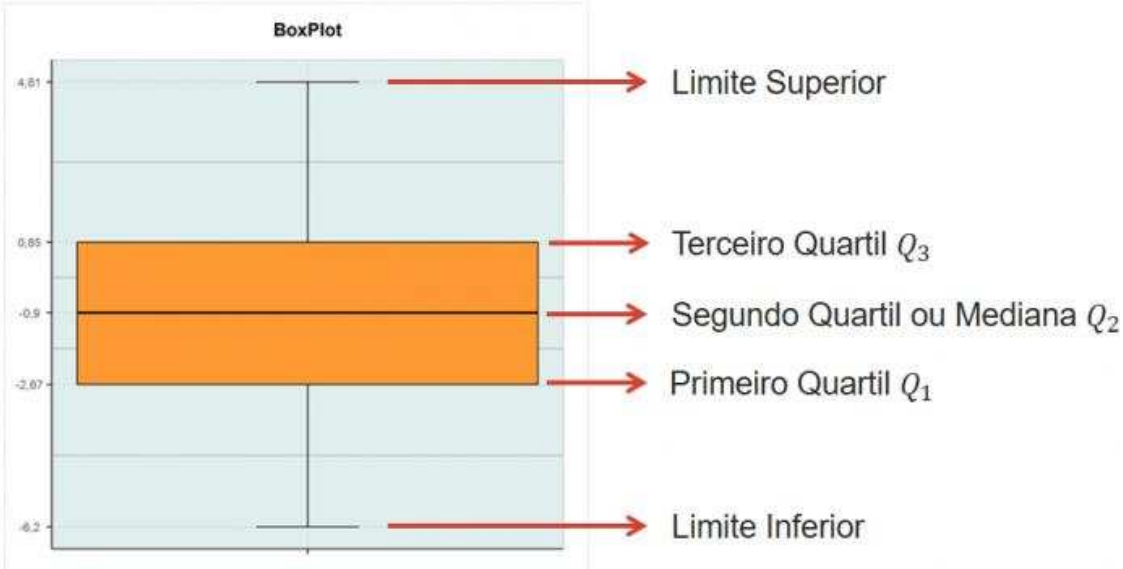

Figura 1 - Apresentação da composição de um Box Plot
Fonte: Portal Action (2016)

Avaliar a qualidade das observações é a primeira etapa de uma análise de dados, o que justifica classificar como crítica qualquer ação posterior.

Muñoz-Garcia et al. (1990) afirmam que é preciso avançar metodologicamente, pois há claramente uma subdivisão na detecção de outliers, detecção individual (em que cada observação precisa ser avaliada) e detecção por valor nominal (em que um valor nominal estabelece um limite de aceitação).

Logo, percebe-se a necessidade da utilização e/ou criação de metodologias que identifiquem individualmente (incorporando ainda a localização espacial) as observações que apresentam uma certa discrepância em relação aos demais valores amostrados.

Entre as metodologias com potencial para o estudo de outliers e necessidade de incorporação da localização geográfica ao processo analítico dos dados, destaca-se a Geoestatística. Tal escolha está baseada em fundamentos históricos e teóricos. Desde o seu surgimento histórico, em 1951, o engenheiro de minas Daniel Gerhardus Krige e o estatístico Herbert Simon Sichel, ao estudarem dados de concentração de ouro, verificaram a existência de dados discrepantes caso não incorporassem mais informações sobre os dados (SILVA et al., 2008).

Segundo Santos et al. (2011) e Yamamoto e Landim (2013), para resolver este problema foi necessário criar um método que utilizasse toda a informação de localização espacial disponível sobre os dados, ponderando-a de forma que a variância de estimação fosse a menor possível. Esse método que garantia a variância mínima foi desenvolvido posteriormente, em 1963 e batizado por Georges Matheron pelo nome de krigagem (uma homenagem a Krige), surgindo então a parte teórica.

Dessa forma, a Geoestatística é uma metodologia que foi criada para a caracterização e modelagem do padrão espacial, além de interpolar sem tendência e com variância mínima.

Ferreira et al. (2013) mostram que, após a análise exploratória dos dados, o estudo do semivariograma é a primeira e mais importante etapa de uma análise geoestatística, pois desempenha o papel central de descrever tanto qualitativa quanto quantitativamente a variação espacial. Além disso, este é o ponto chave na predição geoestatística.

Segundo Isaaks e Srivastava (1989), o semivariograma é um gráfico construído por meio da função de semivariância versus cada valor h (em que h é a distância euclideana entre os pontos de coleta amostral), conforme Figura 2. Já a função de semivariância $\gamma(\mathrm{h})$ é definida como sendo a metade da esperança matemática do quadrado da diferença entre as realizações de duas variáveis localizadas no espaço, separada pelo vetor h, dado pela equação (1)

$$\gamma(h) = \frac{1}{2} E[z(x_i) - z(x_i + h)]^2 \quad (1)$$

em que: $z(x_i)$ é o valor observado da variável no ponto (x_i) e $z(x_i + h)$ é o valor observado no ponto $(x_i + h)$. Se a função $\gamma(h)$ depende da direção do vetor h então a dependência é dita ser anisotrópica, mas, se depende apenas do tamanho do vetor h, então a dependência é do tipo isotrópica – adotando apenas este tipo sem perda de generalidade, fazendo h ser símbolo para o tamanho do vetor, sem depender do sentido.

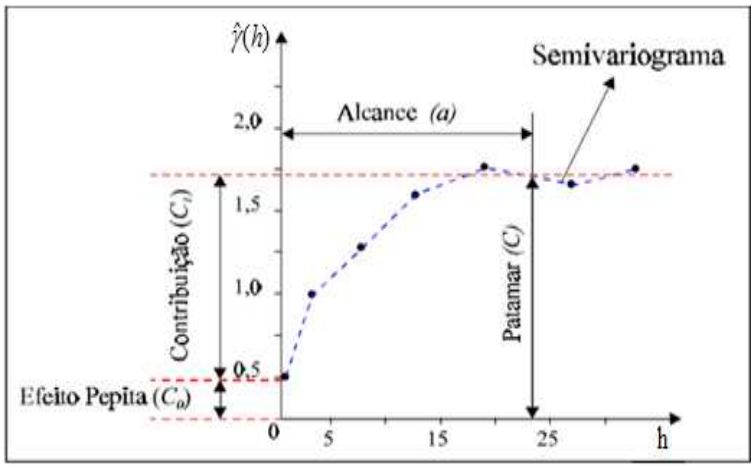

Figura 2 - Apresentação de um exemplo de um semivariograma
Fonte: Câmara e Medeiros (1998)

Segundo Vieira (2000), dentre os estimadores de semivariâncias o mais utilizado é o baseado no método de momentos, proposto por Matheron em 1963, e dado pela seguinte equação:

$$\hat{}(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} [z(x_i) \quad z(x_i + h)]^2 \quad (2)$$

em que: $\hat{}(h)$ é o valor estimado da semivariância na distância h e N(h) é o número de pares de pontos separados entre si por uma distância h.

O semivariograma possui parâmetros que descrevem a dependência espacial do fenômeno em estudo, conforme Figura 2, definidos como:

- Alcance (a): a distância dentro da qual os valores amostrais apresentam-se correlacionadas espacialmente. É o raio de dependência espacial.

- Patamar (C): o valor do semivariograma correspondente ao valor do alcance (a);
- Efeito Pepita (C_0): o valor da semivariância para a distância zero;
- Contribuição (C_1): a diferença entre os valores do patamar (C) e do Efeito Pepita (C_0).

Diferentes modelos teóricos podem ser ajustados a um semivariograma experimental. Contudo, os modelos teóricos mais utilizados são os modelos esférico, exponencial e gaussiano, apresentadas nas equações (3), (4) e (5), respectivamente.

$$\gamma(h) = \begin{cases} 0 & , \quad h = 0 \\ C_0 + C_1 \left[1{,}5\frac{h}{a} - 0{,}5 \left(\frac{h}{a}\right)^3 \right] & , \quad 0 < h < a \\ C_0 + C_1 & , \quad h \geq a \end{cases} \tag{4}$$

$$\gamma(h) = \begin{cases} 0 & , h = 0 \\ C_0 + C_1 \left[1 - \exp\left(-\frac{3h}{a}\right) \right] & , h \neq 0 \end{cases} \tag{5}$$

$$\gamma(h) = \begin{cases} 0 & , h = 0 \\ C_0 + C_1 \left[1 - \exp\left(-\frac{3h^2}{a^2}\right) \right] & , h \neq 0 \end{cases} \tag{6}$$

Uma etapa importante da análise geoestatística, utilizada para avaliar a qualidade de ajuste do modelo escolhido, é a validação cruzada. O modelo teórico que apresentar valores mais satisfatórios para as estatísticas dos resíduos obtidos deve ser escolhido para representar a dependência espacial do fenômeno em estudo (ISAAKS e SRIVASTAVA, 1989).

Assim, tendo sido feita a escolha do modelo do variograma, pode-se proceder à interpolação geoestatística denominada krigagem. Segundo Vieira (2000), esse método permite interpolar valores em qualquer posição no campo em estudo, sem tendência e com variância mínima, desde que exista dependência espacial entre as amostras e o semivariograma da variável seja conhecido.

Existem alguns tipos de krigagens, sendo a krigagem ordinária a mais comum por não exigir o conhecimento da média estacionária e sendo a krigagem simples a mais precisa dos preditores lineares, conforme apresentam e comparam Santos et al. (2011).

Conforme Figura 3, a krigagem interpola a variável em estudo para locais não observados, transformando a região em estudo em um mapa de isocores.

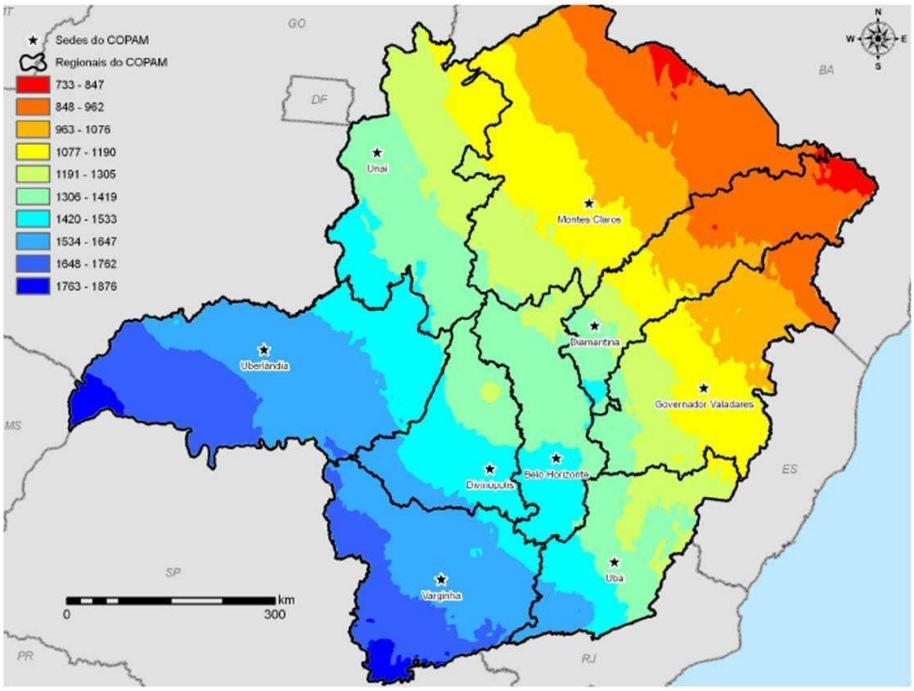

Figura 3 - Mapa de isocores das classes de precipitação pluvial total média anual (mm) para o Estado de Minas Gerais - ZEE/MG
Fonte: Scolforo et al. (2008)

Objetiva-se com este trabalho apresentar a relevância da adoção de metodologias adequadas no tratamento de dados discrepantes em variáveis geoespaciais contínuas; propor dois novos métodos de detecção e identificação de outliers, para variáveis dessa natureza, comparando-os com o tradicional método de Box Plot; e, ainda, propor um critério de decisão para a exclusão ou não dos outliers de uma amostra georreferenciada com dependência espacial. A estrutura teórica que se pretende adotar no trabalho é a combinação de teoremas da Estatística Clássica e o uso da Geoestatística como metodologia de apoio.

Especificamente, apresenta-se neste trabalho três estudos sobre a importância da proposição de métodos de detecção de outliers que incorporam a localização dos dados. No primeiro capítulo propõe-se um novo método de detecção de outliers para dados geoespaciais contínuos, em que um conjunto de dados reais, sabidamente com outliers, foi analisado tanto pelo Box Plot quanto pelo método em proposição, comparando a

eficácia de ambos quanto à detecção dos outliers. No segundo capítulo propõe-se um novo método de detecção de outliers para dados geoespaciais contínuos, cujas variáveis são não-negativas, em que, utilizando um conjunto de dados reais tanto o Box Plot quanto o método proposto têm a eficácia de detecção de outliers comparada. Finalmente, no terceiro capítulo propôe-se um mecanismo metodológico para a decisão de exclusão dos dados com alta probabilidade de discrepância. Neste capítulo são utilizados quatro conjuntos de dados, sendo três simulados computacionalmente e um conjunto de dados reais.

REFERÊNCIAS BIBLIOGRÁFICAS

ANSCOMBE, F.J. **Rejection of outliers.** Technometrics 20 (1960): 123-147.

APPICE, A., GUCCIONE, P., MALERBA, D., & CIAMPI, A. **Dealing. with temporal and spatial correlations to classify outliers in geophysical data streams.** Information Science 285 (2014): 162-80.

BARNETT, V., & LEWIS, T. **Outliers in statistical data.** Biometrical Journal 379 (1994): 256.

BARUA, S., & ALHAJJ, R. **High performance computing for spatial outliers detection using parallel wavelet transform.** Intelligent Data Analysis 11 (2007): 707-730.

BECKMAN, R.J., & COOK, R.D. **Outliers.** Technometrics 25 (1983): 119-149.

CÂMARA, G.; MEDEIROS, J.S. **Geoprocessamento para projetos ambientais.** V.1, Online Book, 1998. São José dos Campos, Brasil. INPE.

FERREIRA, D.F. **Estatística básica**. Lavras, Editora UFLA (2009)

FERREIRA, I.O., SANTOS, G.R., & RODRIGUES, D.D. (2013). **Estudo sobre a utilização adequada da krigagem na representação computacional de superfícies batimétricas.** Revista Brasileira de Cartografia 65 (2013): 831-842.

GRUBBS, F.E. **Procedures for detecting outlying observations in samples**. Technometrics 11 (1969): 1-21.

HONGXING, L., KENNETCH, C.J., & MORTON, E.O. **Detecting outliers in irregularly distributed spatial data sets by locally adaptive and robust statistical analysis and GIS.** International Journal Geographical Information Science 15 (2001): 721-741.

ISAAKS E.H., SRIVASTAVA R.M. **An Introduction to Applied Geostatistics.** N.Y OUP (1989). 561p.

MORETTIN, P.A.; TOLOI, C.M.C; **Análise de séries temporais.** Edgard Blucher (2006). 2 ed. 538p.

MUÑOZ-GARCIA, J., MORENO-REBOLLO, J.L., & PASCUAL-ACOSTA, A. **Outliers: a formal approach.** International Statistical Review 58 (1990): 215-226.

PORTAL ACTION. Estatcamp - Consultoria Estatística e Qualidade. <http://www.portalaction.com.br/>. Último acesso em: 26.01.2017

QIAO, C., HAIBO, H., & HONG, M. **Spatial outlier detection based on iterative self-organizing learning model**. Neurocomputing 117 (2013): 161-172.

ROUSSEUW, P.J., & ZOMEREN, B.C. **Unmasking multivariate outliers and leverage points.** Journal of the American Statistical Association 85 (1990): 633-639.

SANTOS, A. P**. Avaliação da Acurácia Posicional em Dados Espaciais com o uso de Estatística Espacial**. 110p. (Magister Scientiae). Universidade Federal de Viçosa, Viçosa, Minas Gerais, Brasil. 2010.

SANTOS, G.R., OLIVEIRA, M.S, & SANTOS, A.M.R.T. **Krigagem simples versus krigagem universal: qual o preditor mais preciso?** Revista Energia na Agricultura 26 (2011): 49-55.

SCOLFORO, J. R. S.; OLIVEIRA, A. D.; CARVALHO, L. M. T. **Zoneamento ecológico-econômicodoestado de Minas Gerais: componentes geofísico e biótico.** Lavras: UFLA, 2008. 161 p.

SILVA, S.A., LIMA, J.S.S., SOUZA, G.S., & OLIVEIRA, R.B. **Avaliação de interpoladores estatísticos e determinísticos na estimativa de atributos do solo em agricultura de precisão.** Idesia 26 (2008): 75-81

TUKEY, J.W. **Exploratory Data Analysis**. Princeton, Ed. Pearson (1977)

YAMAMOTO, J., & LANDIM, P. **Geoestatística: Conceitos e Aplicações.** São Paulo, Oficina de Textos, (2013)

VIEIRA, S.R. **Geoestatística em estudos de variabilidade espacial do solo.** Tópicos em Ciências do Solo 1(2000): 1-54.

CAPÍTULO 1

DETECÇÃO DE INCONSISTÊNCIAS EM VARIÁVEIS GEOESPACIAIS UTILIZANDO GEOESTATÍSTICA

RESUMO

Provavelmente todo pesquisador já tenha se deparado com um conjunto de dados em que algumas observações se "afastam" das demais, sugerindo, inclusive, que o mecanismo gerador dos dados não é o mesmo, tornando tais informações como pertencentes a outra população. Essas observações são consideradas como inconsistentes, comumente denominadas outliers, ou dados discrepantes. Dessa forma, o objetivo desse trabalho é propor um novo método de detecção de dados inconsistentes para dados geoespaciais contínuos baseado na Geoestatística, independente da causa geradora das inconsistências (erros de medição, erros de execução e variabilidade inerente aos dados). A escolha pela Geoestatística está baseada em suas características ideais, pois adota um procedimento similar às médias móveis com o objetivo de evitar erros sistemáticos. A importância da proposta de um método de detecção de dados inconsistentes está no fato de que parte dos métodos usados no tratamento de dados geoespaciais consideram pressuposições teóricas dificilmente atendidas e/ou verificadas. Utilizando um conjunto de dados de, aproximadamente, 192 mil informações, e 3 subamostras, acerca da altimetria da região de Treynor, estado de Iowa, Estados Unidos, foi possível detectar e mapear dados discrepantes dessa variável. Comparando a nova técnica de detecção de outliers com um método muito utilizado de detecção, o Box Plot, verificou-se a importância e funcionalidade do novo método, já que o Box Plot não detectou nenhum dado como discrepante. Já o método proposto apontou, em média, 1,2% dos dados de possíveis outliers inferiores regionalizados e, em média, 1,4% de possíveis outliers superiores regionalizados.

Palavras-chaves: Detecção de Outliers; Geoestatística; Dados Geoespaciais

ABSTRACT

Almost every researcher has already encountered a data set, that some observations “step away” from the others, suggesting that the generating mechanism of the data is not equal to the others, making these informations as belonging to another population. These observations are considered as inconsistent, commonly called outliers, or discrepant data. That way, the main point of this paper is the creation of a new method of inconsistent data detection for continuous geospatial data in Geostatistics, independent of the cause that generates the inconsistencies (mistakes of measuring systems, mistakes of execution and inherent variability data). The choice made by Geostatistics is based on its ideal characteristics, because it adopts an equal procedure to the mobile media with the main objective of avoiding systematic mistakes. The importance of the creation of a detection method of inconsistent data is the fact that part of the adopted methods on the geospatial treatment consider theoretical assumptions hardly attended and/or verified. Utilizing a data set of, approximately, 192 thousand informations, and 3 subsamples, about altimetry of the region of Treynor, state of Iowa, United States, it was possible to detect and map outliers of this variable. Comparing the new technic of outlier detection with a detection method that is commonly used, BoxPlot, it was verified the importance and the functionality of the new method, as the BoxPlot did not detect any outliers. The proposed method has detected, in average, 1,2% of the data of possible inferior regionalized outliers and, in average, 1,4% of possible superior regionalized outliers.

Keywords: Outliers detection; Geostatistics; Geospatial Data

1. INTRODUÇÃO

Muitos conjuntos de dados são compostos por algumas observações se "afastam" estatisticamente das demais, sugerindo, inclusive, estas pertencem à outra população. Essas observações são consideradas como inconsistentes, comumente denominadas outliers, ou dados discrepantes.

Muitos trabalhos surgiram com o objetivo de contribuírem com o estudo de outliers, como é o caso de Anscombe (1960), Grubbs (1969), Beckman e Cook (1983), Rousseuw e Zomeren (1990), Muñoz-Garcia et al. (1990), Barnett e Lewis (1994), entre outros pioneiros. Alguns destes autores afirmam que a preocupação com dados discrepantes é tão antiga quanto às primeiras tentativas de análises de um conjunto de dados, como é o caso dos comentários de Bernoulli de 1777 a cerca da existência de tais dados.

Recentemente, novas metodologias têm sido criadas para atender as demandas das diversas áreas do conhecimento científico, como é o caso de Hongxing et al. (2001) para dados espaciais distribuídos de forma irregular em malhas amostrais, Barua and Alhajj (2007) para processamento de imagens, Quiao et al. (2013) para dados de satélites e Appice et al. (2014) para fluxo de dados geofísicos.

Muñoz-Garcia et al. (1990) afirmam que estudar a detecção destas informações é importante porque uma das primeiras etapas de uma análise de dados é a avaliação da qualidade das observações. Apesar da consciência da importância da discussão aprofundada em relação à utilização ou não desses dados, é considerada crítica a fase de detecção de dados discrepantes, pois qualquer ação posterior pode ser reprovada.

A importância em se propor um método de detecção de dados inconsistentes está no fato de que uma boa parte dos já então adotados no tratamento de dados geoespaciais consideram pressuposições teóricas dificilmente atendidas e/ou verificadas, ou seja, segundo Mood et al. (1974), as variáveis devem ser independentes e identicamente distribuídas para um tratamento estatístico clássico. Conforme Yamamoto e Landim (2013) citam e referenciam, a maioria das variáveis provenientes de malhas amostrais planejadas são comprovadamente dependentes no espaço (não aleatórias) e de difícil comprovação quanto à distribuição, como é o caso de dados geoespaciais. Dessa forma,

uma nova metodologia que considere amostragem planejada (em forma de malhas regulares ou irregulares) e a estrutura de dependência espacial entre os valores observados, considerando esta estrutura em todo o processo analítico dos dados, é de grande valor científico.

Dessa forma, o objetivo desse trabalho é a proposição de um novo método de detecção de outliers para dados geoespaciais contínuos através da Geoestatística e teoremas da Estatística Clássica, independentemente da causa geradora das inconsistências (erros de medição, erros de execução e variabilidade inerente aos dados).

A escolha pela Geoestatística, como metodologia de apoio, está baseada em suas características ideais, pois desde o seu surgimento histórico, em 1951, o engenheiro de minas Daniel Gerhardus Krige e o estatístico Herbert Simon Sichel, ao estudarem dados de concentração de ouro, verificaram a existência de dados discrepantes, levando-os a adotarem um procedimento similar às médias móveis com o objetivo de evitar erros sistemáticos (SILVA et al., 2008).

Segundo Yamamoto e Landim (2013) e Santos et al. (2011), para resolver este problema foi necessário criar um método que utilizasse toda a informação disponível dos dados, ponderando-a de forma que a variância de estimação fosse a menor possível. Esse método que garante a variância mínima foi desenvolvido posteriormente, em 1961 e batizado por Georges Matheron pelo nome de krigagem (em inglês kriging), uma homenagem a Krige.

Dessa forma, escolheu-se uma metodologia, a Geoestatística, que segundo Vieira (2000) foi criada para interpolar sem tendência e com variância mínima, além de levar em consideração na modelagem e predição a estrutura de dependência espacial das amostras, o que é inerente de dados geoespaciais.

2. MATERIAL E MÉTODOS

Visando atingir os objetivos deste projeto descreve-se em seguida o que será adotado como metodologia.

2.1. Caracterização dos dados

A área de estudo é localizada no estado de Iowa, Estados Unidos, no condado de Pottawattamie, compreendendo uma porção de 34,32 hectares do município de Treynor. Delimita-se a região de estudos pelas latitudes 41°10’23”N e 41°09’53”N, e longitudes 95°38’24”W a 95°38’47”W, conforme mostrado pela Figura 1.

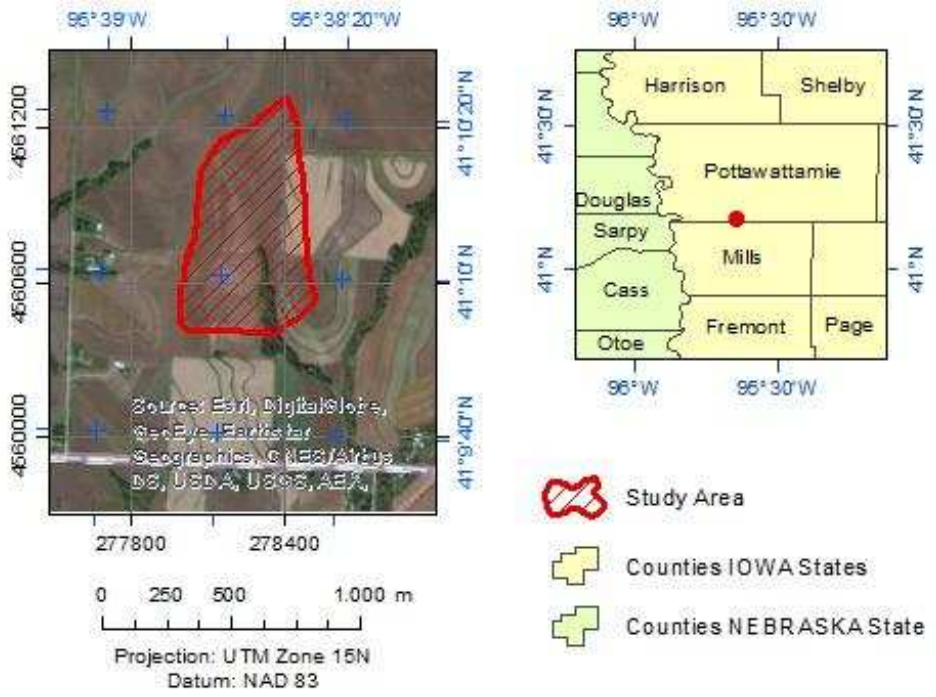

Figura 1 - Apresentação geral da área de estudo, localizada no estado de Iowa, Estados Unidos, no condado de Pottawattamie, compreendendo uma porção de 34,32 hectares do município de Treynor.

Conforme Höhle e Höhle (2009), para mapeamento de escalas grandes e médias, tem-se produzido Modelos Digitais de Elevação (MDE) utilizando principalmente a

tecnologia LiDAR (Light Detection And Ranging). Além de proporcionar alta densidade de pontos planialtimétricos o método supracitado mostra-se acurado e eficiente.

Os dados de altimetria utilizados neste trabalho, provenientes de um mapeamento LiDAR, são referenciados ao sistema geodésico NAD 83 (North American Datum of 1983) e representados na projeção UTM (Universal Transversa de Mercator) fuso 15N. Estes dados totalizam 192.079 mil pontos de altitudes conhecidas, respectivamente, com uma densidade de 0,55 pontos/m² e um espaçamento de aproximadamente 1,7 e 1,2 metros nas direções X e Y, respectivamente. Apresentam valores de altitudes mínimos de 340,8 metros e valores máximos de 385,5 metros.

Realizou-se a pesquisa com 4 conjuntos de dados, representados em 100% dos dados (192.079 pontos), 75% dos dados (144.059 pontos), 50% dos dados (96.040 pontos) e 25% dos dados (48.020 pontos).

2.2. Proposição do método

A proposição de um novo método de detecção de inconsistências para variáveis geoespaciais baseou-se nas pressuposições teóricas dos resíduos de uma modelagem estatística, segundo Rencher e Schaalje (2008). Tais resíduos são caracterizados como ruído branco, ou seja, em sua forma padronizada seguem uma distribuição de probabilidade gaussiana com média nula e variância unitária, em outras palavras, distribuição normal padrão, $\varepsilon''_p \sim Z(0;1)$, em que ε''_p são os resíduos padronizados, conforme VIEIRA (2000).

Buscando atender às pressuposições teóricas dos resíduos, foi adotada a análise geoestatística para os dados geoespaciais, seguindo as recomendações de Yamamoto e Landim (2013), Santos et al. (2011) e Vieira (2000), ou seja, uma metodologia que modela sem tendência e com variância mínima, levando em consideração a estrutura de dependência espacial das amostras.

Para obter os resíduos, conforme orientação de Druck et al. (2004), a variável regionalizada escolhida neste estudo foi decomposta de forma aditiva em três componentes: uma componente estrutural, associada à um valor médio constante ou à uma tendência constante; uma componente aleatória, espacialmente correlacionada; e uma componente residual, também chamada de ruído branco, ruído aleatório ou passeio

aleatório. Considerando o vetor de localização espacial **x**, que **x** representa a posição da variável em uma (**x**=x), duas (**x**=[x,y]) ou mais dimensões (**x**=[x,y,z,…]), a variável regionalizada Y, em **x**, também chamada de função aleatória, pode ser denotada como

$$Y(x) = \mu(x) + \varepsilon'(x) + \varepsilon'' \quad (1)$$

em que: $\mu(x)$ é a função determinística que descreve a componente estrutural de Y em **x**; $\varepsilon'(x)$ é o termo estocástico correlacionado localmente; ε'' é o ruído branco não correlacionado com distribuição normal com média zero e variância σ^2.

Dessa forma, adotando o entendimento da decomposição da equação (1), o que se pretendeu foi utilizar a metodologia geoestatística para analisar os dados geoespaciais com dependência espacial comprovada e caracterizada, e, assim, obter os resíduos dessa modelagem. Tal metodologia é composta por análise exploratória clássica, testes de pressuposições, análise exploratória espacial, análise variográfica, modelagem variográfica, autovalidação e krigagem (FERREIRA et al., 2013).

Adotar esta abordagem geoestatística para os dados geoespaciais significa considerar cada ponto amostral georeferenciado como uma variável aleatória, gerando assim uma função aleatória, ou comumente, processo estocástico (SANTOS et al., 2011; CRESSIE, 1993).

Para que esta metodologia tenha validade estatística é necessário, segundo Yamamoto e Landim (2013) e Vieira (2000), que a pressuposição de estacionariedade do variograma seja assumida, ou seja, que o variograma exista e que seja estacionário para a variável na área de estudo.

Adotando o variograma teórico, populacional, de Vieira (2000), ou seja, $2\gamma(h) = E\{[Y(x) - Y(x+h)]^2\}$, e o estimador citado por Kamimura et al. (2013), dado pela equação

$$\hat{\gamma}(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} \{[Y(x_i) - Y(x_i + h)]^2\} \quad (2)$$

em que: $N(h)$ é o número de pares de valores medidos em (x_i) e $(x_i + h)$; $Y(x_i)$ e $Y(x_i + h)$ representam todas as variáveis aleatórias separadas por um vetor h que geram as amostras e, consequentemente, o principal mecanismo de detecção de dependência

espacial da metodologia geoestatística, o variograma, um gráfico de $\hat{\gamma}(\mathrm{h})$ em função do vetor-distância **h**.

Após a análise geoestatística e consequente obtenção dos resíduos (vindos da autovalidação leave-one-out, ou seja, cada resíduo vem da diferença entre um valor observado e seu respectivo valor predito), as características do ruído branco devem ser testadas, a saber, independência, distribuição normal com média nula e variância constante. Com os resultados satisfatórios, a etapa seguinte é a estimação intervalar dos resíduos. Como os estimadores são considerados variáveis aleatórias, suas estimativas comumente são distintas do valor do parâmetro, ou seja, comumente se comete um erro de estimação. Por esta razão tornou-se necessária a construção de intervalos de confiança com probabilidade (1-α) (FERREIRA, 2009).

Para a estimação intervalar (IC - Intervalo de Confiança) adotou-se a distribuição normal padrão (Z(0, 1)) e nível de significância α de 1% (arbitrário). Assim, em outras palavras, desejou-se determinar o quanto estas estimativas dos resíduos são prováveis (1 - α) de confiança, em que $\alpha \in (0,1)$, conforme equação (3). O nível α é também chamado de a probabilidade de se cometer o erro tipo I, rejeitar uma hipótese nula (que neste caso será $\mathrm{H}_0 : \varepsilon''_{\mathrm{p_i}} = 0$) verdadeira (MOOD, et al., 1974; VIEIRA, 2000; CASELLA E BERGER, 2010).

$$P\left[\bar{x} - Z_\alpha \frac{s}{\sqrt{n}} < \mu < \bar{x} + Z_\alpha \frac{s}{\sqrt{n}}\right] = 1 - \alpha \qquad (3)$$

Assim, todos os valores que não pertencerem ao IC construído, sem viés, com variância mínima e levando em consideração a estrutura de dependência espacial, Silva (2012) mostra que tais valores tornam-se possíveis outliers. Estatisticamente, se $x_i \in IC_{(1-\alpha)}$ então x_i é ruído branco, caso contrário é um provável outlier.

Utilizando os recursos de georreferenciamento dos dados, pretendeu-se ainda apontar quantos, quais e onde estão os resíduos com alta probabilidade de serem outliers.

Para comparação e/ou validação do método serão realizadas comparações do novo método com um dos mais robustos, estatisticamente, e utilizados métodos de detecção atuais, o Box Plot (HOAGLIN et al., 1983).

Toda a parte inovadora da metodologia foi realizada por meio do software livre R (R Development Core Team, 2014), em que a análise geoestatística foi realizada através do pacote geoR, desenvolvido por Ribeiro Júnior e Diggle (2001). Contudo, para a análise de geoprocessamento, utilizou-se o Software Arcgis 10.2 como descrito: após georreferenciar e analisar os dados através da ferramenta Geostatistical Analyst, reduziu-se o tamanho amostral, ou seja, subamostrou-se os dados utilizando malhas amostrais irregulares gerando assim outras três amostras com 75%, 50% e 25% dos dados originais. Estes procedimentos foram feitos através do Toolbox Random Selection.

3. RESULTADOS E DISCUSSÃO

Buscando atingir os objetivos propostos neste trabalho, utilizou-se dos dados de altimetria (elevação) de uma bacia hidrográfica da região da Cidade de Treynor, estado de Iowa, Estados Unidos.

3.1. Caracterização dos dados

A amostragem completa contou com 192.079 pontos. Com a redução do tamanho da amostra em, aproximadamente, 75%, 50% e 25% do tamanho original, contou-se também com mais três conjuntos de dados com 144.059; 96.040 e 48.020 pontos, respectivamente.

Na Figura 2 tem-se as representações tridimensionais dos quatro conjuntos de dados da região de estudo.

Através do destaque das quatro imagens da Figura 2 é possível perceber uma alteração no relevo da região e o que, a princípio, de fato provocou tal alteração, ou seja, se esta alteração ocorre na região de estudo ou na coleta dos dados, gerando prováveis dados discrepantes em relação aos demais.

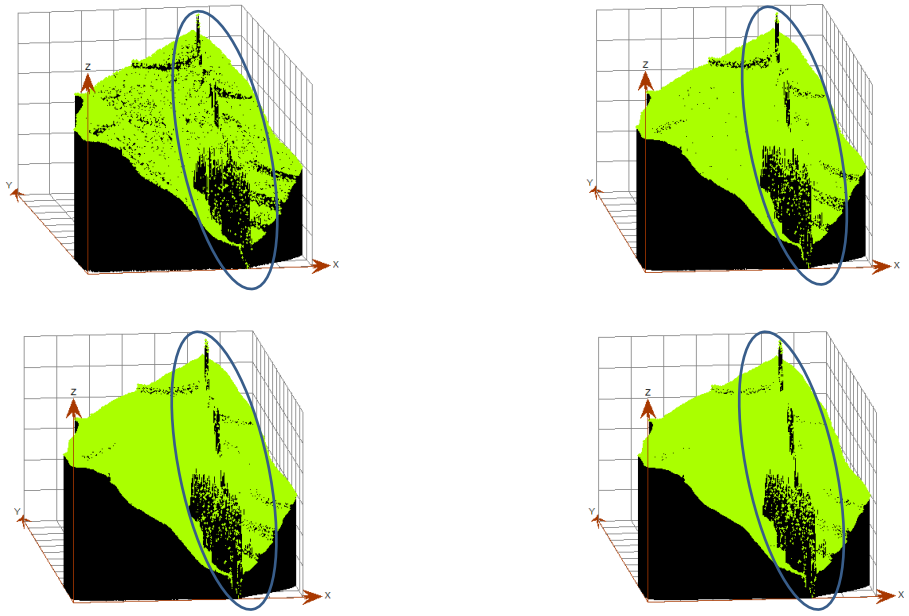

Figura 2 – Apresentação geral dos dados de altimetria obtidos pelo LiDAR Cloud de uma pequena bacia hidrográfica da região de Treynor Town - Iowa - USA, sendo 25% dos dados (superior esquerdo), 50% (superior direito), 75% (inferior esquerdo) e 100% (inferior direito).

3.2. Análise exploratória dos dados

Segundo Yamamoto e Landim (2013) e Ferreira et al. (2013), é importante verificar o comportamento espacial dos dados, ainda como análise exploratória.

Dessa forma, apresenta-se, na Figura 3, os gráficos de quartis (gráficos que utilizam o primeiro quartil, a mediana e o terceiro quartil como divisor de cores) dos quatro conjuntos de dados.

Conforme Figura 3, para os 4 tamanhos amostrais, a cor vermelha refere-se aos dados de maior valor, superiores ao terceiro quartil, presentes nas extremidades das imagens. Em seguida, percebe-se que os valores imediatamente menores, representados pela cor amarela, são os valores que se apresentam entre a mediana e o terceiro quartil. Após estes, estão os valores do segundo quartil, representado pela cor verde e, finalmente, os valores do primeiro quartil, ou seja, menores valores de altimetria, representados pela cor azul nas posições mais centrais das imagens. Nesta forma de visualização dos dados, possíveis valores discrepantes também estão perceptíveis, pois entre os valores de menor altimetria aparecem valores pertencentes a outro quartil.

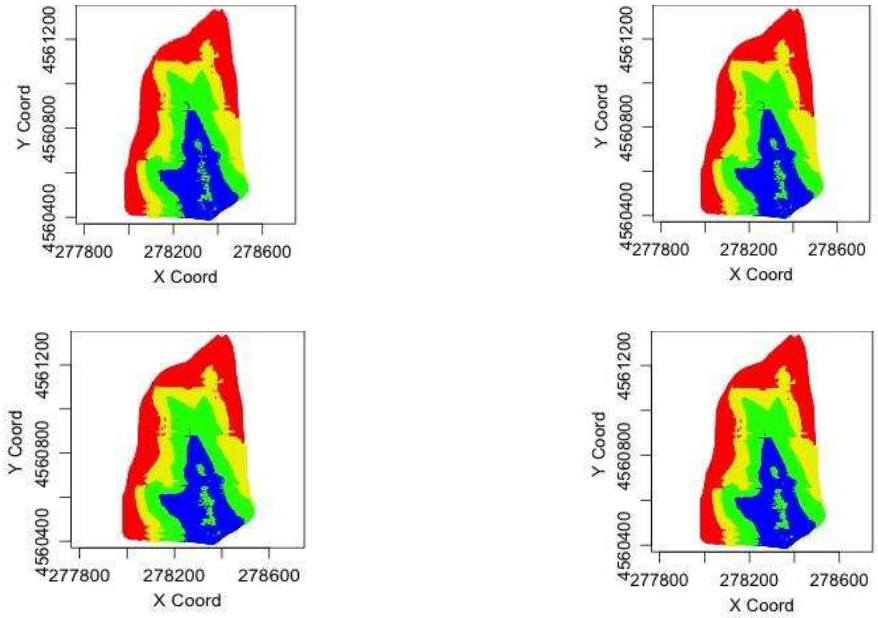

Figura 3 – Apresentação do comportamento dos dados quanto à "intensidade" da altimetria de uma pequena bacia hidrográfica da região de Treynor Town - Iowa - USA, utilizando os gráficos de quartis para 25% dos dados (superior esquerdo), 50% (superior direito), 75% (inferior esquerdo) e 100% (inferior direito).

No intuito de se constatar a existência de dados discrepantes em toda a área de estudo, principalmente, na parte mais central das imagens, utiliza-se o recurso do Box Plot, conforme Figura 4.

Através da Figura 4 pode-se perceber que com esta metodologia de detecção de outliers não foi possível detectar os dados inconsistentes percebidos através das Figuras 2 e 3. Apesar da robustez do BoxPlot, segundo Tukey (1977) e Benjamini (1988), aparentemente o conjunto de dados não apresenta discrepâncias.

3.3. Análise Geoestatística dos dados

Conforme descrito na metodologia do trabalho, para obter os resíduos com propriedades que atendam as pressuposições teóricas da Estatística, os dados foram analisados pela Geoestatística, conforme Tabela 1.

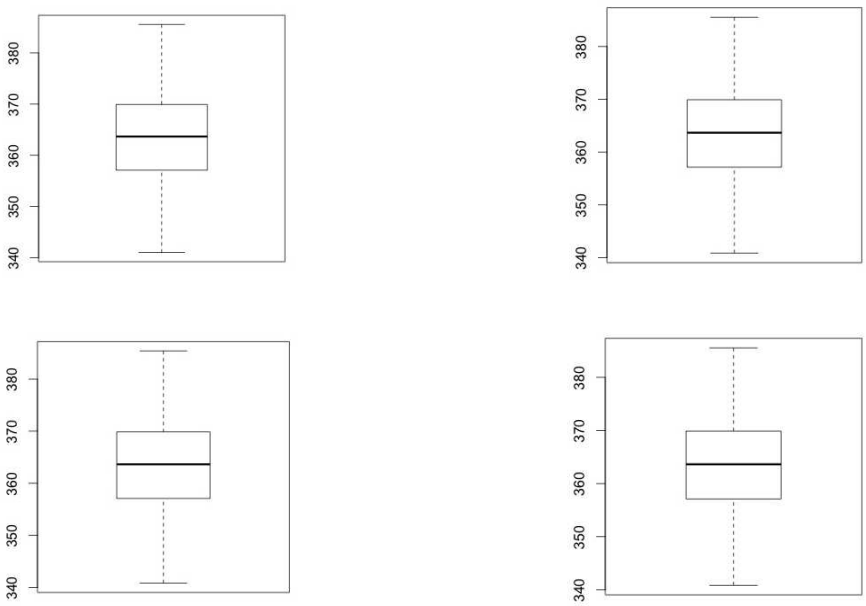

Figura 4 – Apresentação do Box Plot dos dados de altimetria de uma pequena bacia hidrográfica da região de Treynor Town - Iowa - USA para 25% dos dados (superior esquerdo), 50% (superior direito), 75% (inferior esquerdo) e 100% (inferior direito).

Tabela 1 – Principais informações da análise geoestatística dos dados de altimetria.

Medida/Característica	**Estimativas**			
Tamanho da Amostra (%)	25	50	75	100
Média (m)	363,27	363,67	363,26	363,26
Variância (m^2)	62,73	62,25	62,25	62,32
Desvio Padrão (m)	7,92	7,89	7,89	7,89
Anisotropia	Não	Não	Não	Não
Modelo	Gaussiano	Gaussiano	Gaussiano	Gaussiano
Efeito Pepita (m^2)	2,78	2,19	3,02	3,01
Contribuição (m^2)	67,37	62,19	69,57	67,04
Alcance (m)	336,75	310,80	350,87	339,39

Analisando a Tabela 1, pode-se perceber que as principais medidas descritivas da análise foram preservadas nos 3 conjuntos de dados provenientes da amostra completa, a saber: média, variância, desvio-padrão, isotropia e o modelo do semivariograma. Destaca-se também a pequena variação das estimativas dos parâmetros do variograma.

Após a análise variográfica foi possível representar o comportamento da altimetria em toda a região estudada, através da interpolação via krigagem simples, Figura 5, conforme recomenda Santos et al. (2011).

Como a krigagem interpola levando em consideração a estrutura de dependência espacial caracterizada pelos pontos da vizinhança, pode-se perceber que a Figura 5 apresenta a mesma característica percebida de possíveis dados discrepantes. Dessa forma, o mapa de krigagem representa bem o conjunto de dados, mas pode não representar o que fato acontece na região estudada, se for constatada a presença de outliers.

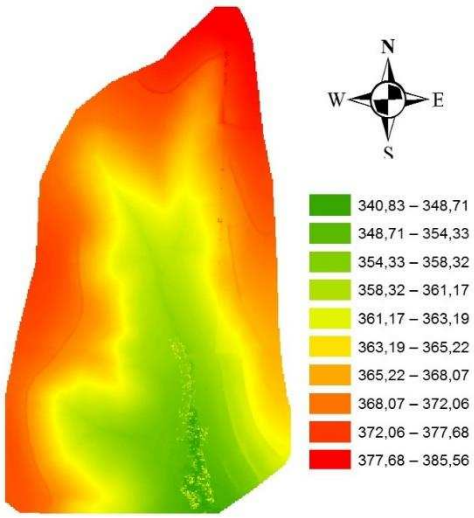

Figura 5 – Krigagem simples dos dados de altimetria obtidos pelo LiDAR Cloud de uma pequena bacia hidrográfica da região de Treynor Town - Iowa - USA.

3.4. Detecção de Outliers via análise de resíduos

Os resíduos provenientes de uma modelagem acurada apresentam características importantes que devem ser testadas, visando a aplicação de metodologias condicionais. Entre elas, destaca-se a independência (testada através da obtenção do efeito pepita-puro no variograma empírico), normalidade (testada através do teste Shapiro and Wilk, 1965) com média nula e variância unitária (para os resíduos padronizados). Todas essas

pressuposições foram constatadas nos resíduos obtidos, logo, os mesmos são classificados como ruído branco (MOOD, et al., 1974).

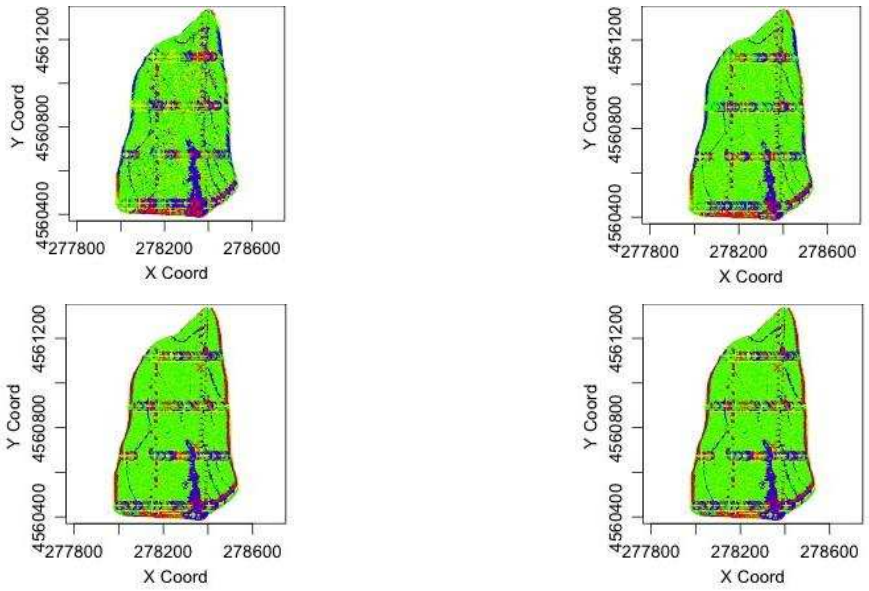

Figura 6 – Apresentação do comportamento de homogeneidade espacial dos resíduos provenientes da autovalidação de uma análise geoestatística, utilizando os gráficos de quartis para 25% dos dados (superior esquerdo), 50% (superior direito), 75% (inferior esquerdo) e 100% (inferior direito).

O comportamento na Figura 6 difere daquele da Figura 3 devido à característica de homogeneidade espacial dos resíduos, contudo a Figura 6 ainda apresenta uma aglomeração (em cor azul na parte central-sul da região) que evidencia o problema de inconsistência.

Assim, como recomenda Ferreira (2009), uma vez que os resíduos apresentam todos os requisitos exigidos pelas pressuposições estatísticas, passa-se à estimação intervalar de 99% de probabilidade, conforme resultados apresentados na Tabela 2.

Tabela 2 – Estimação intervalar a 99% de confiança para os resíduos provenientes da autovalidação de uma análise geoestatística dos dados de altimetria.

Tamanho da Amostra (%)	$IC_{(99\%)}$ [Mín ; Máx] (m)
25	[-14,21; 14,22]
50	[-12,44; 12,44]
75	[-19,45; 19,48]
100	[-21,33; 21,37]

Como os intervalos estimados foram bilaterais, dois tipos de dados discrepantes foram detectados, superiores e inferiores.

Apresenta-se, na Figura 7, os dados com alta probabilidade de serem outliers inferiores. Em outras palavras, estes possíveis outliers vieram da diferença entre os dados observados e os preditos, e resultaram em valores negativos e distantes do esperado teoricamente. Com o uso da Geoestatística, foi possível também apontar a localização desses dados.

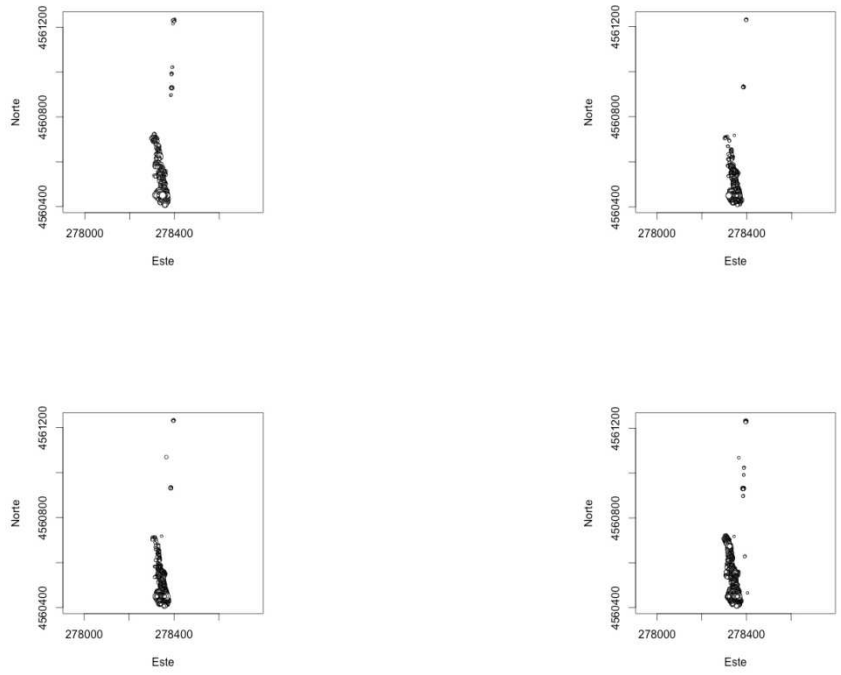

Figura 7 – Apresentação dos valores inferiores com alta probabilidade de serem outliers, para 25% dos dados (superior esquerdo), 50% (superior direito), 75% (inferior esquerdo) e 100% (inferior direito).

Apresenta-se também os dados com alta probabilidade de serem outliers superiores na Figura 8. Em outras palavras, estes possíveis outliers vieram da diferença entre os dados observados e os preditos, e resultaram em valores positivos e distantes do esperado teoricamente. Com o uso da Geoestatística, foi possível também apontar a localização desses dados.

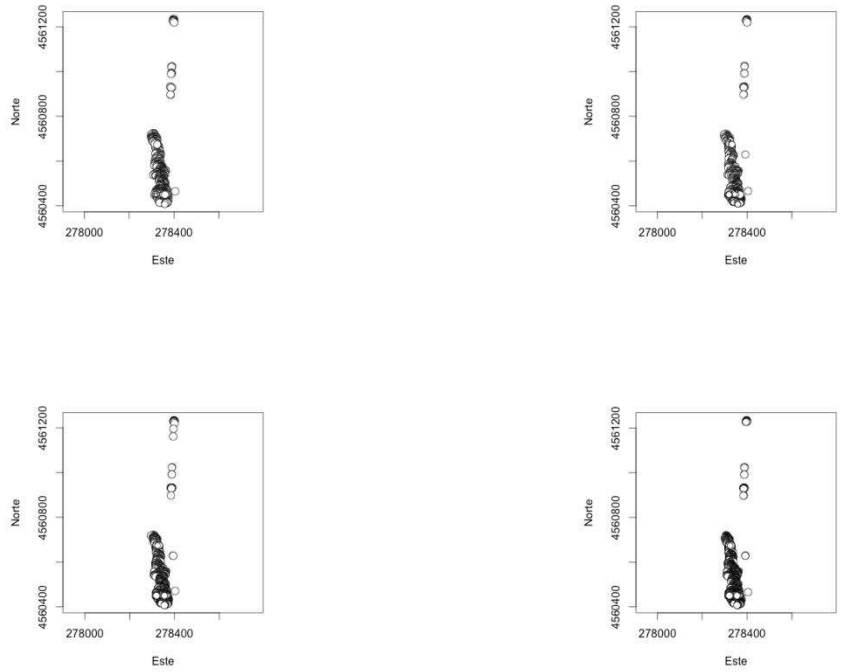

Figura 8 – Apresentação dos valores superiores com alta probabilidade de serem outliers, para 25% dos dados (superior esquerdo), 50% (superior direito), 75% (inferior esquerdo) e 100% (inferior direito).

Resumidamente, apresenta-se o percentual dos dados com alta probabilidade de serem outliers na Tabela 3.

Tabela 3 – Resumo do percentual dos dados com alta probabilidade de serem considerados outliers para a altimetria para os 4 conjuntos de dados do trabalho.

Tamanho da Amostra (%)	Inferiores (%)	Superiores (%)	Total (%)
25	1,67	1,72	3,39
50	1,03	1,22	2,25
75	1,07	1,23	2,30
100	1,12	1,33	2,45

Pode-se perceber, pelos resultados apresentados na Tabela 3 que, mesmo diante da redução significativa do tamanho do conjunto de dados, o percentual de outliers detectados variam de 2,25% a 3,39%. Ainda nesse sentido, avaliando os resultados das figuras desse trabalho, pode-se notar que desde o início da análise geoestatística foi

possível detectar regiões em que os dados eram discrepantes em relação aos demais, o que comprova a importância da adoção metodológica da estatística espacial em dados dessa natureza.

Acerca desta importância, Vieira (2000), Ferreira et al. (2013) e Yamamoto e Landim (2013) afirmam e/ou mostram a importância da utilização da Estatística Espacial para dados geoespaciais, não ignorando a Estatística Clássica.

4. CONCLUSÕES

Conjuntos de dados em que alguns valores diferem dos demais a ponto de gerarem conclusões equivocadas sobre a própria amostra recolhida e a população de origem dos dados são geralmente chamados de outliers e todos os tipos de estudos estão sujeitos à ocorrência dos mesmos.

Conforme visto, independente da causa geradora dessas inconsistências (erros de medição, erros de execução, variabilidade inerente dos dados, entre outros) e do tipo de variáveis em estudo (georeferenciada ou não, univariada ou multivariada) é preciso adotar corretas metodologias de análises, e não adotar metodologias gerais, pois o conjunto de pressuposições teóricas pode diferir fortemente.

Neste trabalho foi adotada a análise geoestatística para variáveis geoespaciais contínuos visando uma modelagem ótima, estatisticamente, de tais dados para que os resíduos atendessem às pressuposições teóricas esperadas, a saber, homogeneidade espacial, independência e normalidade.

Utilizando um conjunto de dados de, aproximadamente, 192 mil informações, e 3 subamostras, acerca da altimetria da região de Treynor, estado de Iowa, Estados Unidos, e intervalos de confiança a 99% de certeza para os resíduos, que atenderam às condições teóricas, foi possível detectar e mapear dados discrepantes dessa variável.

Comparando a metodologia proposta de detecção de outliers com um método muito utilizado de detecção, o Box Plot, verificou-se a importância e funcionalidade do novo método, já que o Box Plot não detectou nenhum dado como discrepante.

Como recomendação para trabalhos futuros, sugere-se avançar nesse método de detecção criando soluções para variáveis em que o intervalo de confiança dos resíduos não pode admitir a bilateralidade do mesmo.

5. REFERÊNCIAS BIBLIOGRÁFICAS

ANSCOMBE, F.J. **Rejection of outliers.** Technometrics 20 (1960): 123-147.

APPICE, A., GUCCIONE, P., MALERBA, D., & CIAMPI, A. **Dealing. with temporal and spatial correlations to classify outliers in geophysical data streams.** Information Science 285 (2014): 162-80.

BARNETT, V., & LEWIS, T. **Outliers in statistical data**. Biometrical Journal 379 (1994): 256.

BARUA, S., & ALHAJJ, R**. High performance computing for spatial outliers detection using parallel wavelet transform**. Intelligent Data Analysis 11 (2007): 707-730.

BECKMAN, R.J., & COOK, R.D. **Outliers.** Technometrics 25 (1983): 119-149.

BENJAMINI, Y., & ADDISON, W. **Opening the Box of a Boxplot.** Journal of the American Statistical Association 42 (1988): 257-262.

CASELLA, G., BERGER, R.L**. Inferência estatística.** Cengage Learning (2010).

CRESSIE, N. **Statistics for spatial data.** Wiley-Interscience (1993).

DRUCK, S., CARVALHO, M.S., CÂMARA, G., & MONTEIRO, A.V.M., (Ed.) **Análise Espacial de Dados Geográficos.** Brasília, Embrapa (2004).

ENVIRONMENTAL SYSTEMS RESEARCH INSTITUTE – **Esri. ArcGIS Desktop: Release 10**. Redlands, CA: 2011.

FERREIRA, D.F. **Estatística básica.** Lavras, Editora UFLA (2009)

FERREIRA, I.O., SANTOS, G.R., & RODRIGUES, D.D. **Estudo sobre a utilização adequada da krigagem na representação computacional de superfícies batimétricas.** Revista Brasileira de Cartografia 65 (2013): 831-842.

GRUBBS, F.E**. Procedures for detecting outlying observations in samples.** Technometrics 11 (1969): 1-21.

HOAGLIN, D.C, MOSTELLER, F., & TUKEY, J.W. **Understanding robust and exploratory data analysis.** New York, J. Wiley (1983)

HÖHLE, J., & HÖHLE, M. **Accuracy assessment of digital elevation models by means of robust statistical methods.** ISPRS Journal of Photogrammetry and Remote Sensing 64 (2009): 398-406.

HONGXING, L., KENNETCH, C.J., & MORTON, E.O. **Detecting outliers in irregularly distributed spatial data sets by locally adaptive and robust statistical analysis and GIS**. International Journal Geographical Information Science 15 (2001): 721-741.

KAMIMURA, K.M., SANTOS, G.R., OLIVEIRA, M.S., DIAS, JR., M.S., & GUIMARÃES, P.T.G. **Variabilidade espacial de atributos físicos de um Latossolo**

Vermelho-Amarelo sob lavoura cafeeira. Revista Brasileira de Ciência do Solo 37 (2013): 877-888.

MOOD, A.M., GRAYBILL, F.A., & BOES, D.C. **Introduction to the theory of statistics**. Kogakusha, McGraw-Hill, (1974)

MUÑOZ-GARCIA, J., MORENO-REBOLLO, J.L., & PASCUAL-ACOSTA, A. **Outliers: a formal approach.** International Statistical Review 58 (1990): 215-226.

QIAO, C., HAIBO, H., & HONG, M. **Spatial outlier detection based on iterative self-organizing learning model.** Neurocomputing 117 (2013): 161-172.

R CORE TEAM. R: **a language and environment for statistical computing.** R Foundation for Statistical Computing, (2014) Vienna, W. Recuperado de http://www.Rproject.org/.

Rencher, A.C., & Schaalje, G.B**. Linear Models in Statistics.** New Jersey, John Wiley & Sons, (2008)

RIBEIRO, J.P.J, & DIGGLE, P.J. **GeoR: a package for geostatistical analysis.** R-News. 1: 15-18.

ROUSSEUW, P.J., & ZOMEREN, B.C. **Unmasking multivariate outliers and leverage points.** Journal of the American Statistical Association 85 (1990): 633-639.

SANTOS, G.R., OLIVEIRA**, M.S, & SANTOS, A.M.R.T. Krigagem** simples versus krigagem universal: qual o preditor mais preciso? Revista Energia na Agricultura 26 (2011): 49-55.

SILVA, S.A., LIMA, J.S.S., SOUZA, G.S., & OLIVEIRA, R.B. **Avaliação de interpoladores estatísticos e determinísticos na estimativa de atributos do solo em agricultura de precisão.** Idesia 26 (2008): 75-81.

SILVA, A. N. **Detecção de outliers em séries espaço-temporais: análise de precipitação em Minas Gerais.** Viçosa, Minas Gerais, 2012.

TUKEY, J.W. **Exploratory Data Analysis.** Princeton, Ed. Pearson (1977)

YAMAMOTO, J., & LANDIM, P. **Geoestatística: Conceitos e Aplicações**. São Paulo, Oficina de Textos, (2013)

VIEIRA, S.R. **Geoestatística em estudos de variabilidade espacial do solo.** Tópicos em Ciências do Solo 1(2000): 1-54.

CAPÍTULO 2

PROPOSIÇÃO DE UM MÉTODO DE DETECÇÃO DE OUTLIERS PARA VARIÁVEIS NÃO-NEGATIVAS UTILIZANDO GEOESTATÍSTICA

RESUMO

Os outliers, ou dados discrepantes, são aqueles que devido à seu comportamento extremo em relação aos demais, podem ser caracterizados como observações anômalas. Dessa forma, objetiva-se com este trabalho a criação de uma proposta de detecção de outliers para dados não-negativos utilizando a robustez teórica da Geoestatística e dos resíduos provenientes da modelagem via BLUP, ou seja, utilizando teoremas estatísticos relacionados aos resíduos provenientes da autovalidação. A importância de um novo método de detecção de outliers para dados não-negativos está no fato de que algumas variáveis geoespaciais contínuas não consideram a presença de dados discrepantes menores que zero (negativos). Utilizando um conjunto de dados de, aproximadamente, 31 mil observações, e 3 subamostras, relacionado à valores do índice NDVI (Normalized Difference Vegetation Index) de áreas com irrigação agrícola por pivô central, dos municípios de Luiz Eduardo Magalhães, Barreiras e Riachão das Neves, estado da Bahia – Brasil, foram detectados, aproximadamente, 1% dos valores observados de NDVI considerados como probabilisticamente outliers. Além da detecção dos outliers, pela metodologia adotada é possível estabelecer também a localização dos mesmos.

Palavras-chave: Dados discrepantes, Geoestatística, NDVI, variáveis não-negativas.

ABSTRACT

Outliers, or discrepant informations, are those that due to their extreme behavior in relation to the others, can be characterized as anomalous observations. Thus, it is aimed with this paper the creation of a proposal of outlier detection method for nonnegative data utilizing the theoretical methodology named Geostatistics and the residuals from the modeling via BLUP, in other words, utilizing statistical theorems applied to the residuals from the geostatistical cross-validation. The importance of a new outlier detection method for nonnegative data is in the fact that some geospatial continuous variables do not consider the presence of discrepant data smaller than zero (negative). Utilizing a data set of, approximately, 31 thousand observations, and 3 subsamples, related to values of the index NDVI (Normalized Difference Vegetation Index) of areas with agricultural irrigation by center pivot, of the counties of Luiz Eduardo Magalhães, Barreiras e Riachão das Neves, state of Bahia - Brazil, there was detected, approximately, 1% of the observed values of NDVI considered as probably outliers. Besides the detection of the outliers, through the adopted methodology it is also possible to establish their location.

Keywords: Discrepant data, Geostatistics, NDVI Non-negative variables.

1. INTRODUÇÃO

Os outliers, ou dados discrepantes, são dados que podem ser classificados como pertencentes a outra população devido seu comportamento extremo em relação aos demais. Dessa forma, na maioria das vezes, quando são detectados outliers em um conjunto de dados, as recomendações em relação ao tratamento desse tipo de dado é a eliminação dos mesmos dos dados observados, a fim de que as estimativas estatísticas não sejam influenciadas (FERREIRA, 2009).

A recomendação de eliminação dos dados, tidos como outliers, deve ser precedida por metodologias adequadas de detecção. Apesar dos prejuízos analíticos que a presença desses dados podem provocar em qualquer estudo estatístico, a eliminação também pode omitir informações preciosas (QIAO et al., 2013).

Trabalhos como os de Anscombe (1960), Grubbs (1969), Beckman e Cook (1983), Rousseuw e Zomeren (1990), Muñoz-Garcia et al. (1990), Barnett e Lewis (1994), entre outros pioneiros, contribuíram com a discussão sobre essa temática, mostrando que a preocupação com esse tipo de dados é muito antiga, pois data de 1777, com os estudos do pesquisador Bernoulli.

Trabalhos mais recentes com novas metodologias têm surgido para atender às demandas das diversas áreas do conhecimento científico, como é o caso de Hongxing, Kennetch and Morton (2001) para dados espaciais distribuídos de forma irregular em malhas amostrais, Barua e Alhajj (2007) para processamento de imagens, Qiao et al. (2013) para dados de satélites e Appice et al. (2014) para fluxo de dados geofísicos.

Além destes, foi apresentada no Capítulo 1 deste trabalho a utilização da Geoestatística para modelar a dependência espacial de dados altimétricos e utilizar os resíduos para a criação de uma metodologia eficiente de detecção de dados discrepantes, utilizando teorias robustas da Estatística.

A escolha pela Geoestatística, como metodologia de modelagem de uma dada variável, está baseada em suas características ideais, as quais foram apresentadas por Vieira (2000), Santos et al. (2011), Yamamoto e Landim (2013) e no Capítulo 1 deste trabalho, também apresenta algumas características interessantes, do ponto de vista preditivo, descrevendo-as como, interpolação sem tendência e com variância mínima,

além de levar em consideração na modelagem e predição a estrutura de dependência espacial das amostras, o que é inerente à variável em estudo.

A importância da criação de um novo método de detecção de outliers para dados não-negativos está no fato de que algumas variáveis geoespaciais contínuas não consideram a presença de dados discrepantes menores que zero (negativos). Além disso, tais variáveis apresentam dependência espacial, o que exige um tratamento estatístico diferenciado, conforme mostram Mood et al. (1974), Yamamoto e Landim (2013).

Assim, objetiva-se com este trabalho a proposta de detecção de outliers para dados de variáveis não-negativas utilizando a robustez teórica da Geoestatística e dos resíduos provenientes da modelagem via BLUP, ou seja, utilizando a aplicação de teoremas estatísticos aos resíduos provenientes da diferença entre os dados observados e dados vindos da modelagem geoestatística (preditos).

2. MATERIAL E MÉTODOS

2.1. Caracterização da área de estudo

A área do estudo do trabalho está localizada na região do Extremo Oeste da Bahia, entre as latitudes 11º36'S e 11º53'S e as longitudes 45º32'O e 45º50'O, totalizando uma área de aproximadamente 965 km^2, inseridas na Bacia do Rio Grande, interceptando partes dos municípios de Luiz Eduardo Magalhães, Barreiras e Riachão das Neves, estado da Bahia - Brasil.

Nas últimas décadas a região de estudo tem sido consideravelmente afetada pelas condições climáticas. Por ser classificada como clima Aw (clima tropical com estação seca de Inverno) no método de Köppen, apresentando precipitação média entre 1100 a 1700 mm/ano com uma estação seca de aproximadamente 5 meses, entre os meses de maio e setembro, houve um condicionamento da região pela ocupação das áreas por plantações extensivas de soja, café, arroz e algodão por meio de uma agricultura mecanizada e de irrigação contínua, destacando-se a irrigação por pivô central (SANTANA et al., 2000).

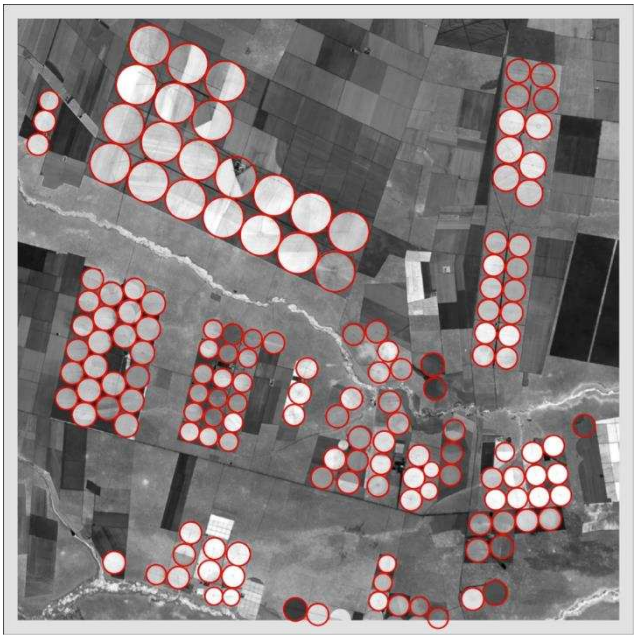

Figura 1 – Representação da localização da área de estudo que intercepta parte dos municípios de Luiz Eduardo Magalhães, Barreiras e Riachão das Neves, estado da Bahia – Brasil (destaque para a irrigação da area por pivô central).
Fonte: Ramos (2016).

2.2. Determinação da variável não-nula

A partir da identificação e determinação das áreas de irrigação por pivô central, da área de estudo, através do processo de extração apresentado em Ramos (2016), com a extração de valores do índice NDVI (Normalized Difference Vegetation Index) para 24 pivôs em 8 datas diferentes, os valores do índice foram relacionados a classes predefinidas em função dos ciclos fenológicos das culturas.

Contudo, como a aplicação do método proposto nesse trabalho não exige uma base de dados espaço-temporal, apenas um pivô e uma única data de amostragem georreferenciada foi utilizada, configurando 30753 pontos amostrais.

O índice NDVI é definido em escala linear com valores entre -1 e +1, porém, para a devida utilização dos dados em Ramos (2016), a escala linear dos valores passou a variar de 0 a +1 (valores não-negativos), ou seja, os dados foram reescalonados.

Com o objetivo de analisar o efeito do tamanho amostral para a metodologia proposta, a amostra original de 30753 valores do índice NDVI foi reduzida três vezes para 361, 314 e 79 observações, conforme limite mínimo recomendado por Modis e Papaodysseus (2006).

2.3. Proposição do método

Santos et al. (2011) mostraram que variáveis regionalizadas, conceitualmente, são representadas como

$$Y(x) = \mu(x) + \varepsilon'(x) + \varepsilon'' \quad (1)$$

em que, $\mu(x)$ é uma função determinística que descreve a componente estrutural de Y em **x**; $\varepsilon'(x)$ é um termo estocástico correlacionado localmente e ε'' é um ruído branco, não correlacionado, com distribuição normal, com média zero e variância σ^2.

Uma vez que a modelagem geoestatística apresenta caracterização e predição com boas propriedades estatísticas, os resíduos provenientes da autovalidação, e padronizados pela média e variância dos mesmos, seguem a distribuição normal padrão (MOOD et al., 1974; VIEIRA, 2000; SANTOS et al., 2011; YAMAMOTO e LANDIM, 2013).

Assim, Mood et al. (1974) apresentam alguns teoremas, demonstrados neste trabalho pela método da função geradora de momentos, doravante fgm, como segue.

Teorema 1. Se $Z_1, \dots, Z_n$ é uma amostra aleatória de uma distribuição normal padrão, então $\sum_{i=1}^{n} Z_i^2$ tem distribuição de qui-quadrado com n graus de liberdade.

Demonstração: utilizando o conceito de fgm para $U = \sum_{i=1}^{n} Z_i^2$, tem-se que:

$$m_U(t) = E[e^{tU}] = E\left[e^{t\sum_{i=1}^{n} Z_i^2}\right] = E\left[e^{t\left(Z_1^2 + Z_2^2 + \cdots + Z_n^2\right)}\right] =$$

$$= E\left[e^{tZ_1^2} e^{tZ_2^2} \dots e^{tZ_n^2}\right] = E\left[\prod_{i=1}^{n} e^{tZ_i^2}\right] =$$

$$= \prod_{i=1}^{n} E\left[e^{tZ_i^2}\right] \quad (2)$$

Pode-se perceber que a expressão (2) é garantida pelo teorema, uma vez que a amostra é dita aleatória, ou seja, independente e identicamente distribuída (iid).

Estes autores mostram ainda que, a partir da expressão (2), apresenta-se o teorema a seguir.

Teorema 2. Se $X_1, \dots, X_n$ são variáveis aleatórias independentes e a fgm de cada variável existe para todo –h < t < h para algum h > 0, e fazendo $Y = \sum_{i=1}^{n} X_i$, então $m_Y(t) = E[exp \sum_{i=1}^{n} tX_i] = \prod_{i=1}^{n} m_{X_i}(t)$, para todo –h < t < h.

Desenvolvendo separadamente $E\left[e^{tZ_i^2}\right]$ e como $Z_i \sim N(0,1)$, sua função densidade de probabilidade, fdp, é $f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$, logo

$$E\left[e^{tZ_i^2}\right] = \int_{-\infty}^{+\infty} e^{tz^2} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz = \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2 + tz^2} dz =$$

$$= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(z^2 - 2tz^2\right)} dz = \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{(1-2t)z^2}{2}} dz =$$

$$= \int_{-\infty}^{+\infty} \frac{\sqrt{1-2t}}{\sqrt{1-2t}} \frac{1}{\sqrt{2\pi}} e^{-\frac{(1-2t)z^2}{2}} dz =$$

$$= \frac{1}{\sqrt{1-2t}} \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} \frac{1}{\frac{1}{\sqrt{1-2t}}} e^{-\frac{1}{2}\left(\frac{z}{\frac{1}{\sqrt{1-2t}}}\right)^2} dz \qquad (3)$$

Mood et al. (1974) mostram que a expressão (3) é igual a $\frac{1}{\sqrt{1-2t}}$, uma vez que $f(z) = \frac{1}{\sqrt{2\pi}} \frac{1}{\frac{1}{\sqrt{1-2t}}} e^{-\frac{1}{2}\left(\frac{z}{\frac{1}{\sqrt{1-2t}}}\right)^2}$ é uma fdp de uma distribuição normal com média zero e variância $\frac{1}{1-2t}$, logo a integral é um.

Portanto,

$$m_U(t) = \prod_{i=1}^{n} E\left[e^{tZ_1^2}\right] = \prod_{i=1}^{n} \frac{1}{\sqrt{1-2t}} = \left(\frac{1}{\sqrt{1-2t}}\right)^n =$$

$$= \left(\frac{1}{1-2t}\right)^{\frac{n}{2}}. \qquad (4)$$

Nota-se que a expressão (4) é a fgm de uma distribuição qui-quadrado com n graus de liberdade.

Conclui-se, então, que a soma de n normais padrão independentes é uma qui-quadrado com n graus de liberdade.

Apesar da conclusão anterior, pode ser interessante uma conclusão mais geral, ou seja, não se basear somente na soma de normais padrão independentes. Assim, faz-se necessário o teorema a seguir.

Teorema 3. Duas variáveis aleatórias conjuntamente distribuídas X e Y são independentes se e somente se $m_{X,Y}(t_1, t_2) = m_X(t_1)m_Y(t_2)$ para todo t_1, t_2, em que $-h < t_1 < h, i = 1{,}2$, e $h > 0$.

Mood, Graybill e Boes (1974) e Casella e Berger (2010) apresentam a demonstração deste teorema (duas variáveis independentes). Como o interesse maior desse trabalho é para k variáveis, passa-se então a descrever tal demonstração, uma vez que estes autores afirmam ser possível a generalização do Teorema 3.

Demonstração:

Usando a fgm conjunta $m_{X_1,\dots,X_k}(t_1, \dots, t_k)$, sob independência, pode-se escrevê-la como o produto das fgm marginais,

$$m_{X_1,\dots,X_k}(t_1, \dots, t_k) = \prod_{i=1}^{k} m_X(t). \quad (5)$$

Como cada fgm marginal é dada por

$$m_X(t) = \left(\frac{1}{1-2t}\right)^{\frac{k}{2}} \quad (6)$$

para todo $t < \frac{1}{2}$ e $k = 1$, pode-se reescrever a equação (5) como

$$m_{X_1,\dots,X_k}(t_1, \dots, t_k) = \prod_{i=1}^{k} m_X(t) = \prod_{i=1}^{k} \left(\frac{1}{1-2t}\right)^{\frac{1}{2}}. \quad (7)$$

Então,

$$m_{X_1,\dots,X_k}(t_1, \dots, t_k) = \left(\frac{1}{1-2t}\right)^{\frac{k}{2}}. \quad (8)$$

De acordo com Mood et al. (1974), a equação (8) é uma distribuição gama $\left(\frac{k}{2}, \frac{1}{2}\right)$. Mas, pode-se perceber ainda que esta é também uma distribuição qui-quadrado com k graus de liberdade, denotado por $\chi^2_{(k)}$.

Portanto, quer para a soma de variáveis aleatórias ou um conjunto de variáveis aleatórias que são iid que seguem a distribuição normal padrão, uma vez que sejam tomados os quadrados dessas variáveis, a distribuição resultante será uma qui-quadrado com k graus de liberdade, denotado por $\chi^2_{(k)}, \forall k = n$.

Dessa forma, adotando o entendimento da decomposição da equação (1), adota-se a metodologia de Ferreira et al. (2013), ou seja, utilizar a análise geoestatística de dados geoespaciais com dependência espacial comprovada e obter os resíduos dessa modelagem.

Adotar esta abordagem geoestatística para os dados geoespaciais significa considerar cada ponto amostral georreferenciado como uma variável aleatória, gerando assim uma função aleatória, ou comumente, processo estocástico (CRESSIE, 1993; SANTOS et al., 2011).

Visando a validade estatística do método, assume-se a pressuposição de estacionariedade do variograma, ou seja, que o mesmo exista e que seja estacionário (Vieira, 2000; Yamamoto e Landim, 2013).

Adota-se ainda o variograma teórico, populacional, de Vieira (2000), ou seja, $2\gamma(h) = E\{[Y(x) - Y(x+h)]^2\}$, e o estimador de Kamimura et al. (2013), dado pela equação

$$\hat{\gamma}(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} \{[Y(x_i) - Y(x_i + h)]^2\} \qquad (9)$$

em que: $N(h)$ é o número de pares de valores medidos em (x_i) e $(x_i + h)$; $Y(x_i)$ e $Y(x_i + h)$ representam todas as variáveis aleatórias separadas por um vetor h que geram as amostras e, consequentemente, o principal mecanismo de detecção de dependência espacial da metodologia geoestatística, o variograma, um gráfico de $\hat{\gamma}(h)$ em função do vetor-distância **h**.

Após a análise geoestatística e consequente obtenção dos resíduos, obtidos a partir da diferença entre valores observados e preditos, as pressuposições foram testadas, a saber, independência, distribuição normal com média nula e variância constante e distribuição de qui-quadrado com n graus de liberdade, sendo todos os resultados satisfatórios.

A etapa seguinte foi a de estimação intervalar unilateral dos resíduos. Como os estimadores são variáveis aleatórias, suas estimativas comumente são distintas do valor do parâmetro, ou seja, comumente se comete um erro de estimação. Por esta razão tornou-se necessário a construção de intervalos de confiança com probabilidade $(1 - \alpha)$ (FERREIRA, 2009).

Para a estimação intervalar unilateral (IC - Intervalo de Confiança) adotou-se a distribuição de qui-quadrado ($\chi^2_{(n)}$) e nível de significância α de 1% (arbitrário), conforme equação (10) . Assim, em outras palavras, desejou-se determinar o quanto estas estimativas dos resíduos são prováveis $(1-\alpha)$ de confiança, em que $\alpha \in (0,1)$ (MOOD et al.,1974; VIEIRA, 2000; CASELLA E BERGER, 2010).

$$P\left[\frac{ns^2}{\chi_2^2} < \sigma^2 < \frac{ns^2}{\chi_1^2}\right] = 1-\alpha \qquad (10)$$

Assim, todos os valores que não pertencerem ao IC construído, sem viés, com variância mínima e levando em consideração a estrutura de dependência espacial, Silva (2012) mostra que tais valores tornam-se possíveis outliers. Estatisticamente, se $x_i \in IC_{(1-\alpha)}$ então x_i é ruído branco, caso contrário é um provável outlier.

Utilizando os recursos de georreferenciamento dos dados, pretendeu-se ainda apontar quantos, quais e onde estão os resíduos com alta probabilidade de serem outliers.

Para comparação e/ou validação do método foram realizadas comparações do novo método com um dos mais robustos, estatisticamente, e utilizados métodos de detecção atuais, o Box Plot (HOAGLIN et al., 1983).

Toda a parte inovadora da metodologia foi realizada através do software livre R (R Development Core Team, 2015), em que a análise geoestatística foi realizada através do pacote geoR, desenvolvido por Ribeiro Júnior e Diggle (2001). Contudo, para a análise de geoprocessamento, utilizou-se o software ArcGIS™ (ESRI, 2013).

3. RESULTADOS E DISCUSSÃO

3.1. Análise exploratória dos dados

Segundo Yamamoto e Landim (2013) e Ferreira et al. (2013), é importante verificar o comportamento espacial dos dados, ainda como análise exploratória.

Dessa forma, apresenta-se na Figura 2 parte da análise exploratória dos dados, através dos gráficos de quartis (gráficos que utilizam o primeiro quartil, a mediana e o terceiro quartil como divisor dos dados).

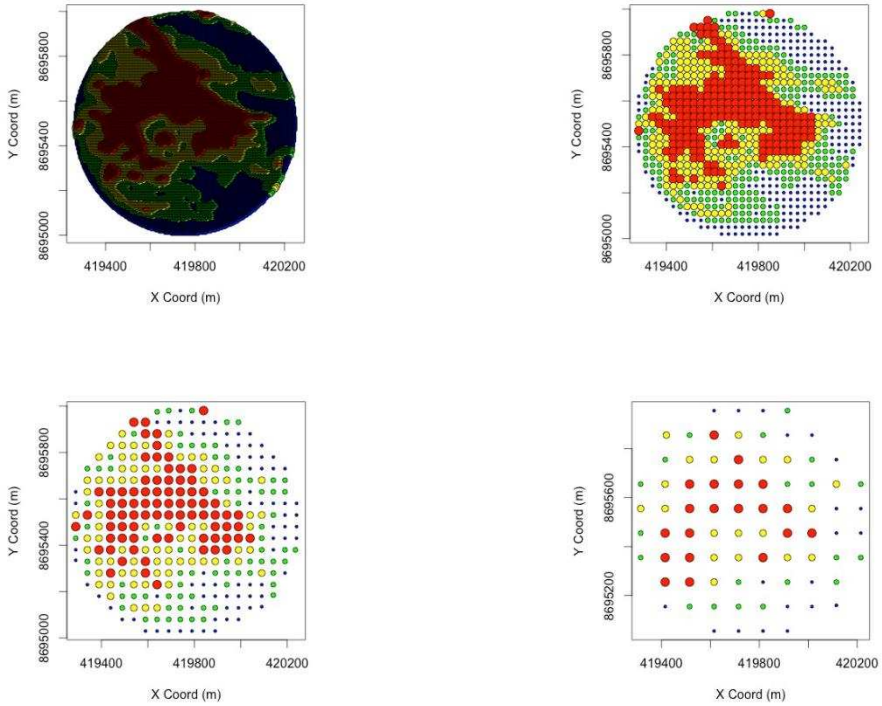

Figura 2 - Apresentação geral dos dados de NDVI (pelo método dos quartis) de irrigação por pivô central, em que as imagens mostram malhas regulares quadráticas com 5 m de distância mínima (superior esquerdo), 30 m (superior direito), 50 m (inferior esquerdo) e 100 m (inferior direito).

As imagens da Figura 2 apresentam as quatro densidades amostrais utilizadas neste trabalho. Estas dimensões constituem, primeiramente 100 % dos dados amostrais, com amostra de 5 em 5 m, totalizando em 30753 pontos, além de uma subamostra de 30 em 30 m, totalizando em 361 pontos, representando 1,17 % dos dados, com uma subamostra de 50 em 50 m, totalizando em 314 pontos, a qual representa 1,02% dos dados, e por fim uma subamostra de 100 em 100 m, totalizando 79 pontos, os quais representam 0,26% dos dados.

A subamostragem de 100 em 100 m (Figura 2) segue a metodologia proposta por Modis e Papaodysseus (2006), cuja densidade amostral para uma malha regular quadrática deve atender a um terço do valor estimado do parâmetro alcance do primeiro variograma obtido na análise geoestatística. Assim, decidiu-se pela distância lateral máxima de cada célula amostral de 100 m.

Ao analisar a Figura 2 nota-se que a ocorrência de possíveis outliers, não é evidente, uma vez que o comportamento da variável NDVI segue um padrão espacial bem definido.

3.2. Análise geoestatística dos dados

Conforme descrito, para obter os resíduos com propriedades que atendam às pressuposições teóricas da Estatística, os dados em relação às amostragens foram analisados pela Geoestatística, conforme resultados apresentados na Tabela 1.

Pode-se perceber que as principais características da análise foram preservadas nos três conjuntos de dados provenientes da amostra original, a saber: média, variância, desvio padrão, isotropia e o modelo do variograma. Destaca-se também a pequena variação das estimativas dos parâmetros do variograma.

Tabela 1 – Principais informações sobre a análise geoestatística dos dados de NDVI.

Medida/Característica	**Estimativas**			
	Tamanho da amostra (%)			
	0,26	1,02	1,17	100
Média (m)	0,54	0,54	0,54	0,54
Variância (m^2)	0,0017	0,0018	0,0017	0,0016
Desvio Padrão (m)	0,041	0,042	0,041	0,040
Anisotropia	Não	Não	Não	Não
Modelo	Esférico	Esférico	Esférico	Esférico
Efeito Pepita (m^2)	0,00018	0,00024	0,00012	0,00011
Contribuição (m^2)	0,0017	0,0019	0,0018	0,0020
Alcance (m)	530	520	510	504

Com base na análise variográfica, foi possível caracterizar o comportamento espacial do NDVI em toda a área estudada, e, em seguida, através da interpolação via krigagem simples, recomendada em Santos et al. (2011), foi possível fazer o mapeamento.

3.3. Detecção de outliers via análise de resíduos

Os resíduos provenientes de uma modelagem acurada apresentam características importantes que devem ser testadas, visando a aplicação de metodologias condicionais. Entre elas, destaca-se a independência (testada através da obtenção do efeito pepita-puro no variograma empírico), normalidade (testada através do teste Shapiro and Wilk, 1965) com média nula e variância unitária (para os resíduos padronizados). Todas essas

pressuposições foram constatadas nos resíduos obtidos, logo, os mesmos são classificados como ruído branco (MOOD, et al., 1974).

Mesmo diante de características preditivas ótimas, a krigagem também subestima e superestima valores observados. Assim, apresenta-se na Figura 3 os gráficos dos valores que mais se excederam na subestimação e superestimação do processo de autovalidação.

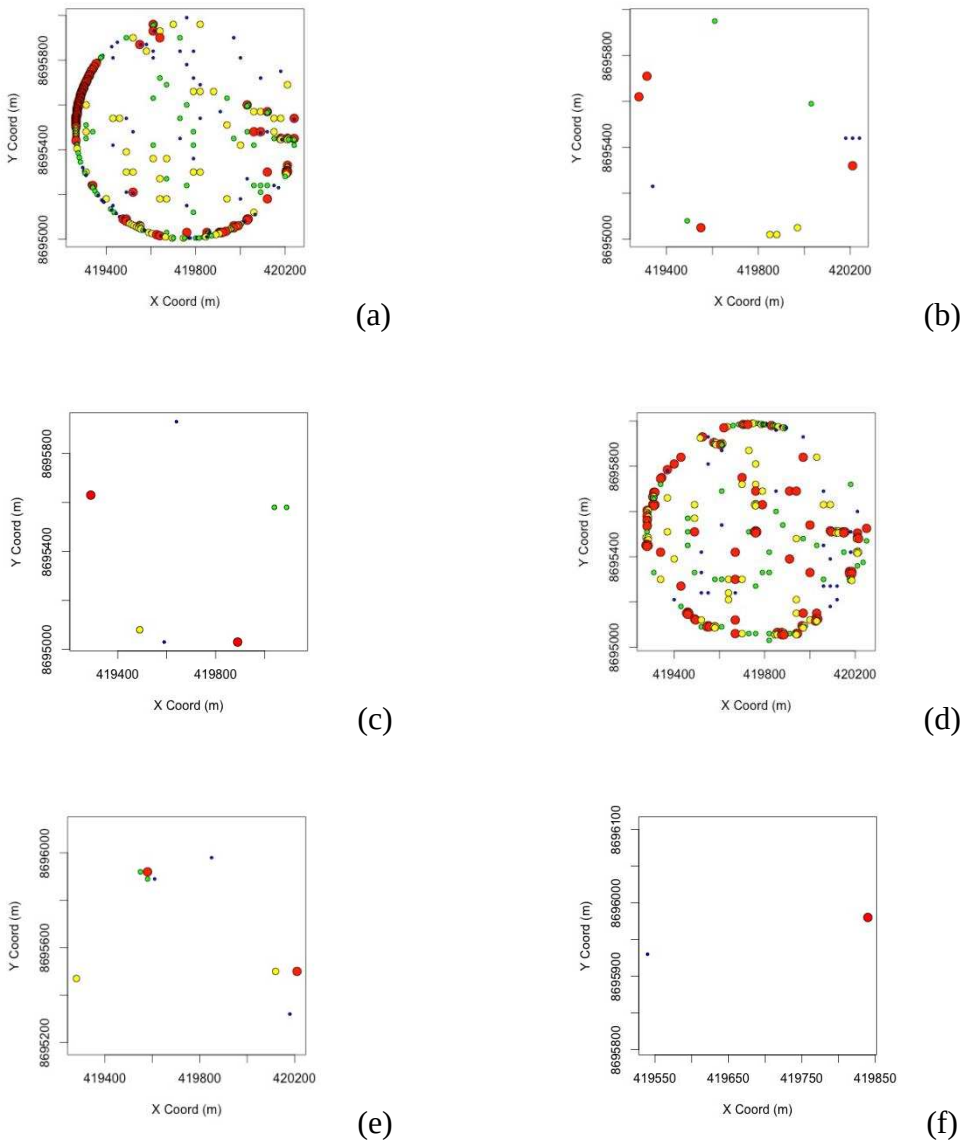

Figura 3 - Apresentação, em quartis, dos resíduos da autovalidação que mostraram os maiores excessos por subestimação e superestimação dos valores observados de NDVI. Em que (a), (b) e (c) são as subestimações excessivas, e (d), (e) e (f) são as superestimações excessivas, para 5, 30 e 50 m de espaçamento, respectivamente.

Pode-se perceber que a amostragem mínima do estudo, cuja densidade amostral foi dada por 100 em 100 m, não apresentou valores excessivos no processo de autovalidação geoestatística dos dados.

Como a Figura 3 apresenta, os resíduos padronizados da autovalidação que estão em destaque, são aqueles que podem exceder de forma bilateral os valores observados. Então, torna-se necessário tomar o Intervalo de Confiança unilateral, ao nível de 99%, dos quadrados desses resíduos para de fato associá-los à probabilidade de serem outliers.

Nota-se ainda, pela Figura 3, que, entre os valores plotados, os mais extremos (representados pela cor vermelha e maior diâmetro) estão próximos do limite da área do pivô central, o que comumente é denominado de "efeito de borda".

Assim, objetivando detectar os possíveis outliers dos conjuntos de dados, aplicou-se os teoremas apresentados anteriormente e obteve-se os possíveis outliers dos conjuntos de dados (não-negativos), além da localização geográficos dos mesmos.

Conforme demonstrado, tomando os quadrados dos resíduos padronizados, obtém-se a distribuição de qui-quadrado com n graus de liberdade. Logo, um Intervalo de Confiança, ao nível $(1-\alpha)$ de certeza pode ser estimado.

A partir daí, como recomenda Ferreira (2009), uma vez que os dados apresentam todos os requisitos exigidos pelas pressuposições estatísticas, passa-se à estimação dos possíveis outliers, com 1% de significância estatística, conforme resultados apresentados na Tabela 2.

Tabela 2 – Estimação dos possíveis outliers com 99% de probabilidade dos resíduos provenientes da autovalidação para os 4 conjuntos de dados do trabalho.

Densidade amostral (m)	**Possíveis outliers (%)**
5 m x 5 m	0,9
30 m x 30 m	1,3
50 m x 50 m	1,2

Conforme Tabela 2, pode-se perceber que as subamostras apresentaram valores coerentes de possíveis outliers, exceto a subamostra com densidade amostral mínima, cujo tamanho amostral foi recomendado por Modis e Papaodysseus (2006). Destaca-se o fato de que esta variável só apresenta possibilidade de outliers superiores, conforme explicita Ramos (2016).

Uma característica apresentada na subestimação e superestimação, representada pelos resíduos, foi confirmada na detecção dos possíveis outliers, ou seja, os limites do

pivô central estudado estão de fato influenciando na existência de valores com fortes indícios de discrepâncias (efeito de borda).

Este efeito de borda acontece na modelagem via Geoestatística devido o número limitado de pontos vizinhos, ou seja, quanto menor o número de pontos vizinhos maior será a imprecisão do modelo geoestatístico, favorecendo o surgimento de observações discrepantes (Yamamoto e Landim, 2013).

Estas discrepâncias podem ocorrer pela ineficácia do modelo preditivo ou pelo comportamento errático do mecanismo de coleta de dados ao se deparar com uma mudança brusca na região de estudo (Ramos, 2016).

Assim, ficou evidente a necessidade de uma metodologia de estudo que explorasse efeitos dessa natureza e a localização geográfica dos mesmos.

Portanto, apesar do Box Plot não apresentar (Figura 4), em nenhum dos conjuntos de dados utilizados no trabalho, possíveis outliers (exceto a amostra mais densa que apontou um único valor como possível outlier), a metodologia proposta neste trabalho foi capaz de mostrar que, aproximadamente, 1% dos valores observados de NDVI são probabilisticamente outliers, além de detectar a localização dos mesmos.

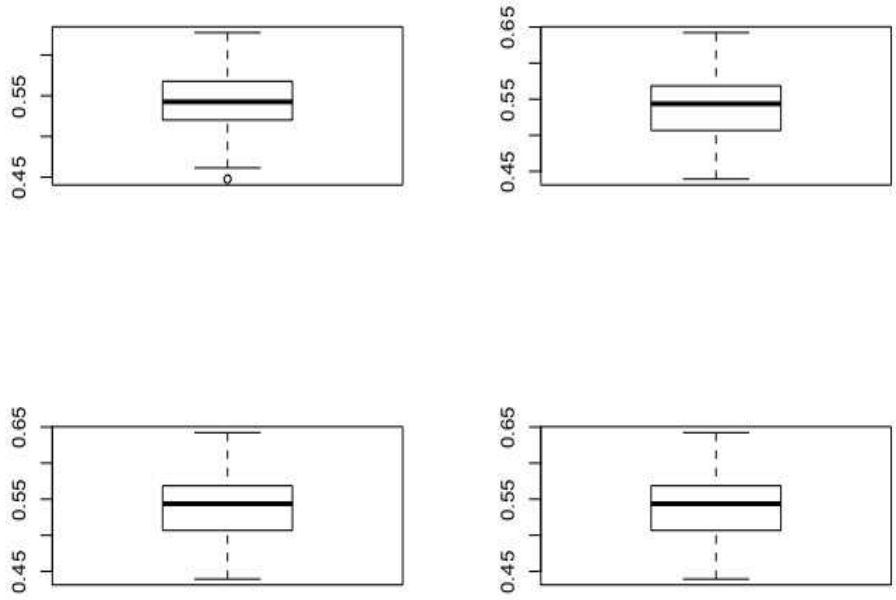

Figura 4 - Apresentação dos quatro Box Plot dos dados de NDVI de irrigação por pivô central, em que as densidades amostrais são dadas por 5 m de distância mínima (superior esquerdo), 30 m (superior direito), 50 m (inferior esquerdo) e 100 m (inferior direito).

Finalmente, nota-se que a metodologia proposta não apresentou perda de eficiência ao reduzir o tamanho amostral, pois partindo de uma amostra de, aproximadamente, 31 mil informações e finalizando em 314, o percentual de prováveis outliers e na mesma localização geográfica foram detectados. Assim, fica evidente a relevância da proposição para lidar com dados geoespaciais e não-negativos, conforme necessidade destacada por Vieira (2000), Ferreira et al. (2013) e Yamamoto e Landim (2013).

4. CONCLUSÕES

Independentemente da causa geradora de inconsistências e do tipo de variável em estudo é preciso adotar metodologias eficientes de detecção de outliers. Assim, para dados geoespaciais cujas observações são não-negativas, pode-se aplicar as teorias apropriadas para a detecção.

Através dos resultados apresentados nesse trabalho foi possível mostrar que o Box Plot, apesar de ser um método muito utilizado, não foi eficiente para a detecção de outliers dos dados apresentados, e que novas proposições metodológicas se fazem necessárias.

Assim, utilizando teoremas da Estatística Clássica e os resíduos obtidos pela análise geoestatística para variáveis geoespaciais contínuas e não-negativas, foi possível propor uma nova metodologia de detecção de outliers que associa altas probabilidades de ocorrência e o posicionamento espacial dos dados.

5. REFERÊNCIAS BIBLIOGRÁFICAS

ANSCOMBE, F.J. **Rejection of outliers.** Technometrics 20 (1960): 123-147.

APPICE, A., GUCCIONE, P., MALERBA, D., & CIAMPI, A**. Dealing. with temporal and spatial correlations to classify outliers in geophysical data streams.** Information Science 285 (2014): 162-80.

BARNETT, V., & LEWIS, T. **Outliers in statistical data.** Biometrical Journal 379 (1994): 256.

BARUA, S., & ALHAJJ, R**. High performance computing for spatial outliers detection using parallel wavelet transform**. Intelligent Data Analysis 11 (2007): 707-730.

BECKMAN, R.J., & COOK, R.D. **Outliers.** Technometrics 25 (1983): 119-149.

BENJAMINI, Y., & ADDISON, W. **Opening the Box of a Boxplot.** Journal of the American Statistical Association 42 (1988): 257-262.

CASELLA, G., BERGER, R.L. **Inferência estatística**. Cengage Learning (2010).

CRESSIE, N**. Statistics for spatial data**. Wiley-Interscience (1993).

ENVIRONMENTAL SYSTEMS RESEARCH INSTITUTE. – **Esri. ArcGIS Desktop: Release 10.** Redlands, CA: 2011.

FERREIRA, D.F. **Estatística básica**. Lavras, Editora UFLA (2009).

FERREIRA, I.O., SANTOS, G.R., & RODRIGUES, D.D. **Estudo sobre a utilização adequada da krigagem na representação computacional de superfícies batimétricas.** Revista Brasileira de Cartografia 65 (2013): 831-842.

GRUBBS, F.E. **Procedures for detecting outlying observations in samples**. Technometrics 11 (1969): 1-21.

HOAGLIN, D.C, MOSTELLER, F., & TUKEY, J.W. **Understanding robust and exploratory data analysis.** New York, J. Wiley (1983)

HONGXING, L., KENNETCH, C.J., & MORTON, E.O. **Detecting outliers in irregularly distributed spatial data sets by locally adaptive and robust statistical analysis and GIS.** International Journal Geographical Information Science 15 (2001): 721-741.

KAMIMURA, K.M., SANTOS, G.R., OLIVEIRA, M.S., DIAS, JR., M.S., & GUIMARÃES, P.T.G. **Variabilidade espacial de atributos físicos de um Latossolo Vermelho-Amarelo sob lavoura cafeeira.** Revista Brasileira de Ciência do Solo 37 (2013): 877-888.

MODIS K. AND PAPAODYSSEUS K. **Theoretical Estimation of the Critical Sampling Size for Homogeneous Orebodies with Small Nugget Effect on.** Mathematical Geology, 38 **(**2006): 489-501.

MOOD, A.M., GRAYBILL, F.A., & BOES, D.C. **Introduction to the theory of statistics.** Kogakusha, McGraw-Hill, (1974)

MUÑOZ-GARCIA, J., MORENO-REBOLLO, J.L., & PASCUAL-ACOSTA, A. **Outliers: a formal approach.** International Statistical Review 58 (1990): 215-226.

QIAO, C., HAIBO, H., & HONG, M. **Spatial outlier detection based on iterative self-organizing learning model.** Neurocomputing 117 (2013): 161-172.

R CORE TEAM. R: **a language and environment for statistical computing.** R Foundation for Statistical Computing, (2014) Vienna, W. Recuperado de http://www.Rproject.org/.

RIBEIRO, J.P.J, & DIGGLE, P.J. GEOR: **a package for geostatistical analysis**. R-News. 1 (2001): 15-18.

ROUSSEUW, P.J., & ZOMEREN, B.C. **Unmasking multivariate outliers and leverage points**. Journal of the American Statistical Association 85 (1990): 633-639.

SANTOS, G.R., OLIVEIRA, M.S, & SANTOS, A.M.R.T. **Krigagem simples versus krigagem universal: qual o preditor mais preciso?** Revista Energia na Agricultura 26 (2011): 49-55.

TUKEY, J.W. **Exploratory Data Analysis Princeton**. Ed. Pearson (1977).

VIEIRA, S.R. **Geoestatística em estudos de variabilidade espacial do solo**. Tópicos em Ciências do Solo 1(2000): 1-54.

YAMAMOTO, J., & LANDIM, P. **Geoestatística: Conceitos e Aplicações**. São Paulo, Oficina de Textos, (2013)

CAPÍTULO 3

AVALIAÇÃO DOS EFEITOS DE OUTLIERS EM VARIÁVEIS GEOESPACIAIS CONTÍNUAS UTILIZANDO GEOESTATÍSTICA

RESUMO

O que fazer com os outliers, ou dados que se apresentam como extremos da distribuição dos dados, é algo bem controverso na literatura técnico-científica. É comum encontrar tanto a recomendação da eliminação destes elementos do conjunto de dados quanto a não eliminação, sendo lógicas as justificativas de ambas as partes. Assim, o objetivo deste trabalho foi apresentar um estudo dos efeitos dos outliers nos parâmetros de qualidade de ajuste de variáveis geoespaciais contínuas, usando a Geoestatística como principal metodologia. Através da autovalidação leave-one-out, foi possível mostrar que eliminando os outliers do conjunto de dados amostrais a acurácia obtida foi significativa estatisticamente, sugerindo assim que o procedimento proposto tem um grande potencial metodológico.

Palavras-chave: Outliers, discrepâncias geoespaciais, autovalidação.

ABSTRACT

What to do with outliers, or data that presents itself as extremes of data distribution, is something very controversial in technical-scientific literacy. It is common to find the recommendation of eliminating of such elements of the data set as it is common to find the non-elimination of such. Thus, the aim of this work was to present a study of the outliers effect in the parameters of adjustment quality of continuous geospatial data, utilizing Geostatistics as main methodology. Through the cross-validation leave-one-out, it was possible to show that by eliminating the outliers of sampling data set the accuracy obtained was statistically meaningful, suggesting that the proposed procedure has a great methodological potential.

Keywords: Outliers, geospatial discrepancies, cross-validation.

1. INTRODUÇÃO

Os outliers, ou dados discrepantes, são dados que podem ser classificados como pertencentes à outra população devido seu comportamento extremo (quanto à distribuição probabilística) em relação aos demais (FERREIRA, 2009).

Dessa forma, na maioria das vezes, encontra-se na literatura a recomendação de eliminação dos mesmos do conjunto de dados para que os resultados das análises estatísticas não sejam comprometidos. Contudo, é comum encontrar também a recomendação da não eliminação desses dados, gerando assim um questionamento sobre a tomada de decisão em relação a este assunto.

Pode-se perceber que este tema é de grande relevância para a ciência, pois a preocupação sobre como lidar com esse tipo de situação é antiga. Estudos clássicos como os de Anscombe (1960), Grubbs (1969), Beckman e Cook (1983), Rousseuw e Zomeren (1990), Muñoz-Garcia et al. (1990), Barnett e Lewis (1994), motivam pesquisadores mais atuais como Hongxing, et al. (2001), Barua e Alhajj (2007), Quiao et al. (2013) e Appice, et al. (2014). Destaca-se também as propostas metodológicas feitas nos capítulos 1 e 2 deste trabalho.

Assim, o objetivo deste trabalho é apresentar um estudo sobre o ganho significativo na acurácia da modelagem do fenômeno estudado, em relação aos efeitos dos outliers nos parâmetros de qualidade de ajuste de dados geoespaciais contínuos, tendo a Geoestatística como principal metodologia. Para validação do estudo, utilizou-se de três conjuntos de dados simulados com diferentes percentuais de contaminação e um conjunto de dados amostrais contaminado com outliers.

Como hipótese, espera-se mostrar que eliminando os outliers dos conjuntos de dados, pela metodologia proposta no Capítulo 1 deste trabalho, a acurácia obtida nos parâmetros de qualidade de ajuste da Geoestatística, segundo Vieira (2000), será significativa estatisticamente.

A escolha pela Geoestatística, como metodologia de modelagem da variável, está baseada em suas características ideais apresentadas por Santos et al. (2011), Yamamoto e Landim (2013) e no Capítulo 1 deste trabalho. Vieira (2000) também apresenta algumas características interessantes, do ponto de vista preditivo, descrevendo-as como interpolação sem tendência e com variância mínima, além de levar em consideração na

modelagem e predição a estrutura de dependência espacial das amostras, o que é inerente à variável em estudo.

2. MATERIAL E MÉTODOS

Os dados utilizados neste trabalho são de duas naturezas diferentes. O primeiro conjunto de dados são informações oriundas de simulações computacionais de três variáveis aleatórias regionalizadas que seguem uma distribuição conhecida e com diferentes percentuais de contaminação, ou seja, 5%, 10% e 20% dos dados dessa variável seguem uma distribuição assimétrica, gerando assim outliers.

O segundo conjunto de dados são informações reais de um levantamento amostral altimétrico de um município do Estado de Iowa – EUA utilizando a tecnologia LiDAR. Sabidamente, conforme descrito no Capítulo 1 deste trabalho, cerca de 1,3% dos dados são outliers.

2.1 Caracterização dos dados simulados

Para a obtenção dos dados simulados, três malhas regulares de 21 x 21 pontos amostrais foram criadas. Assim, os três conjuntos de dados com 441 informações de uma variável fictícia foram georreferenciadas de maneira peculiar para este trabalho.

Para que fossem conhecidos os percentuais de outliers, uma estrutura de covariância espacial foi estabelecida para todos os dados simulados, ou seja, as simulações foram feitas com base em um campo aleatório gaussiano. Em seguida, 5%, 10% e 20% dos dados gaussianos foram substituídos aleatoriamente por um campo aleatório qui-quadrado.

Todas as simulações foram feitas tendo como base principal o pacote RandomFields (Schlather et al., 2015) do Programa R (R Core Team, 2016).

2.2 Caracterização dos dados reais

A área de estudo é localizada no estado de Iowa, Estados Unidos, no condado de Pottawattamie, compreendendo uma porção de 34,3 hectares do município de Treynor.

Delimita-se a região de estudos pelas latitudes 41°10’23”N e 41°09’53”N, e longitudes 95°38’24”W a 95°38’47”W, conforme Figura 1.

Conforme Höhle e Höhle (2009), para mapeamento de áreas com determinadas características é comum o uso de Modelos Digitais de Elevação (MDE), utilizando principalmente a tecnologia LiDAR (Light Detection And Ranging). Além de proporcionar uma alta densidade de pontos tridimensionais, o método LiDAR proporciona um mapeamento acurado e eficiente.

Assim, os dados reais utilizados neste trabalho são provenientes de um mapeamento LiDAR, sem qualquer pré-processamento, referenciados ao sistema geodésico NAD 83 (North American Datum of 1983) e representados na projeção UTM (Universal Transversa de Mercator) fuso 15N. Estes dados totalizam 1920 pontos tridimensionais, com espaçamento mínimo entre os pontos de 0,8 metros. Esta região apresenta valores de altitudes de 340,8 a 385,5 metros.

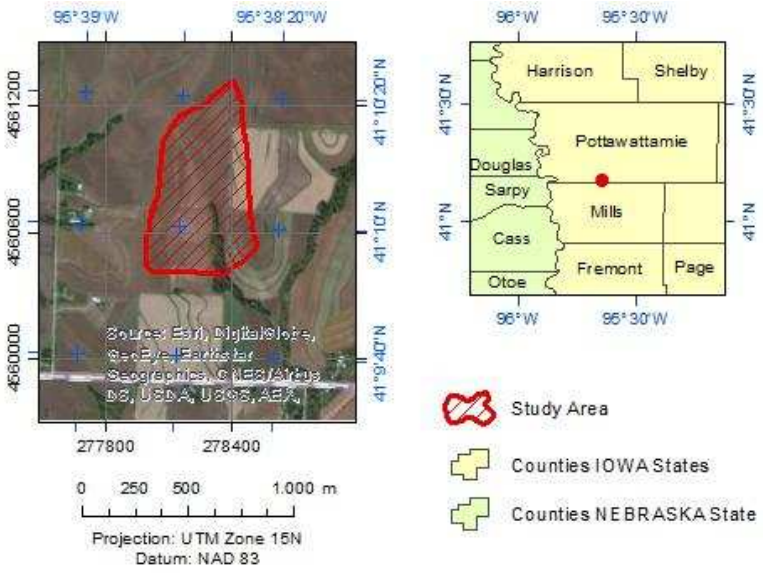

Figura 1 - Apresentação da área de estudos, localizada no estado de Iowa, Estados Unidos, no condado de Pottawattamie, compreendendo uma porção de 34.3209 hectare do município de Treynor.

2.3 Proposição do método

Utilizando a notação geoestatística de Santos et al. (2011) para as variáveis regionalizadas com dependência espacial, ou seja,

$$Y(x) = \mu(x) + \varepsilon'(x) + \varepsilon'' \quad (1)$$

em que, $\mu(x)$ é uma função determinística que descreve a componente estrutural de Y em **x**; $\varepsilon'(x)$ é um termo estocástico correlacionado localmente e ε'' é um ruído branco.

Essa notação apresenta uma característica interessante, ou seja, uma vez que a modelagem geoestatística apresenta caracterização e predição com boas propriedades estatísticas, os resíduos provenientes da autovalidação devem apresentar a característica de ruído branco, ou seja, os resíduos devem ser não correlacionados, seguir a distribuição normal com média nula e variância finita (MOOD et al., 1974; VIEIRA, 2000; SANTOS et al., 2011; YAMAMOTO e LANDIM, 2013).

Assim, utilizou-se a abordagem do Capítulo 1 deste trabalho e o procedimento descrito por Ferreira, et al. (2013) para uma análise geoestatística assistida. Esta metodologia foi utilizada também na recomendação ou não da exclusão dos outliers do conjunto de dados, tendo como base os indicadores de qualidade de ajuste da análise geoestatística adotados por Vieira (2000).

Especificamente, foi assumida como verdadeira a decomposição da equação (1) e utilizada a análise geoestatística de dados geoespaciais com dependência espacial comprovada. Os resíduos dessa modelagem foram obtidos através da autovalidação geoestatística, denominada de leave-one-out.

Visando a validação estatística do método, foi assumida a pressuposição de estacionariedade do variograma, ou seja, que o variograma existe e que seja estacionário (VIEIRA, 2000; YAMAMOTO e LANDIM, 2013).

Adotou-se ainda o variograma teórico de Vieira (2000), cujo estimador citado por Kamimura et al. (2013) é dado pela equação (2),

$$\hat{\gamma}(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} \{[Y(x_i) - Y(x_i + h)]^2\} \quad (2)$$

em que, $N(h)$ é o número de pares de valores medidos em (x_i) e $(x_i + h)$; $Y(x_i)$ e $Y(x_i + h)$ representam todas as variáveis aleatórias separadas por um vetor h que geram as amostras e, consequentemente, o principal mecanismo de detecção de dependência espacial da metodologia geoestatística, o variograma, um gráfico de $\hat{\gamma}(h)$ em função do vetor-distância **h**.

Após a análise geoestatística os resíduos foram obtidos a partir da diferença entre valores observados e preditos. Também foram testadas as pressuposições de ruído branco, conforme recomendam Mood et al. (1974).

Especificamente, os resíduos provenientes de uma modelagem acurada devem apresentar pressuposições importantes como: independência (testada através da obtenção do efeito pepita-puro no variograma empírico), normalidade (testada através do teste Shapiro and Wilk, 1965) com média nula e variância unitária (para os resíduos sejam padronizados). Todas essas pressuposições foram testadas nos resíduos obtidos (MOOD, et al., 1974).

A etapa seguinte foi a de estimação intervalar bilateral (IC) para a média dos resíduos padronizados, cujo nível de confiança do intervalo foi de 99%, ou seja, o nível arbitrário de significância foi $\alpha = 1\%$ (FERREIRA, 2009).

Conforme demonstrado no Capítulo 1 deste trabalho, todos os valores que não pertencerem ao IC construído, sem viés, com variância mínima e levando em consideração a estrutura de dependência espacial, tornaram-se possíveis outliers.

Como sabidamente os quatro conjuntos de dados estavam contaminados com outliers, estimativas dos parâmetros de qualidade de ajuste vindos da autovalidação foram obtidos antes e depois da retirada dos valores discrepantes. Logo em seguida, um teste estatístico (qui-quadrado) foi realizado para verificar a significância da diferença obtida entre as estimativas da variância dos erros e do coeficiente de determinação R^2_{ad} (este coeficiente vem da Regressão Linear Simples entre os valores observados e preditos de Vieira, 2000).

Esta fase da metodologia proposta neste trabalho foi também realizada através do software livre R (R Core Team, 2016), e do pacote geoR, desenvolvido por Ribeiro Júnior e Diggle (2001).

3. RESULTADOS E DISCUSSÃO

Com o intuito de verificação do comportamento descritivo dos dados reais tanto em nível clássico quanto espacial, a amostra de 1920 pontos de altitude (coordenada Z dos pontos tridimensionais), apresentou a média de 363,24 m, variância de 62,09 m^2, distância mínima de 0,8 m entre as observações e distância máxima de 952 m.

Para o comportamento espacial dos dados reais, pode-se perceber, através da Figura 2a (gráfico dos quartis dos dados), que entre os pontos de menor altitude vários valores podem apresentar probabilidade de serem outliers, conforme destaque na imagem.

Conforme descrito, para obter os resíduos com propriedades que atendam às pressuposições dos resíduos classificados como ruído branco, os dados foram analisados pela metodologia geoestatística, ou seja, obtenção e modelagem do variograma experimental, interpolação via krigagem simples da variável em regiões não amostradas e mapeamento das incertezas associadas às interpolações, conforme Figura 2b, 2c e 2d.

Vale destacar que a krigagem realizada foi a krigagem simples, pois, segundo Santos et al. (2011), este preditor linear apresenta maior acurácia em relação à krigagem ordinária e à krigagem universal.

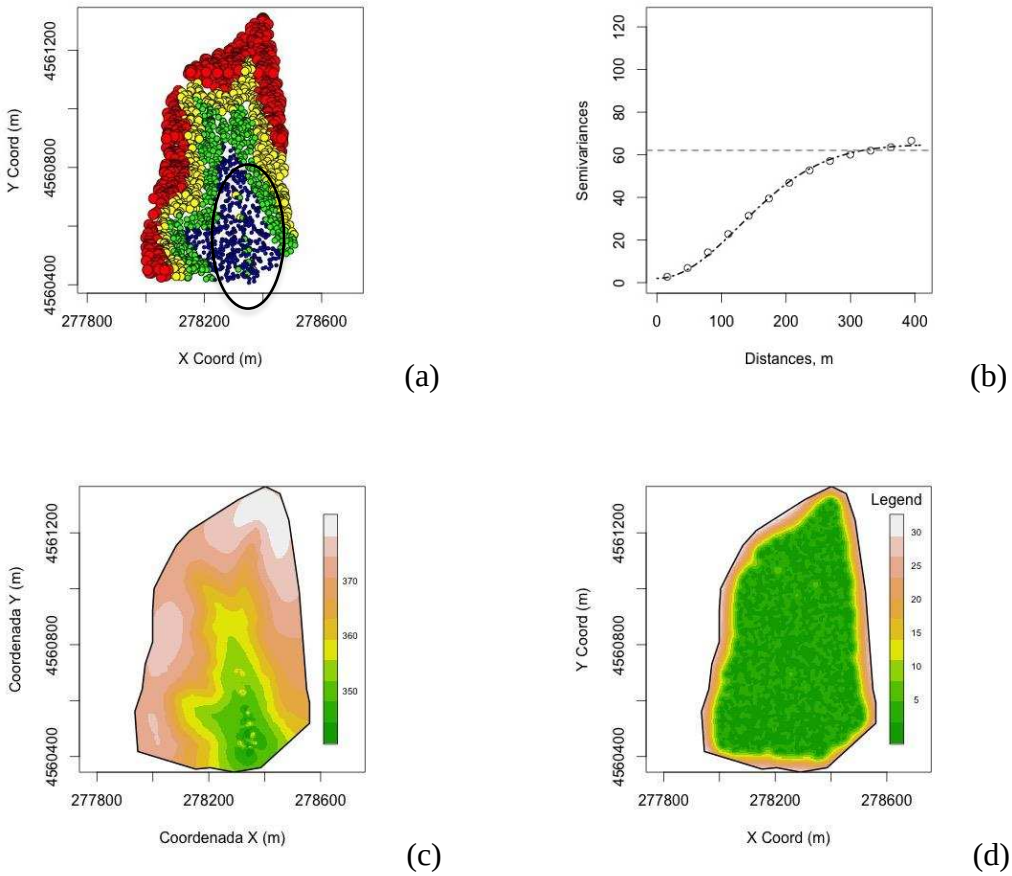

Figura 2 - Análise geoestatística dos dados da altimetria, em uma região da cidade de Treynor – estado de Iowa – USA. Em que, (a) gráfico dos quartis, (b) variograma empírico modelado, (c) krigagem simples, e (d) variância de krigagem.

Os resíduos provenientes de uma modelagem acurada apresentam características importantes que devem ser testadas, visando a aplicação de metodologias condicionais. Entre elas, destaca-se a independência (testada através da obtenção do efeito pepita-puro no variograma empírico), normalidade (testada através do teste Shapiro e Wilk, 1965) com média nula e variância unitária (para os resíduos padronizados). Todas essas pressuposições foram constatadas nos resíduos obtidos, logo, os mesmos são classificados como ruído branco (MOOD et al., 1974).

Através do Capítulo 1 deste trabalho, nota-se que estas características são, de fato, as principais que devem ser testadas para classificar os resíduos como ruído branco. Contudo, estes autores complementam estas pressuposições com a característica denominada homogeneidade espacial, ou seja, os resíduos ficam distribuídos espacialmente de maneira homogênea em toda a área de estudo, uma vez que toda e qualquer tendência de aglomeração foi modelada. Esta característica também foi constatada nos dados através da obtenção do gráfico dos quartis dos resíduos.

Em seguida, a estimação intervalar bilateral para a média dos resíduos padronizados foi realizada, com significância arbitrária de 1%, e, assim, aproximadamente 1,2% dos dados apresentaram alta probabilidade de serem outliers. Todos estes possíveis outliers foram georreferenciados e apresentados na Figura 3a (possíveis outliers inferiores divididos em quartis) e 3b (possíveis outliers superiores divididos em quartis).

Ainda para a Figura 3 (“c” e “d”) foram apresentados o gráfico dos quartis dos dados sem os prováveis outliers e o mapa de krigagem simples obtido a partir do novo conjunto de dados, respectivamente. Nitidamente pode-se perceber uma correção do problema apresentado na Figura 2a e 2c, além de uma coerência quantitativa entre os vizinhos.

Como parte do objetivo deste trabalho é sugerir o que fazer com os prováveis outliers, ou seja, excluir ou não do conjunto de dados, a exclusão foi executada e as estimativas dos parâmetros de qualidade de ajuste de uma análise geoestatística foram obtidos, antes e depois da exclusão.

As estimativas dos parâmetros da qualidade de ajuste foram apresentados na Tabela 1, conforme resultados do Capítulo 1 deste trabalho e recomendações de Vieira (2000), Santos et al., (2011) e Ferreira et al. (2013),

Yfantis et al. (1987), avaliando teoricamente indicadores de qualidade para uma análise geoestatística, afirmam que a variância do erro da autovalidação figura entre os mais indicados parâmetros dessa natureza e, essencialmente, deve ser a menor possível.

Segundo Mood et al. (1974), a variância do erro pode seguir uma distribuição de probabilidade qui-quadrado, logo um teste de hipóteses dessa natureza indica se a diferença dos dados originais e sem outliers é significativa estatisticamente. Logo, ao realizar o teste foi percebida a significância em um nível arbitrário de 5%.

Assim, pela Tabela 1, a variância do erro após a exclusão dos possíveis outliers do conjunto de dados foi, aproximadamente, 70% menor do que os dados originais.

Segundo Vieira (2000), o coeficiente de determinação R^2_{ad} da Regressão Linear Simples entre os valores observados e preditos deve ser estimado o mais próximo possível da unidade. Já Mood et al. (1974) mostram que também este parâmetro pode seguir uma distribuição de probabilidade qui-quadrado. Assim, ao realizar o teste foi percebida a significância em um nível arbitrário de 5%.

Pela Tabela 1, nota-se que o aumento neste coeficiente foi de 20%, aproximadamente, e, pelo teste estatístico, foi constatada a significância da diferença.

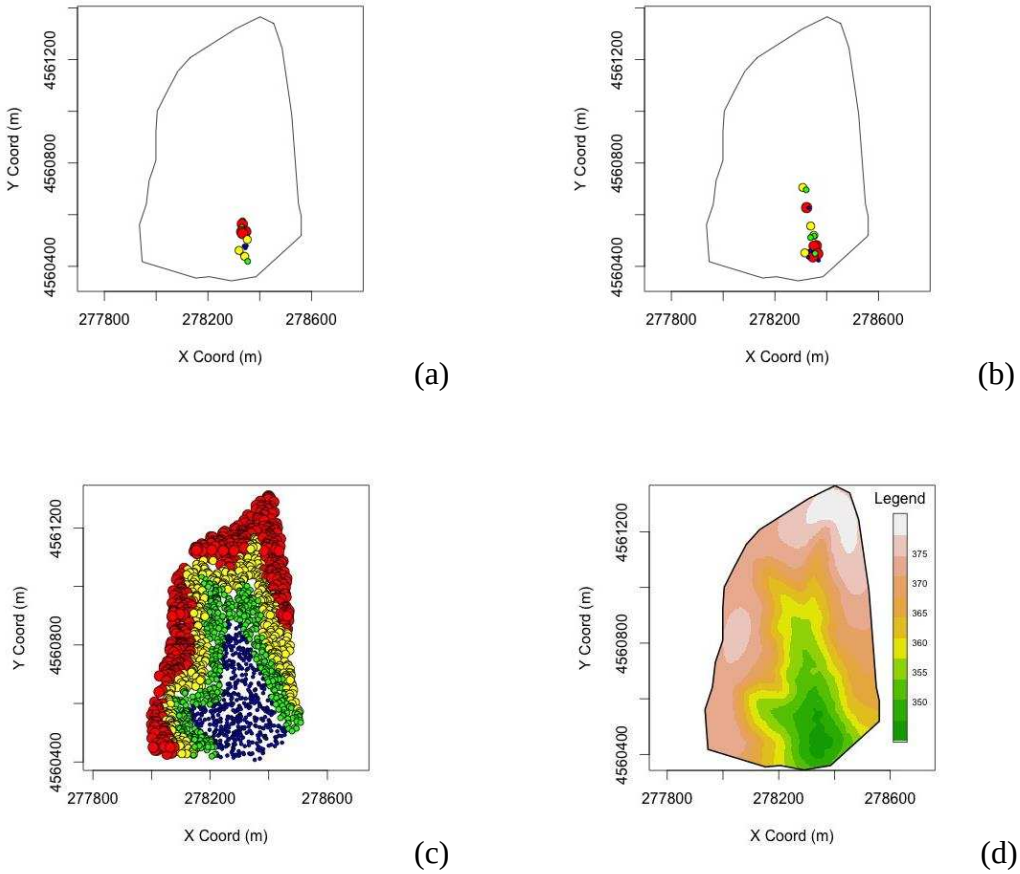

Figura 3 - Detecção dos prováveis outliers e consequências da exclusão destes, via imagens. Em que, (a) prováveis outliers inferiores, (b) prováveis outliers superiores, (c) gráfico dos quartis sem os prováveis outliers, e (d) krigagem simples dos valores sem os prováveis outliers.

Tabela 1 – Principais indicadores de qualidade de uma análise geoestatística, em que os resultados são referentes aos dados de altimetria da cidade de Treynor – IA – USA, antes e depois da exclusão dos prováveis outliers

Indicadores	**Dados originais**	**Dados sem outliers**
$\hat{\sigma}_{\varepsilon}^{2}$	1,84	0,54*
R_{ad}^{2} da RLS	0,55	0,66*

* Estatisticamente significativa a 5% pelo teste de qui-quadrado com (n-1) G.L.

Portanto, de maneira geral, os indicadores de qualidade de ajuste de uma análise geoestatística apresentaram significância estatística ($\alpha = 5\%$) quando os prováveis outliers são excluídos do conjunto de dados.

Para os dados simulados, onde ocorreram 3 níveis de contaminação com outliers (5, 10 e 20%), conforme Tabela 2, a necessidade de exclusão dos dados discrepantes também ficou evidente, pois, adotando-se o mesmo nível de significância dos dados reais,

todos os testes mostraram que a diferença entre os indicadores é estatisticamente significativa.

Assim, conforme método proposto no Capítulo 1 deste trabalho, os outliers foram detectados em proporções muito próximas das contaminações, ou seja, para os dados com 5% de contaminação foram detectados 5,5%, para 10% foram detectados 9,3% e para 20% foram detectados 18,9%.

Portanto estes dados foram eliminados do conjunto de dados simulados e novas estimativas dos parâmetros de qualidade de ajuste foram feitas, conforme apresentado na Tabela 2.

Tabela 2 – Principais indicadores de qualidade de uma análise geoestatística, em que os resultados são referentes aos dados de uma variável fictícia simulada computacionalmente com diferentes níveis de contaminação com outliers

Indicadores \ Contaminação	**5% de Outliers**		**10% de Outliers**		**20% de Outliers**	
	Antes	**Depois**	**Antes**	**Depois**	**Antes**	**Depois**
$\hat{\sigma}_{\varepsilon}^{2}$	2,18	1,68*	3,16	1,64*	3,77	1,70*
R_{ad}^{2} da RLS	0,66	0,72*	0,50	0,67*	0,40	0,65*

* Estatisticamente significativa a 5% pelo teste de qui-quadrado com (n-1) G.L.

Pela Tabela 2, além da diferença significativa entre as estimativas, pode-se perceber ainda que quanto maior o nível de contaminação dos dados pior o comportamento das estimativas dos parâmetros $\hat{\sigma}_{\varepsilon}^{2}$ e R_{ad}^{2}. Contudo, após a exclusão dos dados discrepantes aconteceu uma estabilização das estimativas, ou seja, após a exclusão a variância do erro estimada foi respectivamente 1,68; 1,64 e 1,70.

Portanto, a proposta metodológica sobre a eliminação ou não dos dados tidos como outliers uma análise geoestatística do tipo “antes” e “depois”, adotando os parâmetros da autovalidação leave-one-out variância dos erros e o coeficiente de determinação da Regressão Linear Simples entre os valores observados e preditos, conforme recomendado por Vieira (2000), mostrou-se eficaz. Assim, fica evidente que uma vez que houver significância estatística a um nível arbitrário para o teste de qui-quadrado com n-1 graus de liberdade, recomenda-se a eliminação dos possíveis outliers do conjunto de dados.

4. CONCLUSÕES

Através da aplicação da metodologia geoestatística de análise de variáveis geoespaciais contínuas foi possível mostrar que a mesma apresenta características ótimas (que atendem às pressuposições teóricas) tanto na detecção quanto na orientação da exclusão ou não dos outliers.

A utilização dos indicadores de qualidade de ajuste da metodologia geoestatística, principalmente a variância do erro e o coeficiente de determinação da Regressão Linear Simples entre os valores observados e preditos (ambos obtidos na autovalidação), pode ser uma proposta metodológica eficiente estatisticamente.

Contudo, sugere-se que outras fontes de dados geoespaciais devem ser testadas a fim de avaliar a aplicabilidade da metodologia proposta, devido a eficiência apresentada pela mesma neste trabalho.

5. REFERÊNCIAS BIBLIOGRÁFICAS

ANSCOMBE, F.J. **Rejection of outliers.** Technometrics 20 (1960): 123-147.

APPICE, A., GUCCIONE, P., MALERBA, D., & CIAMPI, A. **Dealing. with temporal and spatial correlations to classify outliers in geophysical data streams.** Information Science 285 (2014): 162-80.

BARNETT, V., & LEWIS, T. **Outliers in statistical data.** Biometrical Journal 379 (1994): 256.

Barua, S., & Alhajj, R. **High performance computing for spatial outliers detection using parallel wavelet transform.** Intelligent Data Analysis 11 (2007): 707-730.

BECKMAN, R.J., & COOK, R.D. **Outliers.** Technometrics 25 (1983): 119-149.

FERREIRA, D.F. **Estatística básica.** Lavras, Editora UFLA (2009).

FERREIRA, I.O., SANTOS, G.R., & RODRIGUES, D.D**. Estudo sobre a utilização adequada da krigagem na representação computacional de superfícies batimétricas.** Revista Brasileira de Cartografia 65 (2013): 831-842.

GRUBBS, F.E. **Procedures for detecting outlying observations in samples.** Technometrics 11 (1969): 1-21.

HÖHLE, J., & HÖHLE, M. **Accuracy assessment of digital elevation models by means of robust statistical methods.** ISPRS Journal of Photogrammetry and Remote Sensing 64 (2009): 398-406.

HONGXING, L., KENNETCH, C.J., & MORTON, E.O. **Detecting outliers in irregularly distributed spatial data sets by locally adaptive and robust statistical analysis and GIS**. International Journal Geographical Information Science 15 (2001): 721-741.

KAMIMURA, K.M., SANTOS, G.R., OLIVEIRA, M.S., DIAS, JR., M.S., & GUIMARÃES, P.T.G. **Variabilidade espacial de atributos físicos de um Latossolo Vermelho-Amarelo sob lavoura cafeeira.** Revista Brasileira de Ciência do Solo 37 (2013): 877-888.

MOOD, A.M., GRAYBILL, F.A., & BOES, D.C. **Introduction to the theory of statistics**. Kogakusha, McGraw-Hill, (1974)

MUÑOZ-GARCIA, J., MORENO-REBOLLO, J.L., & PASCUAL-ACOSTA, A. **Outliers: a formal approach.** International Statistical Review 58 (1990): 215-226.

QIAO, C., HAIBO, H., & HONG, M. **Spatial outlier detection based on iterative self-organizing learning model.** Neurocomputing 117 (2013): 161-172.

R Core Team. R: **a language and environment for statistical computing.** R Foundation for Statistical Computing, (2014) Vienna, W. Recuperado de http://www.Rproject.org/.

RIBEIRO, J.P.J, & DIGGLE, P.J. **GeoR: a package for geostatistical analysis**. R-News. 1: 15-18.

ROUSSEUW, P.J., & ZOMEREN, B.C. **Unmasking multivariate outliers and leverage points.** Journal of the American Statistical Association 85 (1990): 633-639.

SANTOS, G.R., OLIVEIRA, M.S, & SANTOS, A.M.R.T. **Krigagem simples versus krigagem universal: qual o preditor mais preciso?** Revista Energia na Agricultura 26 (2011): 49-55.

SANTOS, A.M.R.T., et al. **Detection of inconsistencies in geospatial data with geostatistics**. Boletim de Ciências Geodésicas.

SANTOS, A.M.R.T., et al. **Proposition of an outlier detection method for nonnegative variables through geostatistics**. Boletim de Ciências Geodésicas.

SCHLATHER, M., MALINOWSKI, A., MENCK, P.J., OESTING, M. AND STROKORB, K. **Analysis, simulation and prediction of multivariate random fields with package RandomFields**. Journal of Statistical Software, 63 (2015): 1-25.

YAMAMOTO, J., & LANDIM, P. **Geoestatística: Conceitos e Aplicações**. São Paulo, Oficina de Textos, (2013)

YFANTIS,E.A., FLATMAN, G.T., BEHEAR, J.V. **Efficiency of kriging estimation for square, triangular and hezagonal grids.** Math Geaol, 19 (1987): 183-205.

VIEIRA, S.R. **Geoestatística em estudos de variabilidade espacial do solo**. Tópicos em Ciências do Solo 1(2000): 1-54.

CONCLUSÕES GERAIS

O presente trabalho demonstra o potencial do uso da Geoestatística para o estudo de variáveis geoespaciais contínuas, pois incorpora a localização geográfica ao processo analítico dos dados. Além disso, é a metodologia ideal na caracterização e modelagem de fenômenos que apresentam dependência espacial.

A utilização dessa metodologia no estudo de dados discrepantes, denominados aqui de outliers, foi satisfatória, conforme os resultados apresentados nos três capítulos do trabalho. Como o mecanismo de interpolação, a krigagem, prediz valores em locais não amostrados, sem tendência e com variância mínima, com base nos pontos observados na região de interpolação, isso possibilita a detecção local de possíveis outliers.

Foram feitas três novas proposições metodológicas no estudo de outliers para variáveis geoespaciais contínuas autocorrelacionadas, buscando sempre a fundamentação teórica da Estatística.

No primeiro capítulo, ao submeter um conjunto de dados reais, sabidamente com outliers, não detectados pela metodologia do Box Plot, uma nova proposta de detecção de discrepâncias foi feita. Especificamente, a Geoestatística foi utilizada na modelagem e mapeamento dos dados, gerando resíduos que deveriam atender determinadas pressuposições teóricas. Através da estimação intervalar bilateral desses resíduos foi possível detectar a presença de outliers no conjunto amostral, e, ainda, a localização geográfica dos mesmos.

Como algumas variáveis não permitem a estimação intervalar bilateral, como é o caso de variáveis não-negativas, no segundo capítulo foi feita uma adaptação metodológica do Capítulo 1 para detecção de outliers utilizando a Geoestatística na modelagem e mapeamento dos dados, e, através de uma combinação de teoremas da Estatística Clássica, foi possível fazer a estimação intervalar unilateral dos resíduos, detectando os outliers desse tipo de variáveis.

No capítulo terceiro, utilizando os parâmetros de indicação de qualidade de ajuste da Geoestatística, foi proposta uma metodologia de orientação sobre a exclusão ou não dos dados, com alta probabilidade de serem outliers, do conjunto amostral original.

Apesar da grande utilização do mecanismo Box Plot para a detecção de outliers, foi possível concluir que este não se mostrou propriado para as variáveis geoespaciais estudadas.