Evaluation of Filtering Approaches and Spatial Relationships with Gaps in Sugarcane Yield Maps

Yuri de Lacerda Barbosa[1] ; Vítor Almeida Ferraz[1*] ; Eudocio Rafael Otavio da Silva[1] ; José Paulo Molin [1]

[1] Laboratory of Precision Agriculture, "Luiz de Queiroz" College of Agriculture – University of São Paulo, SP, Brazil

A paper from the Proceedings of the
17th International Conference on Precision Agriculture and 11th Brazilian Congress on Precision and Digital Agriculture
July 13–15, 2026
Porto Alegre, Brazil

ABSTRACT

Yield maps are indispensable tools in Precision Agriculture (PA) for the quantitative and qualitative characterization of agricultural fields. However, data derived from yield monitors do not always reflect actual field productivity, as they are subject to measurement errors, operational errors, and/or equipment malfunction. These distortions compromise analytical accuracy and, consequently, decision-making. Given this challenge, the present study aimed to evaluate the performance of filtering approaches applied to sugarcane yield data, as well as to investigate the spatial relationships between harvester records and the occurrence of sugarcane row gaps. To achieve these objectives, yield data were obtained from a sugarcane field in the Center-South region of Brazil using harvesters equipped with yield monitors. Anomalous observations were artificially inserted into this dataset at six distinct magnitudes. Subsequently, nine filtering approaches (Z-Score, MapFilter 2.0, Standard Deviation, Interquartile Range, Chauvenet's Criterion, Kernel, Mahalanobis, DBSCAN, and K-means) were applied under two different scenarios: to the artificially contaminated dataset to evaluate the detection rate of synthetic outliers, and to the original baseline data. Additionally, a vector layer of sugarcane row gaps was spatially overlaid onto the original and filtered maps to identify inconsistencies, classifying harvest records located over these gaps (areas without plants) as noise. The results indicated that filtering using MapFilter 2.0 and Chauvenet's Criterion exhibited higher performance in identifying and removing artificial outliers, particularly for discrepancies of greater magnitude. On the other hand, the Kernel, Mahalanobis, and DBSCAN approaches showed inferior performance when compared with the aforementioned methods. Furthermore, a strong correlation was observed between the performance hierarchy of the filtering approaches and their respective percentages of record removal over the gap layer, attesting to the robustness of the findings. It is concluded that data filtering based on robust statistics and local spatial dependence is superior for processing this type of data. Although these approaches require prior parameterization of density variables and the spatial range of the crop, they possess a greater ability to capture and preserve intrinsic agronomic variability, overcoming the limitations of filters based purely on conventional statistics. Finally, the overlay with sugarcane row gaps highlights the need to develop further studies exploring filters sensitive to the anisotropic spatial structure characteristic of sugarcane along its rows.

Keywords

Yield sensors, MapFilter, Yield map, Outlier, Spatial variability.

INTRODUCTION

Yield maps are an important Precision Agriculture (PA) tool that reflect crop responses to management practices and enable production characterization (Lund et al., 2016). Such maps also allow the identification of spatial variations and the investigation of the factors affecting them (Higgins et al., 2019), supporting decision-making and the implementation of site-specific management strategies (Monteiro et al., 2021). The quality of this information is important because the agronomic intelligence extracted from yield maps guides complex interventions, going far beyond simple operational harvest tracking.

The implementation of automated systems results in the collection of a vast amount of data. However, outliers frequently differ from the observations within the dataset and may be attributed to measurement errors, harvester feeding dynamics, headland maneuvers, and calibration issues (Lyle et al., 2014; Maldaner et al., 2021; Silva et al., 2024). There are also true outliers, which indicate real spatial phenomena not previously identified (Smiti, 2020). Faced with the challenge of the absence of an absolute reference defining what constitutes a discrepant observation within a dataset, the insertion of artificial discrepant data has become a methodology adopted in some studies to assess the performance of filtering algorithms, since the artificially inserted data are known (Liu et al., 2021; Silva et al., 2024).

Sugarcane (Saccharum spp.) exhibits high spatial variability both between and within rows, and due to the specific challenges involved in generating yield maps, the filtering process for these data is difficult, as it requires an approach capable of removing outliers while preserving spatial variability within the sugarcane field (Usaborisut, 2018; Mutran et al., 2019; Maldaner and Molin, 2019). At the same time, row-level monitoring allows for the assessment of how gaps affect yield (Stolf, 1989; Maldaner et al., 2024). In this crop, a gap is defined as any segment of the planting row where the observed plant density is lower than expected, indicating the absence of plants in locations defined by the crop's spatial arrangement (da Silva, 2025). These gaps can be mapped using different approaches and are currently identified by sugarcane mills using unmanned aerial vehicles (UAVs). In this context, exploring the use of multiple data layers acquired at the row level in sugarcane management—such as in data filtering processes—has the potential to generate reliable maps that accurately represent field conditions.

Therefore, the present study aimed to evaluate the performance of nine different filtering approaches for yield data in sugarcane fields using artificial discrepancy data. The specific objectives included comparing the statistical and spatial performance of the algorithms and investigating the spatial relationships between yield layers obtained from harvesters and sugarcane row gaps extracted from UAV imagery.

MATERIALS AND METHODS

Study area and data collection

The dataset was obtained from the 2021 harvest of four fields (23.1, 23.8, 30.1, and 12.9 ha) cultivated with the sugarcane variety RB966928, following the processing and analysis steps illustrated in Figure 1. The study area is located in the municipality of Mirassolândia (20°35'29.4" S, 49°26'56.2" W), São Paulo State, Brazil. The single-row harvesters were equipped with a Global Navigation Satellite System (GNSS) and a commercial yield sensing system based on the feed pressure of the chopper drum (Solinftec, Araçatuba, Brazil), operating at an average data collection frequency of 0.20 Hz.

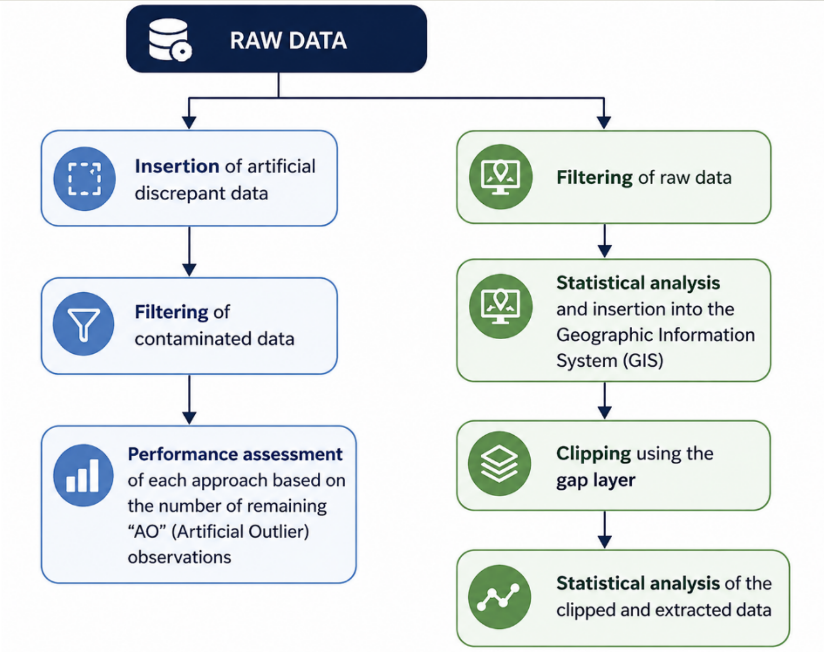

Fig. 1. Flowchart of the main steps of the adopted methodology. AO: Artificial outliers

Filtering approaches

Nine filtering approaches were selected and tested on the yield maps.

- *Approach 1: Z-score.*
  Z-score filtering identified outliers based on the number of standard deviations by which each observation differed from the dataset mean (Hodge and Austin, 2004).
- *Approach 2: MapFilter 2.0.*
  Proposed by Maldaner et al. (2021), the algorithm performed both local and global spatial analyses. The global analysis removed data outside the upper and lower limits, calculated based on the median (Equations 1 and 2):
  $$Upper\ limit = median + (Variation\ threshold * median) \quad (1)$$
  $$Lower\ limit = median + (Variation\ threshold * median) \quad (2)$$
  A value of 30% was adopted for the parametric variation threshold, and a spatial dependency radius of 200 m was used for the local isotropic analysis.

- *Approach 3: Standard Deviation.*
  A normal distribution was assumed, and outliers were removed based on multiple thresholds calculated according to Equation 3 (Beck and Kühn, 2017):
  $$s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n}} \quad (3)$$
  where x_i is the observed value, $\bar{x}$ is the mean, and n is the number of observations.

- *Approach 4: Interquartile Range.*
  The measure of data dispersion based on quartiles was used (Montgomery and Runger, 2010), defining the limits through the interquartile range (iqr), calculated according to Equation 4:
  $$iqr = q3 - q1 \quad (4)$$
  The upper and lower acceptance thresholds were defined by Equations 5 and 6:
  $$Upper\ limit = q3 + 1{,}5 * iqr \quad (5)$$
  $$Lower\ limit = q1 - 1{,}5 * iqr \quad (6)$$

- *Approach 5: Chauvenet's Criterion.*
  Filtering was based on the standard deviation, mean, and sample size (Lin and Sherman, 2007). Observations were rejected according to the criterion established in Equation 7:
  $$If\ \frac{|x_i - \mu|}{\sigma} > R, reject\ x_i \quad (7)$$

where R is the tabulated critical value of Chauvenet's criterion, x_i is the analyzed value, μ is the mean, and σ is the sample standard deviation.

- *Approach 6: Kernel Density Estimation.*
  The method was based on spatial probability density to identify isolated data in low-density regions (Silverman, 1986), calculated according to Equation 8:
  $$f(x) - \frac{1}{nh^d}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{8}$$
  where $f(x)$ is the estimated density, h is the smoothing bandwidth, d is the data dimension, and K is the kernel function.

- *Approach 7: Mahalanobis Distance*.
  The distance of each georeferenced point to the centroid of the multivariate data was measured (Mahalanobis, 1936), as expressed in Equation 9:
  $$D^m = \sqrt{(x - \mu)^{\mathrm{T}} S^{-1} (x - \mu)} \tag{9}$$
  Where x is the feature vector, μ is the mean vector, S^{-1} is the inverse covariance matrix, and T denotes the mathematical transpose.

- *Approach 8: DBSCAN.*
  The density-based clustering algorithm (Ester et al. 1996) was used, with a neighborhood radius parameterization (ϵ) and a minimum number of neighboring points $(minPts)$ set to 5. The methodological classification was governed by Equation 10:
  $$Cluster(p) = \begin{cases} Core & se\ |\mathrm{N}_\epsilon(p)| \geq minPts \\ Border & se\ p \in \mathrm{N}_\epsilon(q) para\ core\ point\ q \\ Noise & caso\ contrário \end{cases} \tag{10}$$

- Approach 9: K-means.
  The partitioning algorithm (Lloyd 1982) was employed, focused on minimizing the sum of the squared distances to the specified centroids (Equation 11):
  $$D(C) = \sum_{i=1}^{K} \sum_{x \epsilon c_i} \|x - \mu_i\|^2 \tag{11}$$
  The parameter corresponding to the number of clusters (K) was set to 2.

Contamination with artificial outliers
The original dataset was deliberately contaminated with artificial outliers to test the sensitivity of the algorithms (Liu et al. 2021). Approximately 3% of the sugarcane harvest dataset was randomly selected, and the formulation in Equation 12 was applied:
$$AO = Z + (OM * Z) \tag{12}$$
where AO is the inserted artificial outlier value, Z is the original yield value of the observation, and OM represents the magnitude of outliers, with OM = (±0.01, ±0.05, ±0.10, ±0.50, ±0.80, and ±1.00), representing outliers of small (±0.01, ±0.05, and ±0.10), medium (±0.50), and large (±0.80 and ±1.00) magnitude. The analysis was performed using the R programming environment v. 4.5.0 (R Core Team, Vienna, Austria).

Spatialization and intersection with the fault layer
The fault indicator layer in the sugarcane field was obtained through photogrammetry using an eBee UAV (senseFly, Cheseaux-sur-Lausanne, Switzerland) equipped with an S.O.D.A RGB sensor, with a focal length of 25.40 mm and a spatial resolution of 20 megapixels, resulting in a spatial resolution of 30 mm. The aerial survey was conducted at an altitude of 100 m with georeferenced control based on a Topcon GR-3 receiver (Topcon Corporation, Tokyo, Japan) operating in an RTK network, with an accuracy of ±30 mm. Orthorectification and mosaicking were performed using PhotoScan v. 1.5 software (Agisoft LLC, St. Petersburg, Russia), following procedures described in detail in Maldaner et al. (2025).

Linear spacings greater than 0.58 m were identified as stand gaps. The raw yield data were imported into the QGIS 3.28.15 Geographic Information System (Open Source Geospatial Foundation, Chicago, USA) and spatially intersected with the gap feature plus a rectangular buffer zone of 0.75 m. Harvester records positioned exactly within these plant-free areas were identified, classified, and removed as operational anomalies.

RESULTS AND DISCUSSION

Descriptive statistics and spatial analysis

The raw yield data exhibited a wide range and high coefficients of variation (CV), with minimum values equal to zero and maximum values exceeding 600.00 Mg ha^{-1}, as detailed in Figure 2. The spatial distribution of these raw records for the four experimental areas is presented in Figure 3, highlighting the heterogeneity and the presence of operational noise prior to processing.

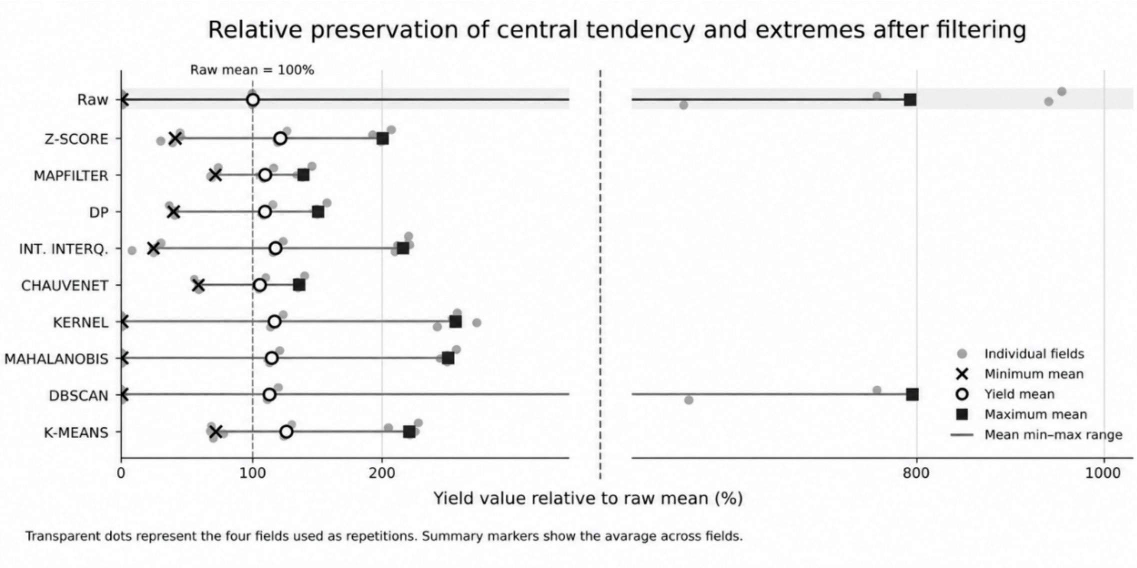

Fig. 2 Descriptive statistics for the raw and filtered data from each area

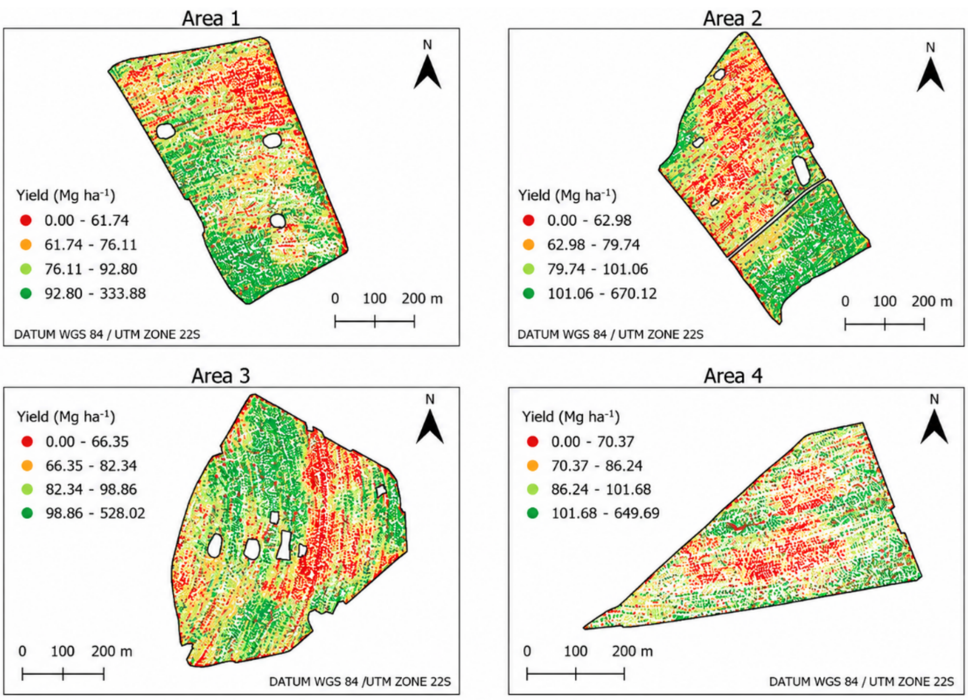

Fig. 3 Spatial distribution of raw sugarcane yield data

After applying the filtering approaches, a general reduction in data dispersion was observed across all plots. The spatial representation of the yield data after being processed by the nine approaches is illustrated in Figure 4. It was noted that Approach 2 (MapFilter 2.0) showed the highest percentage of removal of inconsistent data, which resulted in the lowest values of standard deviation and CV, followed by Approach 5 (Chauvenet Criterion). These results indicate that methods that consider local spatial variability or robust statistics of the observational data tend to be more restrictive and effective in eliminating data that deviate from the agronomic reality of the crop.

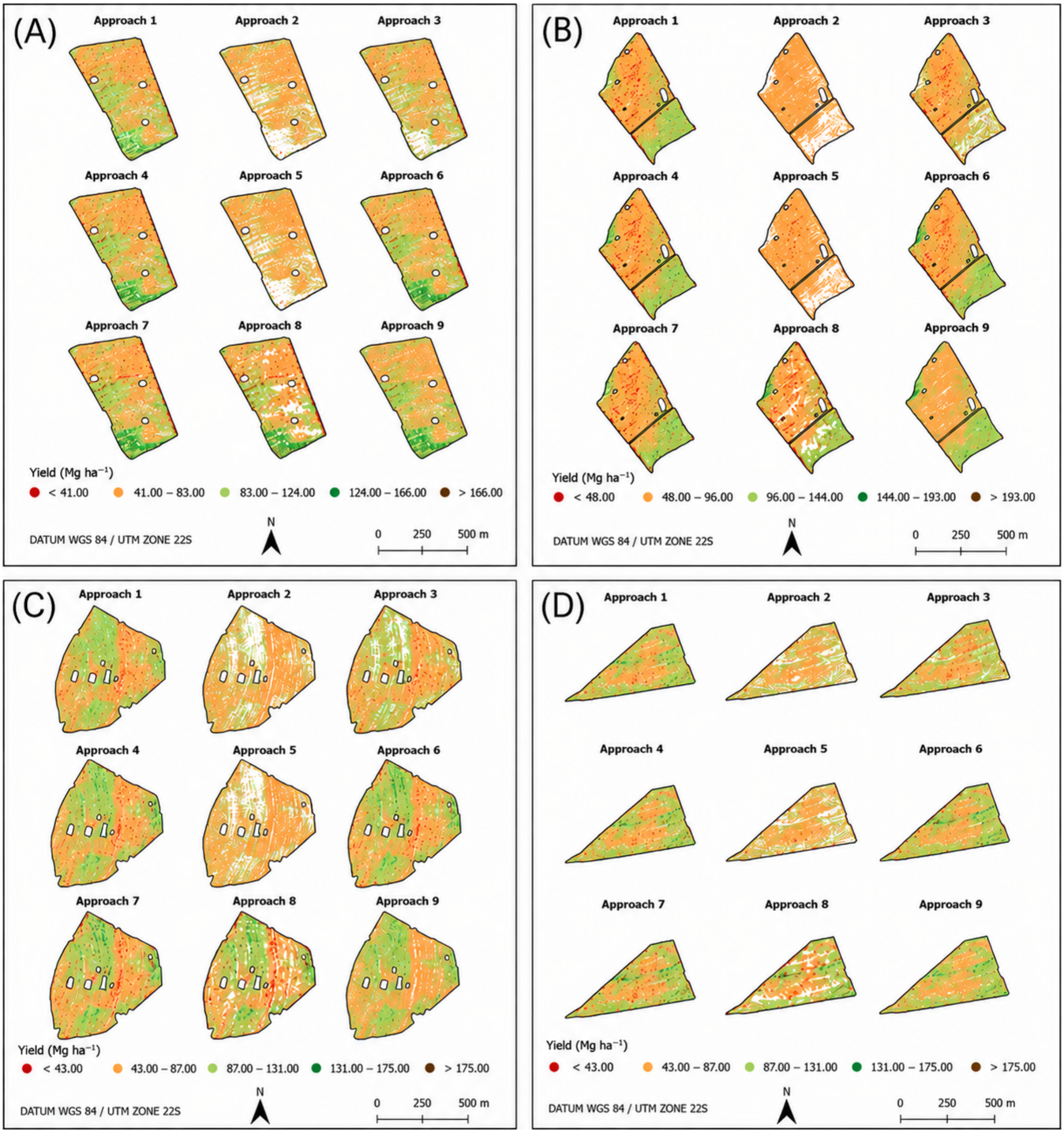

Fig. 4 Spatialization of yield data after filtering using the nine approaches of the study

Performance of the approaches in detecting artificial outliers

The performance of each approach in identifying artificial outliers inserted into the database is presented in Figure 5. It was found in all filtering approaches that performance was directly influenced by the magnitude of the inserted error (OM); the greater the magnitude, the higher the detection of synthetic outliers; with the exception of Approach 8 (DBSCAN), which was unable to detect anomalies in all tested areas and magnitudes.

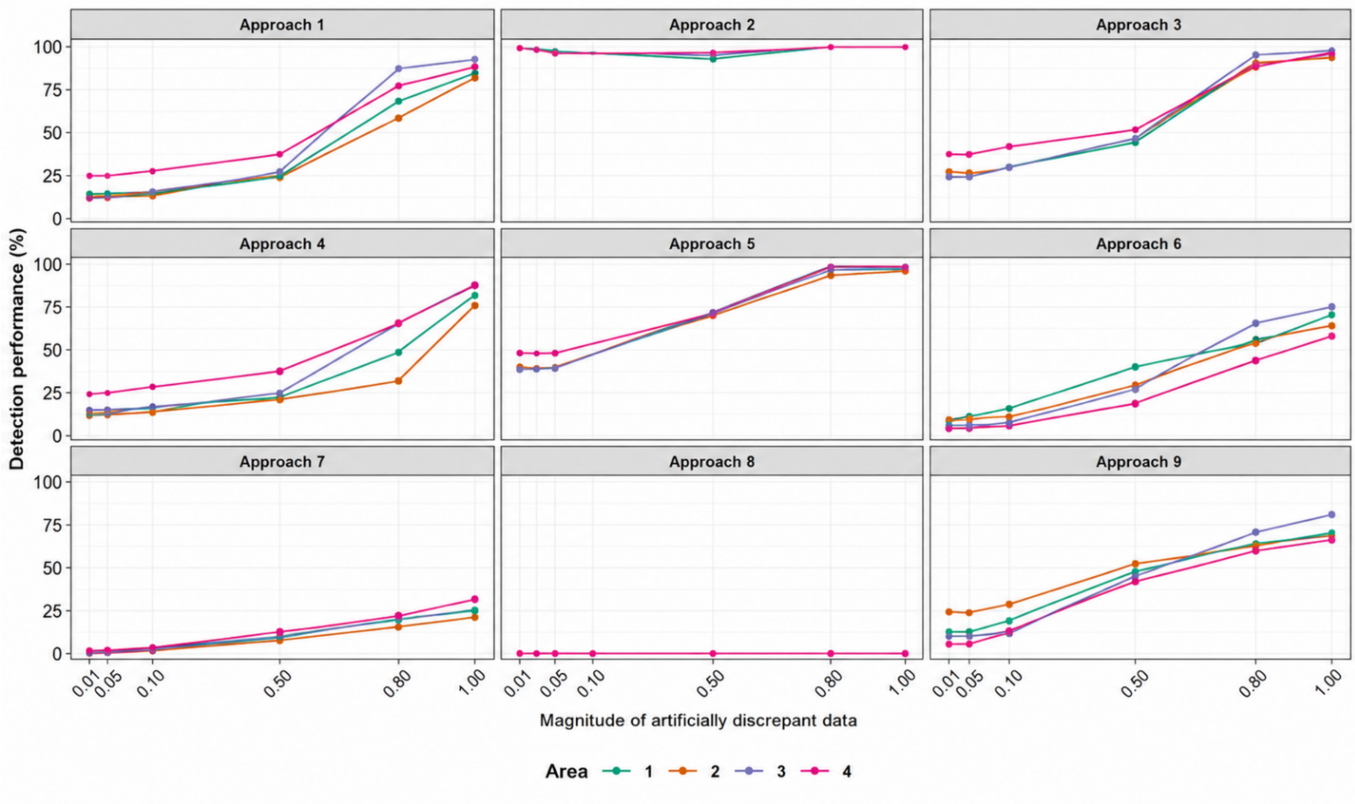

Fig. 5 Detection performance of artificial outliers by the proposed filtering approaches

Approaches 6 (Kernel) and 7 (Mahalanobis) demonstrated low sensitivity in detecting artificial outliers, with performance below 2% for low magnitudes of AO. In contrast, Approach 2 showed superior performance, achieving 100% detection for high-magnitude AO and maintaining high performance (93% to 98%) for medium- and low-magnitude AO. The superior performance in AO detection compared to other filtering approaches can be attributed to the global and local analyses performed in MapFilter 2.0, which, in addition to filtering data based on yield values across the entire field (global filter), includes a second step that filters data in a single direction (anisotropic filter) and also includes filtering based on neighbors within a specified radius (Maldaner et al., 2021), incorporating the values of neighboring rows into the calculation of the upper and lower limits (isotropic filter). This allows for capturing yield variability more accurately than the other tested approaches.

Spatial relationships with the row failure layer

The maps of row failures in the sugarcane field are presented in Figure 6. The average length of the failures ranged from 1.36 m to 1.63 m in the studied areas, with Area 3 concentrating the largest volume of raw data overlapping failure zones.

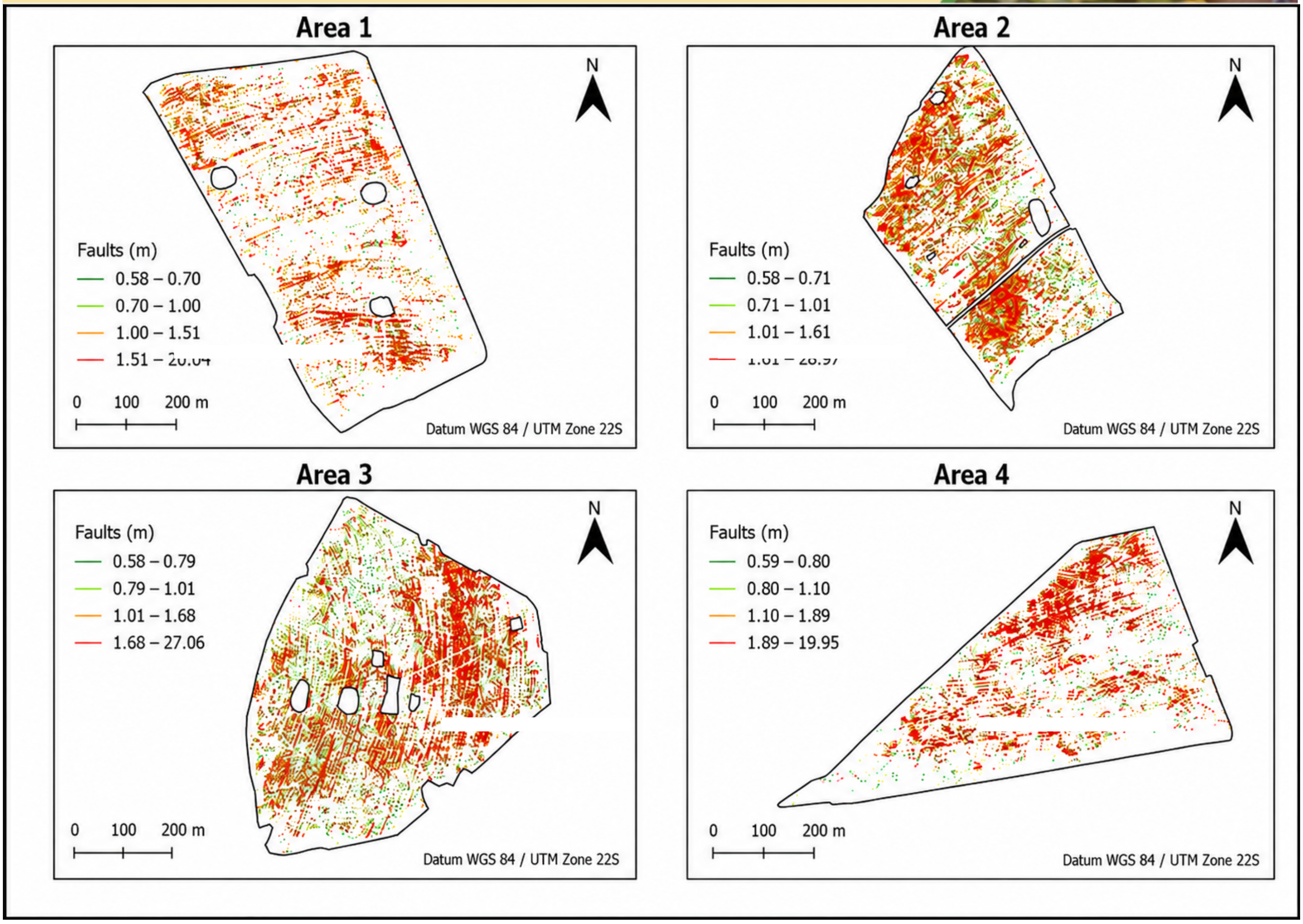

Fig. 6 Maps of row-level faults in sugarcane extracted from UAV imagery

The descriptive statistics of the yield data after intersection with the gap layer are presented in Figure 7. A direct correlation was observed between the detection and removal of outliers located in areas without plants, with gaps. Figure 8 details the records that remained within the gap zones after filtering. Approach 2 was the most effective in cleaning these critical areas, which reinforces the importance of integrating external spatial information (such as UAV-derived fault maps) to enhance the reliability of productivity maps and support more precise site-specific management strategies.

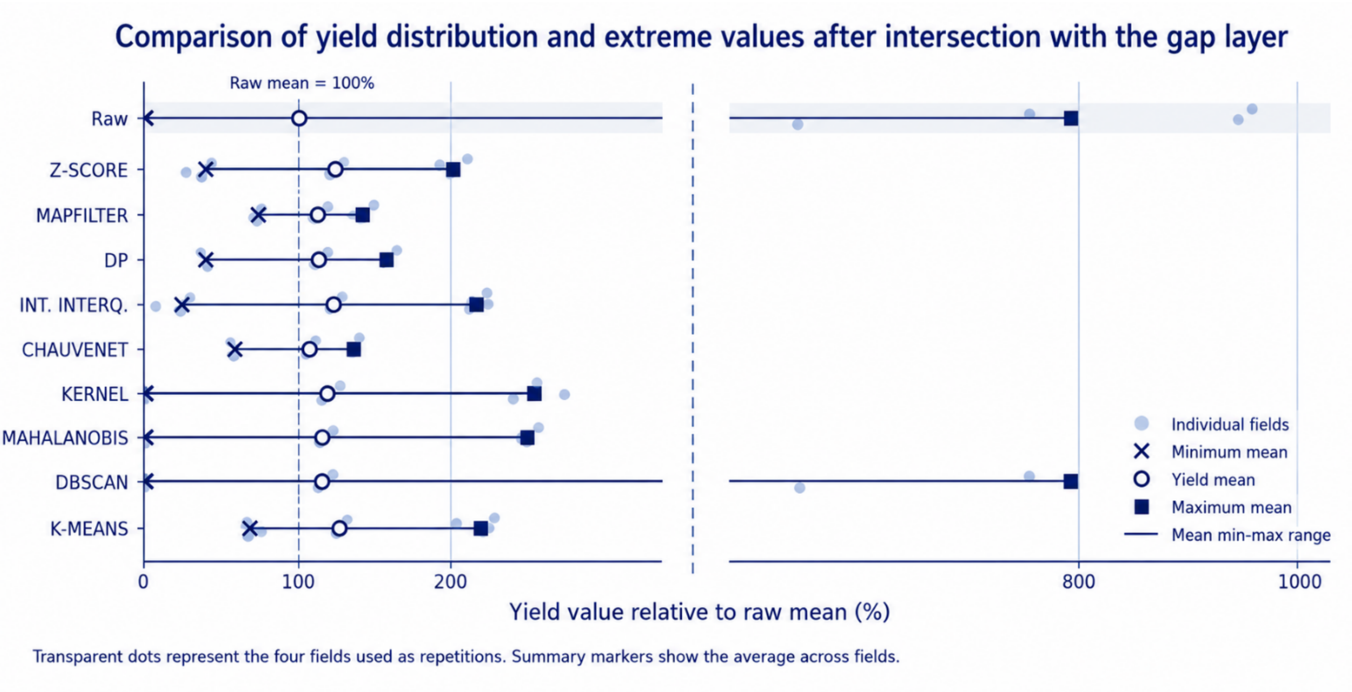

Fig. 7 Descriptive statistics for raw and filtered data cropped by the fault layer

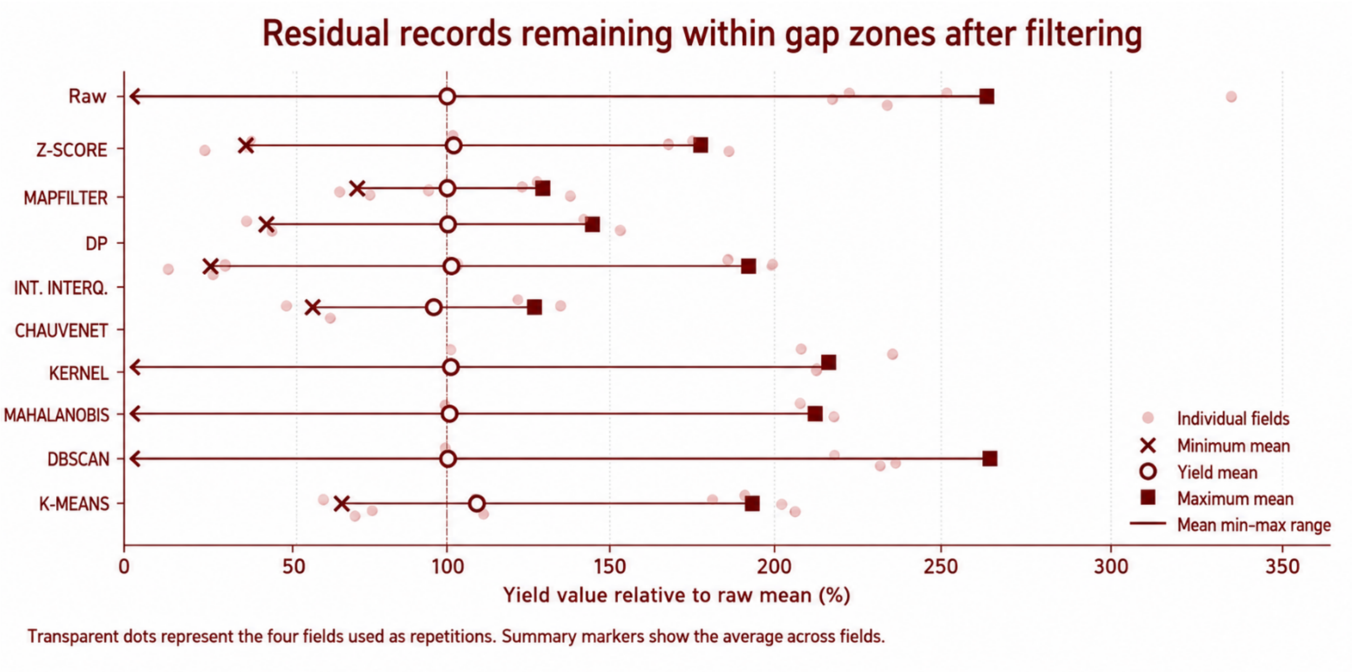

Fig. 8 Descriptive statistics for raw and filtered data extracted at the intersection with faults

CONCLUSIONS

All of the approaches evaluated demonstrated the technical capability to filter raw sugarcane productivity data, significantly reducing the total volume of records. It was found, however, that performance in detecting anomalies varied greatly depending on the mathematical premise of each algorithm. Methodologies based on a combination of local and global spatial analysis (MapFilter

2.0) and robust probability statistics (Chauvenet Criterion) demonstrated superiority in the technical cleaning of yield maps, as evidenced by their superior performance in identifying artificial outliers of low to high magnitudes.

Overlaying harvester data with the row-level error layer proved to be an indispensable spatial validation strategy. The persistence of yield records in areas proven to be devoid of plants, even after applying purely numerical filters (such as Z-score or Standard Deviation), highlights the limitation of addressing the error solely from a statistical perspective. It is concluded that the creation of yield maps that reflect the true agronomic intelligence of the crop requires the use of filtering methods sensitive to the structural variability and anisotropic spatial distribution inherent to sugarcane cultivation.

REFERENCES

Beck, H., & Kühn, M. (2017). Dynamic data filtering of long-range Doppler LiDAR wind speed measurements. Remote Sensing, 9(6), 561.

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (pp. 226–231). Menlo Park, USA: AAAI Press.

Higgins, S., Schellberg, J., & Bailey, J. S. (2019). 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, 106, 67-74.

Hodge, V. J., & Austin, J. (2004). A survey of outlier detection methodologies. Artificial Intelligence Review, 22(2), 85–126.

Lin, L., & Sherman, P. D. (2007). Cleaning data the Chauvenet way. In Proceedings of the Southeast SAS Users Group Conference. Cary, USA: SAS Institute.

Liu, J., Liu, L., Du, J., & Sang, J. (2021). TLE outlier detection based on expectation maximization algorithm. Advances in Space Research, 68(7), 2695-2712.

Lloyd, S. P. (1982). Least squares quantization in PCM. IEEE Transactions on Information Theory, 28(2), 129-137.

Lund, E. D., Maxton, C. R., & Lund, T. J. (2016). A data fusion method for yield and soil sensor maps. In Proceedings of the 13th International Conference on Precision Agriculture. Monticello, USA: International Society of Precision Agriculture.

Lyle, G., Bryan, B. A., & Ostendorf, B. (2014). Post-processing methods to eliminate erroneous grain yield measurements: Review and directions for future development. Precision Agriculture, 15(4), 377-402.

Mahalanobis, P. C. (1936). On the generalized distance in statistics. Proceedings of the National Institute of Sciences of India, 2(1), 49-55.

Maldaner, L. F., & Molin, J. P. (2019). Data processing within rows for sugarcane yield mapping. Scientia Agricola, 76, e20180391.

Maldaner, L. F., Molin, J. P., & Spekken, M. (2021). Methodology to filter out outliers in high spatial density data to improve map reliability. Scientia Agricola, 79(1), e20200178.

Maldaner, L. F., Molin, J. P., & Silva, E. R. O. (2024). Spatial-temporal analysis to investigate the influence of in-row plant spacing on sugarcane yield. Sugar Tech, 26, 194–206.

Maldaner, L. F., Molin, J. P., Dias, C. T. S., & Silva, E. R. O. (2025). When ratoon longevity matters: Spatial and temporal analysis of sugarcane plant population patterns using features extracted from UAV images. Remote Sensing Applications: Society and Environment, 101703.

Molin, J. P., & Veiga, J. P. S. (2016). Spatial variability of sugarcane row gaps: measurement and mapping. Science and Agrotechnology, 40(1), 89-96.

Monteiro, A., Santos, S., & Gonçalves, P. (2021). Precision agriculture for crop and livestock farming—brief review. Animals, 11(8), 2345.

Montgomery, D. C., & Runger, G. C. (2010). Applied statistics and probability for engineers. Hoboken, USA: John Wiley & Sons.

Mutran, V. M., Ribeiro, C. O., Nascimento, C. O. A., & Chachuat, B. (2019). Risk-conscious approach to optimizing bioenergy investments in the Brazilian sugarcane industry. In Computer Aided Chemical Engineering (pp. 361-366). Amsterdam, Netherlands: Elsevier.

Silva, E. R. O., Molin, J. P., Wei, M. C. F., & Canal Filho, R. (2024). Automatic filtering of sugarcane yield data. AgriEngineering, 6(4), 4812.

Silverman, B. W. (1986). Density estimation for statistics and data analysis. London, UK: Chapman & Hall.

Smiti, A. (2020). A critical overview of outlier detection methods. Computer Science Review, 38, 100306.

Stolf, R. (1989). An explanatory model of competition between sugarcane stalks and the valley of death. STAB, 8(2), 27-34.

Usaborisut, P. (2018). Progress in mechanization of sugarcane farms in Thailand. Sugar Tech, 20, 116-121.