AVALIÇÃO DE MÉTODOS DE REMOÇÃO DE OUTLIERS E SEUS IMPACTOS NA PRECISÃO DOS MÉTODOS DE INTERPOLAÇÃO.

Guilherme Mossi Bento[1*], Raul Teruel dos Santos[2]

[1,] Graduando de Sistemas de informação, Instituto de Computação, Universidade Federal de Mato Grosso, Cuiabá, Mato Grosso, Brasil. [2,] Doutorado, Instituto de Computação, Universidade Federal de Mato Grosso, Cuiabá, Mato Grosso, Brasil.
*E-mail: guilhermeissom@outlook.com

RESUMO: Os outliers são dados fora da normalidade que afetam negativamente a análise estatística e o entendimento das informações sendo necessário detectá-los e removê-los. No campo matemático há diversas formas de remover outliers, o que acaba tornando complexo o processo de escolha dos mesmos para a aplicação. Neste artigo serão abordados os métodos mais encontrados na literatura sendo eles Tukey, Standard Deviation e Zscore, e seus impactos na acurácia dos métodos de interpolação Spline e o Inverso da Distância ao Quadrado. A análise abordou a remoção dos outliers e seus impactos na precisão dos resultados da interpolação. A predição foi medida através do erro quadrático médio. Para todos os métodos que conseguiram remover outliers foi notório o aumento da acurácia.
Palavra-chave: Agricultura de precisão, Zscore, Boxplot, Standard Deviation.

The impact of outliers on the accuracy of interpolation methods.

ABSTRACT: *The outliers are data out of normality that negatively affects statistical analysis and the undertanding of information generating the need to detect and remove them. However in mathematical field there are many ways to remove outliers, thus making the choosing process complex. In this paper discusses the methods most commnly found in the literature, such as Tukey, Standard Deviation and Zscore, and their impact on the accuracy of the interpolation methods Spline and the Inverse Distance to Square. The analysis addressed the correction and non-correction of outliers. The quality of the prediction was measured using the mean square error. For all methods that were able to remove outliers, the increase in accuracy was notorious.*
Keywords: *Precision Agriculture, Zscore, Boxplot, Standard Deviation.*

1. INTRODUÇÃO

Devido ao alto custo por amostra de solo coletada, em busca de economia, o proprietário busca extrair poucas amostras, portanto é necessário métodos de interpolação para aumentar a densidade amostral do mapa, considerando a dependência espacial para interpolar valores em locais não amostrados, e assim construir mapas com maior densidade amostral a partir de grades amostrais menos densas (GREGO et al., 2014). Porém, dados amostrados podem conter erros ocasionados pelo processo de coleta, reduzindo assim a precisão dos mapas obtidos a partir dos métodos de interpolação, acarretando em prejuízos tanto financeiros quanto ambientais.

Segundo a definição de Silva (2004) um outlier é uma amostra que desvia de um padrão do conjunto de dados. Eles são geralmente causados por erro humano, como erros de coleta, gravação ou de entrada e é necessário o tratamento desses para não interferir negativamente na acurácia (OSBOME; OVERBAY, 2004). Entretanto há vários métodos para remoção dessas anomalias no âmbito matemático, sendo difícil escolher o método para usar no conjunto de dados. O objetivo desse trabalho é uma análise comparativa entre os métodos de remoção de outliers mais encontrados na literatura, e o seu impacto na acurácia dos métodos determinístico de interpolação *Inverse distance weighted* (IDW) e *Spline*.

2. MATERIAL E MÉTODOS

Os dados usados nessa pesquisa foram retirados de um talhão de 133,96 ha, próximo ao município de Tibagi no estado do Paraná contendo 62 pontos de coleta. Para o estudo foram escolhidos os atributos PH, pois segundo Faraco et al. (2001) a acidez dos solos é um problema mundial que afeta negativamente as áreas agrícolas nas atividades microbiota do solo e o Magnésio (Mg) por sua função de ativar a produção dos fotossimiladores necessários para a planta se desenvolver (Hawksford et al., 2012). Devido ao fato de apresentarem diferentes variâncias e desvio padrão foi possível observar os comportamentos dos métodos, essa variação dos dados pode ser observada na Tabela 1 referente aos dados de PH e Tabela 2, referente aos dados de Magnésio.

Tabela 1. Analise estatística descritiva dos dados de PH.

Medidas	Valores
Mínimo	4,300
Média	5,195
Mediana	5,200
Máxima	5,900
Desvio padrão	0,358241
Variância	0,128336

Tabela 2. Analise estatística descritiva dos dados de Mg.

Medidas	Valores
Mínimo	5,200
Média	17,740
Mediana	18,150
Máxima	30,500
Desvio padrão	4,73737
Variância	22,4427

Após a escolha do atributo, os pontos foram divididos em 10 grupos de pontos para aplicação da validação cruzada, a qual se caracteriza por remover temporariamente o valor da amostra original e feito a interpolação sem aquele dado descartado, após a interpolação é comparado o valor da amostra original juntamente com o valor estimado pelo método de interpolação, (ISAAKS & SRIVASTAVA, 1989 apud FARACO et al., 2008), é feito a comparação dos dois dados e para estimar a acurácia essas informações são aplicadas no Erro Quadrático Médio (Eq. 1) ou *Root Mean Square Error* (RMSE), que segundo Hallak e Pereira Filho (2011) esse método de análise de acurácia é comumente usada por resultar em valores nas mesmas dimensões da variável analisada.

$$V = \sqrt{\left[\frac{\sum_{i=1}^{N}(d-d_i)^2}{N}\right]} \quad (1)$$

em que: N é a quantidade de resultados, V é o valor do RMSE obtido, d *e* d_i são respectivamente o dado obtido pela validação e o dado original.

Essa escolha de grupo foi adotada para não ocorrer pontos vizinhos pertencentes ao mesmo grupo, evitando assim grandes lacunas de dados que poderiam impactar no processo de interpolação pela utilização da validação cruzada. A Figura 1 apresenta a distribuição dos grupos definidos para validação cruzada.

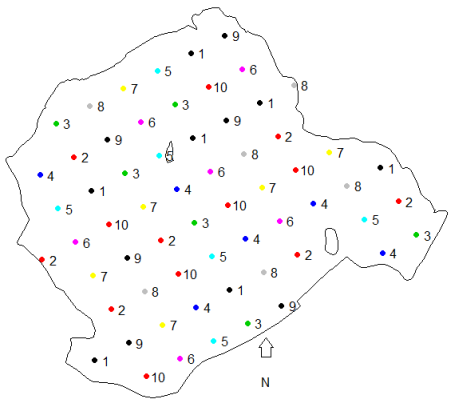

Figura 1. Grupos definidos para a validação cruzada.

Para a remoção de outliers foram escolhidos os métodos Tukey, Standard Deviation e Zscore, os quais são os mais encontrados na literatura (SEO, 2006).

O método de Tukey ou mais conhecido como boxplot define limites inferior (Eq. 2) e superior (Eq. 3) a partir do interquartil (IQR) e os primeiros e terceiros quartis.

$$L_{inf} = Q1 - (1.5 * IQR) \quad (2)$$

$$L_{sup} = Q3 + (1.5 * IQR) \quad (3)$$

em que: Q1 é o primeiro quartil, Q3 o terceiro quartil e o L_{sup} limite superior e L_{inf} limite inferior.

Os dados que estiverem fora desses limites serão determinados como outlier (ROUSSEEUW; HUBERT, 2017).

Segundo Seo (2006) o método Standard Deviation (SD) (ou desvio padrão tradução livre) (Eq. 4) se baseia na seguinte formula:

$$X \pm 2\,SD \quad (4)$$

em que: X é a média dos dados e SD é o desvio padrão calculado pela raiz quadrada da variância.

É frequentemente usado segundo Leys et al. (2013), 2 ou 2.5 como limites dependendo da análise que o pesquisador precisa, sendo qualquer valor maior que esses caraterizados de outliers, entretanto, esse método é sensível a valores extremos, podendo não identificar possíveis outliers, isso por usar a média como uma das etapas para o cálculo.

O ZScore (Eq. 5) pode ser descrito como (X_i) sendo o resultado dos dados conhecidos, subtraído pelo valor médio do conjunto de dados ($\bar{X}$), o resultado disso é dividido pelo desvio padrão (DP), para uma análise mais precisa é recomendado usar os valores iguais ou menores que 2 para o resultado de Z (TRIPATHY et al., 2013).

$$Z = \frac{(X_i - \bar{X})}{DP} \quad (5)$$

Os métodos de interpolação mais encontrados são IDW, Spline e Krigagem, entretanto nesse trabalho será tratado somente de métodos determinísticos.

O método de interpolação inverso do quadrado da distância ou mais conhecido como *Inverse Distance Weighting* (IDW) calcula os valores não amostrados usando a média ponderada dos valores dos pontos conhecidos. Esse método foi criado usando a média ponderada e essa média é calculada pelo inverso da distância dos pontos. A equação foi definida por Shepard (1968, apud Mei et al., 2017), para calcular o peso (Eq. 6), e após calculado o peso será aplicado na equação geral do método (Eq.7).

$$\mathrm{w(x)} = \frac{1}{d(x,x_i)^{\alpha}} \quad (6)$$

em que: x indica o ponto que quer ser amostrado, e x_i é o ponto do dado amostrado e o valor de α usualmente é igual a dois, mas é um valor arbitrário.

$$Z_o = \frac{\sum_{i=1}^{N} z_i.\mathrm{w(x)}}{\sum_{i=1}^{N} \mathrm{w(x)}} \quad (7)$$

em que: o N é o número total de pontos usados na interpolação, Z_1 significa o valor da amostra no ponto i e Z_o é o valor estimado da variável Z no ponto i.

Segundo Mazzini e Schettini (2009), o IDW é considerado um interpolador exato ou suavizante variando no coeficiente α.

Quanto ao *Spline* bivariado ou TPS (*Thin-Plate Spline*), há uma grande vantagem de usá-lo em regiões que pequena variância, assim ele com poucos pontos pode produzir resultados mais precisos que outros métodos e a superfície criada por ele é visualmente satisfatória mesmo usando poucos pontos (Wu & Hung, 2016). Entretanto, segundo Yang (2015) e Wu & Hung (2016) sua função por ser de grande suavização dos dados, gerando dados imprecisos quando há pontos extremos, devido sua equação ser gerada da integral do quadrado da segunda derivada.

Primeiramente os dados foram interpolados antes da remoção de outliers, feito isso foi levada em consideração para a análise da acurácia o método de validação cruzada dos dados.

Após isso foram aplicados cada método de remoção de outlier mencionado e novamente realizada a interpolação sem os outlier, com essa mudança nos dados foi refeito a análise estatística dos dados, calculando a variância e o desvio padrão, e para cada método de interpolação foram coletadas as informações dos RMSE, dessa forma podendo comparar a exatidão dos métodos de interpolação antes e depois do tratamento de outliers.

3. RESULTADOS E DISCUSSÃO

Com a aplicação dos métodos de remoção de outliers foram coletadas as informações sobre as diferenças nos dados e nos métodos de interpolação.

Após a remoção de outliers foi observado a diferença na variância e no desvio padrão, essas informações e a quantidade de outliers removidos são apresentados nas Tabelas 3 e 4, PH e Mg respectivamente.

Tabela 3. Resultado da remoção de outliers para o PH.

Método	Qtd. outliers	Variância	Desv. Padrão
SD	2	0,107581	0,327996
Zscore	1	0,327996	0,341908
Tukey	0	*	*

*sem mudanças nos dados

Tabela 4. Resultado da remoção de outliers para o Mg.

Método	Qtd. outliers	Variância	Desv. Padrão
SD	3	15,71216	3,963856
Zscore	5	13,18034	3,630474
Tukey	3	15,71216	3,963856

Com essa diferença nos dados foi possível observar a diferença o aumento da acurácia, ou seja, a diminuição do RMRE nos métodos de interpolação que são demonstrados nas Tabelas 5 e 6.

Tabela 5. RMSE dos dados de PH para os métodos de interpolação.

Método de interpolação	Com outlier	SD	Tukey	Zscore
IDW	0,26877	0,25618	*	0,26648
SPLINE	0,29898	0,28980	*	0,29643

*valores não alterados

Tabela 6. RMSE dos dados de Mg para os métodos de interpolação.

Método de interpolação	Com outlier	SD	Tukey	Zscore
IDW	4,14604	3,76517	3,76517	3,41562
SPLINE	4,92040	4,49049	4,49049	4,12617

O método SD nos dados de PH removeu 50% a mais de outliers que o Zscore, entretanto nos dados de Mg, acabou removeu 60% a menos que o método Zscore resultando num ganho de aproximadamente 9% de acurácia nos dois métodos de interpolação.

O método Tukey funcionou de modo estável, ou seja, se adaptou de acordo com os dados não sendo afetado por valores extremos, isso por usar medidas robustas no seu cálculo, mas em dados com variação, possíveis outliers não foram identificados removendo a mesma quantidade de outliers que o SD nos dados de Mg, seu ganho em exatidão nesses dados foi igual ao ganho do método SD.

O Zscore mesmo usando como parâmetros o desvio padrão e média dos dados, funcionou de forma estável diferente do Método SD em que o cálculo é pela divisão entre a diferença do erro e o desvio padrão.

O Zscore conseguiu encontrar cinco outliers nos dados de Mg tendo um ganho de aproximadamente 17,6% na acurácia dos métodos de interpolação e um outlier nos dados de PH e assim um ganho de apenas 1% em ambos os métodos.

Portanto nessas condições, o método Zscore teve melhor desempenho nos dados de Mg, aumentando a exatidão dos dados interpolados cerca de 50% a mais que os outros métodos.

4. CONCLUSÕES

Após a remoção de outliers foi comprovado que o método SD tem uma grande diferença na sua estabilidade quando encontra um valor extremo, por usar a média como um parâmetro base, isso pode ser observado quando nos dados de PH removeu dois outliers superando os outros métodos. Entretanto quando aplicado nos dados de Mg, o método removeu menos que o Zscore, isso porquê os dados de Mg possuem dados extremos. Seu aumento na precisão foi de 4,6% passando de 0,26877 com outliers para 0,25618 após remove-los com esse método.

Já o Zscore, nos dados de PH removeu apenas um outliers e a acurácia do método Zscore foi menor que o método SD nessa análise, mas quando aplicado aos dados de Mg sua acurácia foi a maior entre os 3 métodos comparados chegando a quase 17,6% de ganho, ou seja, de diminuição no RMSE, que diminuiu de 4,14604 para 3,76517.

O método Tukey se mostrou o mais estável dos três métodos removendo nenhum outlier no primeiro conjunto de dados e 3 no segundo, isso por usar métodos mais complexos de cálculo removendo apenas os pontos mais extremos, mesmo assim foi possível um aumento na exatidão de 10%.

Para a escolha do método mais adequado deve ser feito uma observação nos dados analisando se possui uma grande variância, assim tornando inviável o uso do método SD. O método Zscore por usar também como parâmetro dos seus cálculos medidas simples, pode ocorrer uma variação em outras situações de dados, mas é um método

mais estável que o SD. Esse método teve a melhor performance nos dados de Mg, ou seja, mesmo com maior variância, aumentou a exatidão dos dados interpolados em uma diferença de aproximadamente 50% a mais que os métodos Tukey e SD.

Em geral, foi observado que mesmo os outros métodos que tiveram um baixo desempenho na remoção de outliers, aumentaram a eficácia dos métodos de interpolação, dessa forma evitando que insumos ou informações sejam interpretados de forma errônea.

6. REFERÊNCIAS

Osborne, J. W., & Overbay, A. The Power of Outliers (and Why Researchers Should Always Check for Them). **Practical Assessment, Research & Evaluation**. North Carolina State University. Raleigh, USA, V. 9, N. 6, 2004.

SILVA, F. R. **Uma Abordagem para Detecção de Outliers em Dados Categóricos**. 2004. 81f. Dissertação (Mestrado em Ciência da Computação) – Universidade Estadual de Campinas UNICAMP, Campinas, 2004.

FARACO, M.A.; URIBE-OPAZO, M.A.; SILVA, E.A.A.; MIGUEL, D. L.; MOREIRA, F. M. S.. Influência do pH do meio de cultivo e da turfa no comportamento de estirpes de Bradyrhizobium. **Rev. Bras. Ciênc. Solo**, Viçosa, v. 25, n. 4, p. 873-883, Dec. 2001

FARACO, M. A., URIBE-OPAZO, M. A., SILVA, A. A., JOHANN, J. A., BORSSOI, J. A. Seleção de modelos de variabilidade espacial para elaboração de mapas temáticos de atributos físicos do solo e produtividade da soja**. Revista Brasileira de Ciência do Solo**. Viçosa, MG, v. 32, n. 2, p. 463-476, mar./abr. 2008.

HALLAK, Ricardo; PEREIRA FILHO, Augusto José. Metodologia para análise de desempenho de simulações de sistemas convectivos na região metropolitana de São Paulo com o modelo ARPS: sensibilidade a variações com os esquemas de advecção e assimilação de dados. **Rev. bras. meteorol.** São Paulo, v. 26, n. 4, p. 591-608, Dec. 2011.

MAZZINI, P. L. F.; SCHETTINI, C. A. F.Avaliação de metodologias de interpolação espacial aplicadas a dados hidrográficos costeiros quase sinóticos. **Brazilian Journal of Aquatic Science and Technology**, v.13, n.1, p.53-64, 2009.

Mei, G.; Xu, L.; Xu, N. Accelerating adaptive inverse distance weighting interpolation algorithm on a graphics processing unit.**R. Soc. open sci.** 2017.

ROUSSEEUW, Peter J.; HUBERT, Mia. Anomaly detection by robust statistics. **WIREs Data Mining and Knowledge Discovery**. Lovaina, Belgium, p. 3-4, nov. 2017.

SEO, Songwon. **A Review and Comparison of Methods for Detecting Outliers in Univariate Data Sets.** 2006. 56 p. Dissertation (Master of Science) - Graduate School of Public Health in partial fulfillment, UNIVERSITY OF PITTSBURGH, Pittsburgh, Pennsylvania, 2006.

YANG, Mengxi. **Benchmarking railfall interpolarion over the nethelands.** 2015. 59 p. Dissertation (Master of Science in Geo-information Science and Earth Observation) – Water Resourses and Environmental Management, University of Twente, Enschede, The Netherlands, 2015.

Wu, Yi-Hwa (Eva); Hung, Ming-Chih. **Comparison of Spatial Interpolation Techniques Using Visualization and Quantitative Assessment**. Department of Humanities and Social Sciences, Northwest Missouri State University, University Drive, Maryville, Missouri, USA. Chapter 2, 2016.

Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. Detecting outliers: do not use standard deviation around the mean, use absolute deviation around the median. **Journal Of Experimental Social Psychology**. 764–766, 2013.

Tripathy S. S., Saxena R. K., Gupta P. K., Comparison of Statistical Methods for Outlier Detection in Proficiency Testing Data on Analysis of Lead in Aqueous Solution, **American Journal of Theoretical and Applied Statistics**. Vol. 2, No. 6, pp. 233-242, 2013.

Hawksford M, Horst W, Kichey T, Lambers H, Schjoerring J, Moller IS, White P. **Functions of macronutrients.** In: Marschner P. ed. Marschner's mineral nutrition of higher plants. 3rd ed. Elsevier; 2012. p.135-189.

GREGO, Célia R.; OLIVEIRA, Ronaldo P; VIEIRA, Sidney R**. Geoestatística aplicada a agricultura de precisão. Agricultura de precisão: resultados de um novo olhar.** Brasília, DF: Embrapa, n. Cap 5, p. 74-83, dez. 2014.