DESEMPENHO DE MÉTODOS DE FILTRAGEM PARA DADOS DE PRODUTIVIDADE EM CANAVIAIS UTILIZANDO DADOS DISCREPANTES ARTIFICIAIS

YURI DE LACERDA BARBOSA[1], EUDOCIO RAFAEL OTAVIO DA SILVA[2], MARCELO CHAN, FU WEI[2], JOSÉ PAULO MOLIN[3]

[1] Graduando em Eng° Agronômica, Lab. de Agricultura de Precisão, Depto. Engenharia de Biossistemas, Escola Superior de Agricultura Luiz de Queiroz, ESALQ, Universidade de São Paulo, Piracicaba – SP, Fone: (19) 996046475, yurilacerdabarbosa@usp.brEng.
[2] Eng. Agrônomo, Doutorando, Lab. de Agricultura de Precisão, Depto. de Engenharia de Biossistemas, Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba – SP.
[3] Eng. Agrícola, Professor Titular, Lab. de Agricultura de Precisão, Depto. de Engenharia de Biossistemas, Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba – SP.

Apresentado no
Congresso Brasileiro de Agricultura de Precisão e Digital - ConBAP 2024
Porto Alegre, RS, 2 a 4 de julho de 2024

RESUMO: O mapeamento da produtividade vem adquirindo importância no setor canavieiro. Como em outras culturas, os dados coletados devem ser submetidos a processos de detecção e remoção de erros. Neste trabalho avalia-se o desempenho de métodos de filtragem para dados de produtividade em canaviais utilizando dados discrepantes artificiais. O conjunto de dados é proveniente de quatro talhões comerciais de cana-de-açúcar obtidos na colheita de 2021. Para isto, foram inseridos dados discrepantes artificias aos dados brutos de maneira aleatória. As filtragens são baseadas em cinco métodos estatísticos existentes: Z-score, desvio padrão, intervalo interquartílico, critério de Chauvenet além de um baseado em análise global e local anisotrópica e isotópica, o aplicativo MapFilter 2.0. Os dados filtrados foram submetidos a estatística descritiva para visualização e os métodos foram comparados baseados no seu desempenho de detecção de dados discrepantes artificiais. O MapFilter 2.0 teve o melhor desempenho em identificar os dados discrepantes artificiais seguido pelo método de Chauvenet. Os outros métodos apresentaram um desempenho igual e inferior a 50,00% na detecção de dados discrepantes. Esses desempenhos refletiram fielmente o resultado da filtragem dos dados brutos originais. Mais estudos são necessários para determinar se todos os dados filtrados eram de fato discrepantes com a realidade.

PALAVRAS-CHAVE: Mapfilter, sensores de colheita, Chauvenet.

PERFORMANCE OF FILTERING METHODS FOR YIELD DATA IN SUGARCANE FIELDS USING ARTIFICIAL OUTLIERS

ABSTRACT: Yield mapping has been gaining importance in the sugarcane sector. As in other crops, the collected data must undergo processes of error detection and removal. In this study, the performance of filtering methods for yield data in sugarcane fields using artificial outliers is evaluated. The dataset comes from four commercial sugarcane plots obtained during the 2021 harvest. For this purpose, artificial outliers were inserted into the raw data randomly. Filtering is based on five existing statistical methods: Z-score, standard deviation, interquartile range, Chauvenet's criterion, as well as one based on global and local anisotropic and isotropic analysis, the MapFilter 2.0 application. The filtered data were subjected to descriptive statistics for visualization, and the methods were compared based on their performance in detecting artificial outliers. MapFilter 2.0 achieved the best performance in identifying artificial outliers followed by Chauvenet's method. The other methods showed performance equal to or less than 50.00% in detecting outliers. These performances accurately reflected the result of filtering the original raw data. Further studies are needed to determine if all the filtered data were indeed outliers in reality.

KEYWORDS: Mapfilter, Harvest sensors, Chauvenet

INTRODUÇÃO: Os mapas de produtividade são uma importante ferramenta de agricultura de precisão, no qual reflete respostas das plantas às práticas de manejo, permite realizar caracterização qualitativa e quantitativa

da produção (LUND et al., 2016), identificar variações e investigar fatores que as afetam (HIGGINS et al., 2019), e auxilia na tomada de decisão e implementação de estratégias de manejo específicas para cada local (MONTEIRO et al., 2021). A implementação de sistemas automatizados em colhedoras para o monitoramento da produtividade resulta na coleta de uma grande quantidade de dados. No entanto, é importante exercer cuidado ao utilizar esses dados, pois nem todos os pontos registrados refletem a produtividade real. Dados discrepantes (em inglês, *outliers*) são aqueles que diferem das outras observações do mesmo conjunto de dados e podem ser atribuídos a erros de medição ou registro, como os causados pelo tempo de demora entre o corte e o enchimento da colhedora, erros de amostragem e há casos de valores excepcionais, porém verdadeiros (LYLE et al., 2014; MALDANER et al., 2021). Logo, cabe ao analista de dados o desafio de remover esses dados, visto que podem distorcer a análise, ou mantê-los pois podem representar um fenômeno que esteja acontecendo no campo (SMITI, 2020). Frente ao desafio da ausência de uma referência que defina o que é um dado discrepante, alguns estudos, como o de Liu et al. (2021), têm inserido dados discrepantes artificiais, que são valores conhecidos e inseridos pelo usuário do arquivo de dados, no conjunto de dados.

Não existe, atualmente, um consenso sobre qual o melhor método a ser utilizado na filtragem de dados oriundos de cana-de-açúcar. Sendo assim, este estudo tem como objetivo avaliar o desempenho de métodos de filtragem para dados de produtividade em canaviais utilizando dados discrepantes artificiais, tendo como objetivos específicos: i) realizar a filtragem de dados de produtividade de cana-de-açúcar utilizando cinco diferentes métodos de filtragem de dados; e ii) comparar seus desempenhos estatisticamente e espacialmente entre si.

MATERIAL E MÉTODOS: O conjunto de dados é proveniente da colheita em 2021 de quatro talhões (23,1; 23,8; 30,1 e 12,9 ha) cultivados com a variedade de cana-de-açúcar RB966928 em solo classificado como Latossolo Vermelho Álico localizados no estado de São Paulo, Brasil. As colhedoras, todas de fileira única, estavam equipadas com Sistema Global de Navegação por Satélite (GNSS) e com sistema de sensor comercial (Solinftec, Araçatuba, Brasil), como o usado por Maldaner; Canata e Molin (2022) na avaliação da acurácia de previsão da massa de cana-de-açúcar por um sensor de pressão de propulsão instalado no tambor picador de uma colhedora. A lavoura estava no seu terceiro corte e seu estado de maturação precoce. Esses dados estavam divididos de acordo com suas respectivas áreas.

O conjunto de dados original de cada talhão foi contaminado por dados discrepantes artificiais, baseados em Liu et al. (2021). A inserção desses dados foi realizada a fim de conferir o desempenho dos filtros usados, já que no banco de dados original nem sempre é possível julgar se o dado filtrado era ou não representativo da realidade. Os dados foram inseridos conforme a equação 1.

$$AO = Z + OM * Z \quad (I)$$

em que: AO - valor periférico artificial; Z - valor original de um elemento; OM - valor de magnitude do dado discrepante, que podem ser: ±0,01, ±0,05, ±0,10, ±0,50, ±0,80, ±1,00.

A seleção de Z, ou seja, dos valores das linhas no banco de dados foi aleatória. Após filtrar com cada método os dados originais contaminados com os dados discrepantes artificiais, foi então calculado, a partir da quantidade de dados "AO" remanescentes, o desempenho na detecção de outliers de cada método. Esta etapa foi realizada em ambiente virtual JupyterLab, utilizando a linguagem de programação Python v. 3.10.5 (KLUYVER et al., 2016).

Os métodos de filtragem de dados selecionados foram: Z-Score (método 1) (HODGE & AUSTIN, 2004), MapFilter 2.0 (método 2) (MALDANER et al., 2021), desvio padrão (método 3) (BECK & KÜHN, 2017), intervalo interquartílico (método 4) (JEONG et al., 2017) e Critério de Chauvenet (método 5) (LIN & SHERMAN, 2007). O método 1 utiliza os cálculos da média e desvio padrão dos dados medidos como definição de outlier. Se a diferença entre aquele dado e a média for maior que X vezes o desvio padrão, então aquela medição é definida como outlier (HODGE & AUSTIN, 2004). O método 2 é um método robusto de análise de dados de alta densidade, processado através do programa MapFilter 2.0. Neste filtro são realizadas análise global dos dados e uma análise local isotrópica e anisotrópica levando em consideração os valores da vizinhança. A análise global remove os dados fora do limite superior e inferior, e estes limites são calculados conforme as equações 2 e 3.

Limite superior = mediana + (Limite de variação (%) ∗ mediana) (2)
Limite inferior = mediana − (Limite de variação (%) ∗ mediana) (3)

em que: mediana - calculada a partir do parâmetro que está sendo filtrado; e limite de variação -valor em porcentagem selecionado dentro do programa.

A análise local anisotrópica é feita calculando uma mediana a partir da seleção de todos os pontos que estão dentro da mesma fileira, e a partir dessa mediana calcula-se os limites superiores e inferiores locais para cada fileira. Após essa filtragem local por linhas, é feita uma segunda filtragem, a isotrópica, onde define-se o valor do raio de dependência espacial para cada ponto. A partir disso, é calculada uma nova mediana, novo limite superior e inferior local (repetindo a mesma fórmula anterior) considerando todos os pontos dentro do raio de cada ponto analisado. Os pontos são removidos caso estejam fora desses novos limites calculados. O método 3 é calculado a partir da equação 4:

desvio padrão: $\sqrt{\frac{\sum(x_i - \bar{x})^2}{n}}$ (4)

em que: x; -valor observado; - média daquele conjunto de dados; e n - número de dados daquele conjunto.

A partir disso foram criados dois limites, um superior (média + desvio padrão) e um inferior (média - desvio padrão), para definir o limiar do que é ou não dados discrepantes (BECK & KÜHN, 2017). O método 4 utiliza valores estimados por métodos de regressão, definindo seus valores limites de aceitação ou rejeição do valor medido (JEONG et al., 2017). Os limites superior e inferior nesse método são definidos conforme as equações 5 e 6.

limite superior = q3 + 1,5 iqr* (5)
limite inferior q1 - 1,5 iqr (6)

em que: q1 - valores que se encontram nos primeiros 25% dos dados ordenados (primeiro quartil); q3 - valores que estão nos últimos 25% dos dados ordenados (terceiro quartil);

O valor de iqr que é definido conforme a equação 7.

iqr = q3 - q1 (7)

O método 5 utiliza o desvio padrão, a média e o tamanho conhecido de uma amostra de dados para definir se um valor medido é aceitável ou não (LIN & SHERMAN, 2007). Este método pode ser definido de a acordo com a equação 8.

Se x $\frac{|x_i - \mu|}{\sigma} > R$, *rejeite* x_i (8)

em que: R-valor tabelado pelo critério de Chauvenet para a quantidade de amostras utilizadas; x; - tamanho conhecido de uma amostra; μ- média; e σ - desvio padrão.

A filtragem dos dados foi realizada no ambiente R (R CORE TEAM, 2023), com exceção do método 2, na qual possui interface própria. Foi realizado cálculo estatistico para os dados filtrados indicando sua média, desvio padrão (DP), coeficiente de variação (CV), valor máximo e mínimo para cada método e área estudada. Os dados filtrados foram inseridos no Sistema de Informação Geográfico QGIS 3.28.15 e, então, seus respectivos mapas foram gerados.

RESULTADOS E DISCUSSÃO: A estatística dos dados está descrita na Tabela 1 e a partir dela é possível observar que os dados brutos de todas as áreas apresentam valores minimos iguais a zero, valores máximos elevados para a cultura da cana-de-açúcar, chegando a valores superiores a 600,00 Mg ha e elevados

coeficientes de variação. O cenário mudou após esses dados serem filtrados e seu comportamento variou de acordo com cada método.

O método de filtragem que apresentou os maiores valores de coeficiente de variação foi o método 4, seguido do método 1. O método 4 utilizou os maiores valores de limite superior e os menores de limite inferior, resultando e classificando menores valores determinados como discrepantes. Um comportamento que se pode observar é de que quanto maior a porcentagem de remoção de dados por cada método, menor o coeficiente de variação resultante da filtragem do mesmo.

TABELA 1, Estatística descritiva para os dados brutos e filtrados de cada área nos diferentes métodos de filtragem.

Área 1						
		Método				
Parâmetros	Brutos	1	2	3	4	5
n	19807	14145	8870	12148	14490	9875
% de remoção	0,00	28,59	55,22	38,67	26,84	50,14
Máximo	333,88	129,69	92,74	102,62	138,12	90,60
Mínimo	0,00	25,96	49,97	28,92	16,68	40,79
Média	65,69	78,78	73,31	73,54	79,57	70,60
DP	36,94	20,63	10,09	15,70	22,07	11,84
CV	56,23	26,19	13,76	21,35	27,73	16,77
Área 2						
n	21711	15752	8620	12975	16141	10562
% de remoção	0,00	27,45	60,30	40,24	25,66	51,35
Máximo	670,12	141,55	95,60	109,91	157,35	97,07
Mínimo	0,00	21,44	51,56	31,10	5,29	43,89
Média	70,46	83,39	74,05	76,42	84,35	73,34
DP	39,46	24,42	10,63	17,76	26,16	13,30
CV	56,00	29,28	14,36	23,25	31,01	18,13
Área 3						
n	24949	17892	11226	15190	18416	11886
% de remoção	0,00	28,29	55,00	39,12	26,19	52,36
Máximo	528,02	132,03	99,24	106,95	145,78	94,52
Mínimo	0,00	33,10	53,49	30,69	19,43	43,11
Média	68,77	84,59	78,98	78,67	85,05	74,71
DP	38,18	20,39	11,00	16,42	22,10	12,45
CV	55,53	24,10	13,93	20,87	25,99	16,67
Área 4						
n	14071	8566	5265	7394	8710	5596
% de remoção	0,00	39,12	62,58	47,45	38,10	60,23
Máximo	649,69	139,70	100,66	109,94	149,50	96,19
Mínimo	0,00	31,36	54,32	25,72	21,20	39,14
Média	67,65	87,33	81,72	81,67	87,61	76,19
DP	42,33	20,74	10,91	17,11	21,88	13,43
CV	62,58	23,75	13,35	20,95	24,98	17,63

n: número de observações; % de remoção: porcentagem de dados removidos; DP: Desvio padrão; CV: Coeficiente de variação.

Na Figura 1 é possível visualizar espacialmente o lato coeficiente de variação dos dados brutos como indicado na Tabela 1.

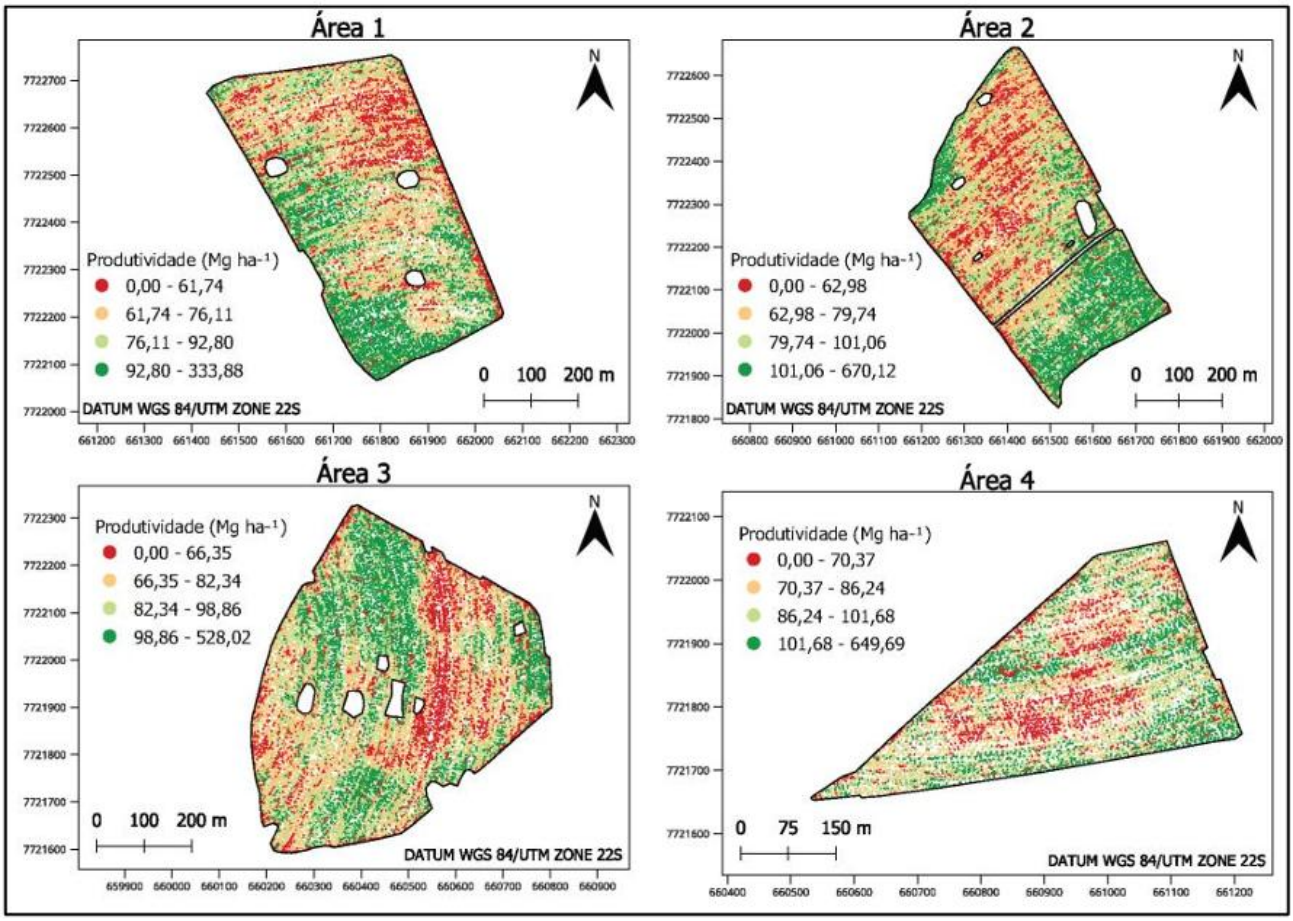

FIGURA 1. Mapas de pontos da produtividade de cana-de-açúcar em todas as áreas com seus respectivos dados brutos, sem a adição de dados discrepantes artificiais.

A Figura 2 mostra a espacialidade dos dados sem os dados discrepantes artificiais de cada área após serem filtrados pelos cinco métodos e nela é possível visualizar espacialmente como o método 2 se diferenciou dos outros. Observando a Figura 2 e a Tabela 1, pode-se afirmar que o método 2 foi o que mais filtrou os dados, utilizando limites superiores menores e inferiores maiores em todas as áreas analisadas. Como resultado, tem-se os maiores valores de mínimo e menores para os de máximo, desvio padrão e coeficientes de variação. Porém, observando a baixa amplitude de dados resultantes, há a possibilidade de que nem todos os dados filtrados eram discrepantes.

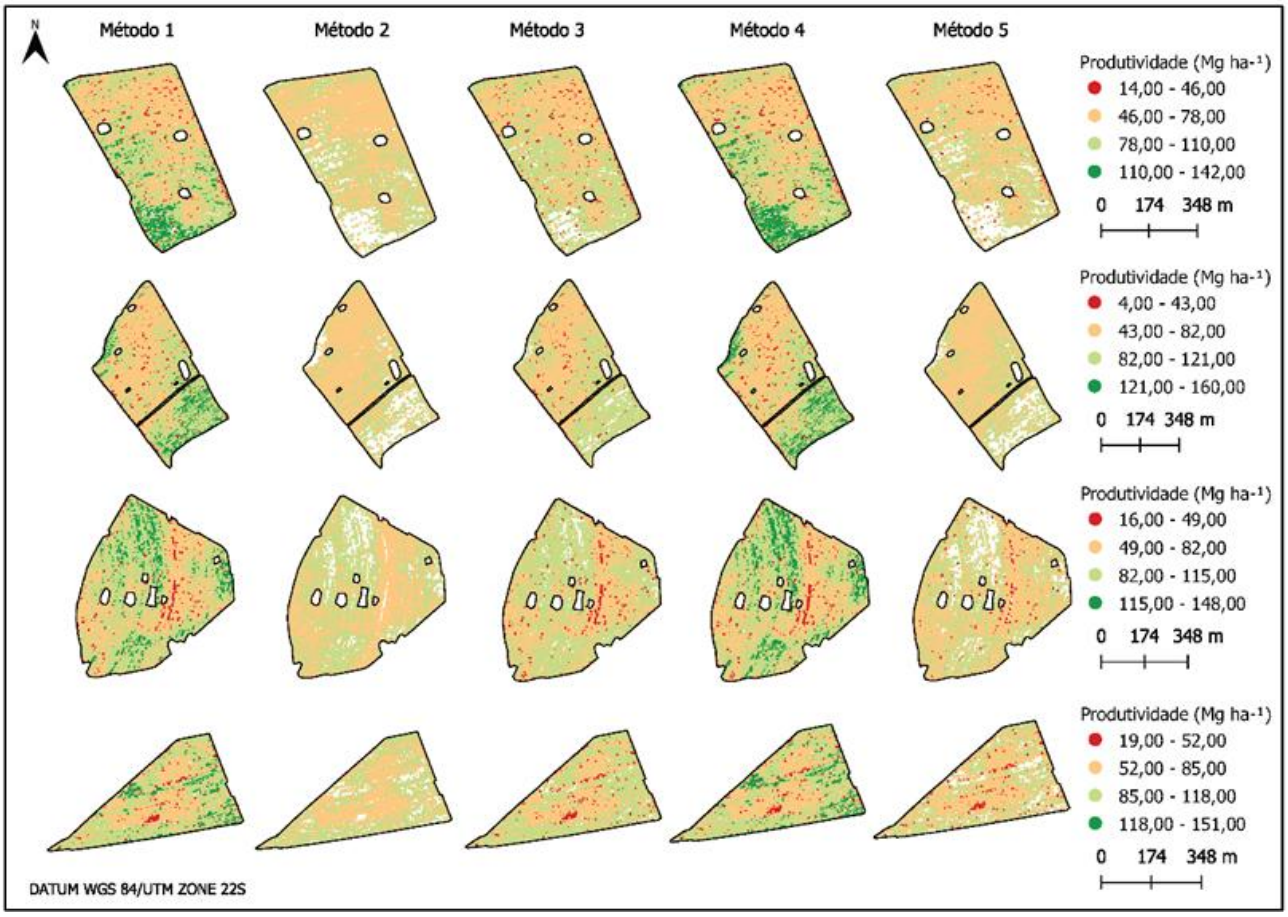

FIGURA 2. Mapas de pontos da produtividade de cana-de-açúcar em todas as áreas com seus respectivos dados filtrados pelos respectivos métodos, sem a adição de dados discrepantes artificiais.

O desempenho de cada método de filtragem encontra é mostrado na Figura 3 e é baseado na porcentagem de dados discrepantes artificiais que cada método foi capaz de identificar e remover. Pode-se afirmar que um comportamento em comum entre todos os métodos de filtragem foi de que quanto maior o valor de OM (magnitude de dado discrepante), maior a porcentagem de detecção alcançada pelos métodos e, consequentemente, maior seu desempenho.
Em média, os métodos 1 e 4 foram os que atingiram as menores porcentagens de desempenho, sendo na maioria das vezes incapazes de atingir um desempenho superior a 90% mesmo com o valor de ± 1,00 de OM. Estes dois métodos foram apenas capazes de atingir um desempenho de detecção acima de 20,00% nas áreas 1, 2 e 3 quando o valor de OM era igual ou superior a ± 0,50. O método 4 quase sempre apresentou um desempenho inferior ao do método 1, sendo igual somente duas vezes e nunca maior.

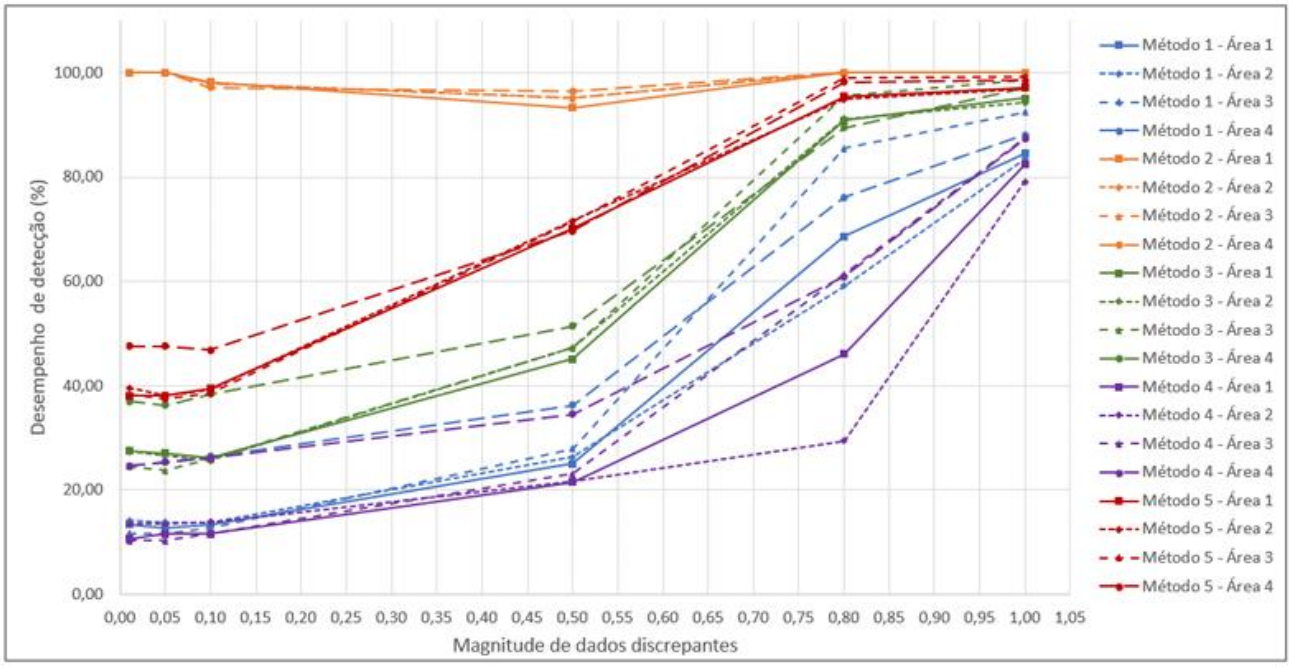

FIGURA 3. Desempenho de detecção de dados discrepantes artificiais (%) pelos métodos de filtragem propostos.

Sendo assim, os métodos 1 e 4 foram os de menor desempenho de detecção de dados discrepantes artificiais neste estudo, isso é esperado pois ambos os métodos são eficazes apenas para dados simetricamente distribuídos em torno de uma média constante (CASTELEYN et al., 2023; DASH et al., 2023). Os métodos 2, 3 e 5 foram os de maior desempenho neste estudo. O método 3 apresentou um desempenho acima de 90,00% apenas para valores de magnitude de dados discrepantes acima de ± 0,80, sendo incapaz de apresentar um desempenho acima de 38,00% em todas as áreas para valores de ± 0,01 OM, ultrapassando esse valor apenas para OM acima de ± 0,10. O método 5, por outro lado, alcançou e ultrapassou essa porcentagem de desempenho para os valores de menor magnitude de dados discrepantes (± 0,01 de OM), porém apresentou um desempenho acima de 90% apenas nos maiores valores de magnitude de dados discrepantes. O método 2 atingiu os maiores desempenhos de detecção de dados discrepantes neste estudo, sendo sempre capaz de atingir 100% de desempenho para valores de magnitude (OM) de ± 0,01, ± 0,05, ± 0,80 e ± 1,00. Em nenhum momento este método apresentou uma porcentagem de desempenho acima de 99,00% para valores de + 0,10e + 0,50 de OM, porém sempre se manteve entre 93,00 e 98,00%.

Essa discrepância de desempenho entre o método 2 e os outros métodos se deve ao fato de que o método 2 realiza análises locais, diferentes dos outros métodos baseados puramente em estatística. Logo, em estudos futuros é necessário compará-lo com métodos de filtragem baseados em distância, cluster, agrupamento e com janela móvel.

CONCLUSÃO: Todos os cinco métodos foram capazes de filtrar os dados de todas as áreas, apresentando diferentes porcentagens de remoção de dados. O método 1 e o método 4 foram o de menor desempenho na detecção de dados discrepantes artificias, enquanto o método 2, método 3 e método 5 foram o de maior desempenho nesta ordem.

A contaminação de conjunto de dados com dados artificias é uma abordagem promissora de identificação de dados discrepantes, entretanto ainda não é possível afirmar com convicção que quanto maior o desempenho do método em filtrar estes dados, mais confiáveis são os dados resultantes da filtragem. Logo, são necessários mais estudos que comprovem se todos os dados filtrados por diversos métodos são de fato discrepantes com a realidade.

AGRADECIMENTO: À empresa Solinftec pelo suporte na obtenção dos dados.

REFERÊNCIAS

BECK, H.; KÜHN, M. Dynamic Data Filtering of Long-Range Doppler LiDAR Wind Speed Measurements. **Remote Sensing** 2017, Vol. 9, Page 561, v. 9, n. 6, p. 561, 4 jun. 2017.

CASTELEYN, S.; OMETOV, A.; TORRES-SOSPEDRA, J.; SHEHU YARO, A.; MALY, F.; PRAZAK, P. Outlier Detection in Time-Series Receive Signal Strength Observation Using Z-Score Method with Sn Scale Estimator for Indoor Localization. **Applied Sciences** 2023, Vol. 13, Page 3900, v. 13, n. 6, p. 3900, 19 mar. 2023.

DASH, C. S. K.; BEHERA, A. K.; DEHURI, S.; GHOSH, A. An outliers detection and elimination framework in classification task of data mining. **Decision Analytics Journal**, v. 6, p. 100164, 1 mar. 2023.

HIGGINS, S.; SCHELLBERG, J.; BAILEY, J. S. Improving productivity and increasing the efficiency of soil nutrient management on grassland farms in the UK and Ireland using precision agriculture technology. **European Journal of Agronomy**, v. 106, p. 67-74, 1 maio 2019.

HODGE, V. J.; AUSTIN, J. A Survey of Outlier Detection Methodologies. **Artificial Intelligence Review 2004** 22:2, v. 22, n. 2, p. 85-126, out. 2004. Disponível em:

JEONG, J.; PARK, E.; HAN, W. S.; KIM, K.; CHOUNG, S.; CHUNG, I. M. Identifying outliers of non-Gaussian groundwater state data based on ensemble estimation for long-term trends. **Journal of Hydrology**, v. 548, p. 135-144, 1 maio 2017.

KLUYVER, T.; RAGAN-KELLEY, B.; PÉREZ, F.; GRANGER, B.; BUSSONNIER, M.; FREDERIC, J.; KELLEY, K.; HAMRICK, J.; GROUT, J.; CORLAY, S.; IVANOV, P.; AVILA, D.; ABDALLA, S.; WILLING, C. Jupyter Notebooks - a publishing format for reproducible computational workflows. **Positioning and Power in Academic Publishing: Players, Agents and Agendas - Proceedings of the 20th International Conference on Electronic Publishing, ELPUB 2016**, p. 87–90, 2016.

LIN, L.; SHERMAN, P. D. Cleaning data the Chauvenet way. **Group - The Proceedings of the SouthEast SAS Users**, 2007.

LIU, J.; LIU, L.; DU, J.; SANG, J. TLE outlier detection based on expectation maximization algorithm. **Advances in Space Research**, v. 68, n. 7, p. 2695-2712, 1 out. 2021.

LUND, E.D.; MAXTON, C.R.; LUND, T. J. A Data Fusion Method for Yield and Soil Sensor Maps. **13th International Conference on Precision Agriculture**, St. Louis, Missouri, EUA, 2016.

LYLE, G.; BRYAN, B. A.; OSTENDORF, B. Post-processing methods to eliminate erroneous grain yield measurements: Review and directions for future development. **Precision Agriculture**, v. 15, n. 4, p. 377- 402, 1 nov. 2014.

MALDANER, L. F.; MOLIN, J. P.; SPEKKEN, M. Methodology to filter out outliers in high spatial density data to improve maps reliability. **Scientia Agricola**, v. 79, n. 1, p. e20200178, 18 jan. 2021..

MALDANER, L. F.; CANATA, T. F.; MOLIN, J. P. An Approach to Sugarcane Yield Estimation Using Sensors in the Harvester and ZigBee Technology. **Sugar Tech**, v. 24, n. 3, p. 813-821, 1 jun. 2022.

MONTEIRO, A.; SANTOS, S.; GONÇALVES, P. Precision Agriculture for Crop and Livestock Farming Brief Review. **Animals 2021**, Vol. 11, Page 2345, v. 11, n. 8, p. 2345, 9 ago. 2021.

R Core Team (2023). _R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.

SMITI, A. A critical overview of outlier detection methods. **Computer Science Review**, v. 38, p. 100306, 1 nov. 2020.