The International Society of Precision Agriculture presents the
15th International Conference on
Precision Agriculture
26–29 JUNE 2022
Minneapolis Marriott City Center | Minneapolis, Minnesota USA

Optimization of Batch Processing of High-Density Anisotropic Distributed Proximal Soil Sensing Data for Precision Agriculture Purposes

Felippe H. S. Karp[1], Viacheslav Adamchuk[1], Alexei Melnitchouck[2], Pierre Dutilleul[3]

[1]Bioresource Engineering Department, McGill University, Canada; [2]Olds College, Canada; [3]Department of Plant Science, McGill University, Canada

A paper from the Proceedings of the
15th International Conference on Precision Agriculture
June 26-29, 2022
Minneapolis, Minnesota, United States

Abstract.

The amount of spatial data collected in agricultural fields has been increasing over the last decade. Advances in computer processing capacity have resulted in data analytics and artificial intelligence becoming hot topics in agriculture. Nevertheless, the proper processing of spatial data is often neglected, and the evaluation of methods that efficiently process agricultural spatial data remains limited. Yield monitor data is a good example of a well-established methodology for data processing that could be used as a guide to determine data processing strategies. However, data processing methods for proximal soil sensors (PSS) are not as well-known as for yield, even though sensors are widely used in precision agriculture and their data often applied in predictive models for soil spatial variability characterization. The main objective of this study was to identify suitable methodologies for processing PSS data and apply them to a particular dataset. It was determined that properly processing any spatial dataset required that the following steps must be taken: (1) data projection, (2) position offset correction, (3) global and (4) local filtering, and (5) interpolation. Based on a literature review, the most suitable methods for each step are listed and discussed, and frameworks are proposed. These methods were applied to 4 different types of PSS data (gamma ray spectrometry, ground-penetrating radar, galvanic contact and electromagnetic induction soil apparent electric conductivity). To evaluate the accuracy of each processing framework, a cross-validation procedure was used. Overall, the proposed batch processing framework improved the value of PSS data by highlighting spatial variability in the field that was previously masked by the presence of erroneous data. Also, when performing an analysis in the measurement's maps, discrepancies were found between the raw versus processed data, thus, emphasizing the need for properly processed PSS datasets.

Keywords. *Filtering, Interpolation, Proximal Soil Sensors, Processing Evaluation*

The authors are solely responsible for the content of this paper, which is not a refereed publication. Citation of this work should state that it is from the Proceedings of the 15th International Conference on Precision Agriculture. EXAMPLE: Last Name, A. B. & Coauthor, C. D. (2018). Title of paper. In Proceedings of the 15th International Conference on Precision Agriculture (unpaginated, online). Monticello, IL: International Society of Precision Agriculture.

Introduction

The amount of spatial data collected in agricultural fields has increased over the last decade, and with the increase in computational power, further developments in machine learning, and data analytics, great advances have been made in extracting practical information from this data (Colaço et al., 2021; Fulton and Port, 2018; Roberton et al., 2021; Willam et al., 2022). However, the importance of proper processing of spatial data is often neglected; this may be related to the lack of research focused on testing and/or developing different processing methodologies and frameworks to handle this type of data for precision agriculture practices.

Yield monitor data is a good example of a well-established processing methodology that has been discussed and studied by many researchers (Arslan and Colvin, 2002; Blackmore and Moore, 1999; Leroux et al., 2018; Lyle et al., 2014; Menegatti and Molin, 2004; Ping and Dobermann, 2005; Simbahan et al., 2004; Sudduth and Drummond, 2007; Sun et al., 2013; Vega et al., 2019). Alternatively, processing strategies for proximal soil sensors (PSS) are not as well-established as as those for yield. To our knowledge, very little research exists on this topic, even though PSS data is widely used in research and often it is used as a guide to determine management zones within agricultural fields (Adamchuk et al., 2004; Ji et al., 2018; Saifuzzaman et al., 2019; Schepers et al., 2004).

Some researchers provided general strategies that focused on spatial data processing (Shekhar et al., 2003; Singh and Lalitha, 2018; Spekken et al., 2013) but they were not tested on PSS data. Others, like Maldaner et al. (2022), have subjected PSS data to their methodology and obtained promising results, but still, the main focus of the processing is on spatial data filtering. Overall, studies involving the use of PSS often refer to yield monitor frameworks when processing soil sensing data (*e.g.,* Rodrigues et al., 2015). Thus, there is a need to establish specific processing algorithms for PSS data, and considering the large amount of information already collected, special focus should be given to methodologies that allow the joint processing of multiple files (sensors data). Therefore, the main objective of this study was to identify suitable methodologies for batch processing PSS data and apply them to a particular dataset.

Materials and Methods

The literature review that we performed on processing agricultural spatial data led to the definition of important steps for processing on-the-go sensing data. These steps are detailed below. They are presented in the order in which they should be applied. It must be noted that in listing, testing, and evaluating the available methodologies for each of the processing steps, more attention was given to methodologies using the least possible variables, computational power, and user input. Because PSS data often presents an anisotropic distribution with a higher density of data acquisition within a pass than between passes, more focus was also given to methods that consider the existence of within-row variability.

Local Projection

Most of the data acquired by PSS are exported from sensors systems with non-projected coordinates (geographic latitude and longitude). Converting this data to a projected coordinate system is often needed for the calculation of distances, which are necessary for other procedures involved in the data processing, *e.g.,* interpolation. Universal Transverse Mercator (UTM) is one of the most used projected coordinate systems. Sometimes, more localized systems, such as the Modified Transverse Mercator, are also used for specific locations.

However, a known issue with projection systems is distortion. In addition, a field might be located over two projection zones. In this case, one zone should be chosen to continue with the data processing, even though an error would be associated with this decision. To avoid and/or reduce these issues, we propose the use of a custom localized Cartesian coordinate system, described under the International Organization for Standardization 12188-1 Annex A (ISO, 2010). Equations 1 and 2 show the calculation for the conversion factors proposed in the standard above mentioned, while Equations 3 and 4 the conversation formulas

$$F_{lon} = \frac{\pi}{180°}\left(\frac{a^2}{\sqrt{a^2\, cos^2\, \varphi + b^2 sin^2\, \varphi}} + h\right)\cos\varphi \tag{1}$$

$$F_{lat} = \frac{\pi}{180°}\left(\frac{a^2 b^2}{(a^2\, cos^2\, \varphi + b^2 sin^2\, \varphi)^{\frac{3}{2}}} + h\right) \tag{2}$$

$$X = (Long - Long_{min}) * F_{lon} \tag{3}$$

$$Y = (Lat - Lat_{min}) * F_{lat} \tag{4}$$

where F_{lat}, F_{lon} are the location-specific conversion factors, in meters per degree, for latitude and longitude, respectively; φ is the latitude location for the center of the field, in degrees; *h* is the average elevation for the field above the ellipsoid, in meters; *a* is the semi-major axis of the ellipsoid, in meters; and *b* is the semi-minor axis of the ellipsoid, in meters; *X* and *Y* are the easting and northing projected coordinates, in meters; $Long_{min}$ and Lat_{min} are the geographic lowest values of longitude and latitude observed in the dataset, in degrees; *Long* and *Lat* are the geographic longitude and latitude, in degrees.

Position Offset

Projection is not the only source of error for the geographic placement of PSS data. Often during data collection, the Global Navigation Satellite System (GNSS) receiver cannot be placed right above the sensor. A standard practice is to place it at a high location free of obstructions , (*e.g.,* above the cabin of the vehicle), but still centered with the sensor. However, this procedure creates an offset between the data collected and the GNSS location recorded. Also, depending on the type of receiver used during data collection, a known pass-to-pass error can also affect the uncertainty of the geolocation of the data. Any such displacement in the data must be corrected. Even though some sensors' collection systems allow an adjustment of the offset distance between the GNSS receiver and sensor, this option is often neglected, or the adjustment is forgotten when making changes to the collection settings (*e.g.*, using another vehicle). Thus, an automatic procedure was considered and tested to calculate and apply this correction to the data.

The methodology proposed by Lee et al. (2012) was employed for this purpose. Their methodology uses phase correlation analysis to automatically calculate the necessary offset. This analysis includes two main steps: rasterization of the points and application of the phase correlation analysis. According to the authors, the pixel size used in the first step can affect the results of the analysis. Thus, they proposed a methodology that automatically determines this value. After determining the best pixel size, a range of offset values was applied to the data, the phase correlation analysis performed, and the phase correlation coefficient obtained. The offset value representing the highest phase correlation coefficient was selected to be applied to the data. To our knowledge, this methodology has not been tested on the identification of offset values for PSS, but is employed in the yield processing software Yield Editor 2.0 (Sudduth et al., 2012), to estimate the delay for yield and moisture sensors.

Other methodologies (*e.g.*, Chung et al., 2002) are available to estimate the offset mentioned above. However, based on the conclusions of Lee et al. (2012), phase correlation analysis produces results similar to those of the geostatistical method, while reducing the processing time, which is an important variable when batch processing data.

Operational, Global, and Local Filtering

Assuming that any data displacements or projection distortions are corrected or reduced, sensors' readings may still be subjected to random factors causing erroneous observed values, or outliers (*e.g.*, a high reading by an electromagnetic induction sensor due to the presence of a metal piece, temporary malfunctioning of the sensor because of field operations). Ignoring outliers in the data can lead to higher uncertainty on interpolation or predictive processes, and affect management decisions made based on the data. Many papers that focused on yield monitor data processing highlighted the issues in final thematic maps and decision-making procedures that may arise from the presence of erroneous measurements (Leroux et al., 2018; Lyle et al., 2014; Maldaner et al., 2022; Ping and Dobermann, 2005; Sudduth et al., 2012).

The literature on yield monitor data processing also emphasizes the need to perform not only a global outlier detection and removal step, but also a local analysis aimed at comparing an observation to a pre-determined neighborhood. In the conditions of application of Geostatistics, it is assumed that an observation is spatially correlated with those of its neighbors. Thus, if an observation is distinct from neighborhood values, that observation is considered a spatial outlier.

Accordingly, based on the developed frameworks for yield monitor data processing, we propose three major steps for global and local filtering of PSS data: operational (removal of operation related patterns, *e.g.*, maneuvers, changes in travel speed), global-statistical (removal of erroneous and extreme values), and local-statistical (removal of spatial outliers).

Operational Filtering

Changes in travel speed and direction (heading), stops, etc. are inevitable during field data collection. However, these operational factors become part of the data and can affect its quality. Co-located points can occur due to a stop or error on the sensor's logger and affect the data interpolation. Maneuvers can cause the tilt or loss of contact between the sensor and soil, also generating erroneous measurements.

The removal of such operational factors is important to increase the overall data quality. Four filters are proposed to remove such data points. For the operational filters involving the setting of thresholds for the identification of extreme values (maneuvers and travel speed), a non-parametric methodology was adopted to reduce the need for user inputs. This methodology uses the identification of outliers proposed by Hubert and Van Der Veeken (2008), which identifies extreme values by calculating an adjusted *outlyingness* index (AO) and cutoff, both based on a skewness-adjusted boxplot. The four filters can be detailed as follows:

1. Co-located points: any points with the same coordinates are removed.
2. Headland: a negative 5-m buffer is applied to an auto-generated field's boundary and any observation outside the buffered area is removed. A pre-generated boundary could be used in this step but was avoided because it would be another user input.
3. Maneuvers: differences in calculated heading direction between consecutive points are computed. In sequence, the adjusted boxplot, AO, and threshold are obtained, and any observation exceeding the threshold is considered a maneuver and is removed.
4. Travel speed: not every data collection system exports information on travel speed or frequency of data collection, but the latter is usually constant, allowing the use of distances between consecutive points to estimate changes in travel speed. Thus, a procedure similar to the maneuver filtering is adopted, while the Euclidean distance between consecutive points is calculated and used in the analysis for outliers. Eventually, points that are too close (lower travel speed) or too far (higher travel speed) are removed.

Global and Local Statistical Filtering

Two recently published methodologies for spatial data filtering include global and local steps and were selected to be tested. Other methodologies are available in the literature, but the two evaluated here consider spatial outliers in one direction (the data collection direction) and in all directions horizontally (Figure 1).

One methodology was proposed by Leroux et al. (2018), and can be summarized in 11 steps:

1. Global statistical filtering: based on Hubert and Van Der Veeken's (2008) AO, cutoff, and skewness-adjusted boxplot (described under Operational Filtering).
2. Median calculation for unidirectional and omnidirectional neighborhoods from each observation within two-pass widths.
3. *Outlyingness* metric calculation: $h_A = f_A - g_A$, where f_A is the observation value, and g_A, is the median of the neighborhood values for the two types of neighborhoods.

4. Repeat Steps 2 and 3 for two more neighborhood settings: three and four times the pass width.
5. Average the results of unidirectional and omnidirectional *outlyingness* metrics over the three neighborhood settings, and do a biplot (unidirectional vs. omnidirectional *outlyingness* metrics).

   The next steps no longer use spatial information (coordinates), but the distribution observed in the biplot obtained from the unidirectional vs. omnidirectional *outlyingness* metrics.
6. Construct the matrix of distances between points in the biplot, and estimate the most frequent distance (ε) using Kernel Density Estimation.
7. Determine the number of neighbors within ε for each observation, estimate their frequency by Kernel Density Estimation, and identify the first local minimum.
8. Apply a Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm (Ester et al., 1996) to the biplot, using ε as neighborhood radius and the number of local minimum found in Step 7 as the minimum number of points.
9. Extract the cluster containing normal observations (excluding outliers).
10. Refine the detection of outliers by repeating Steps 2–9 and comparing the observations flagged as outliers with their neighborhood after removing the previously flagged outliers.
11. Outliers flagged twice are considered spatial outliers and removed from the data.

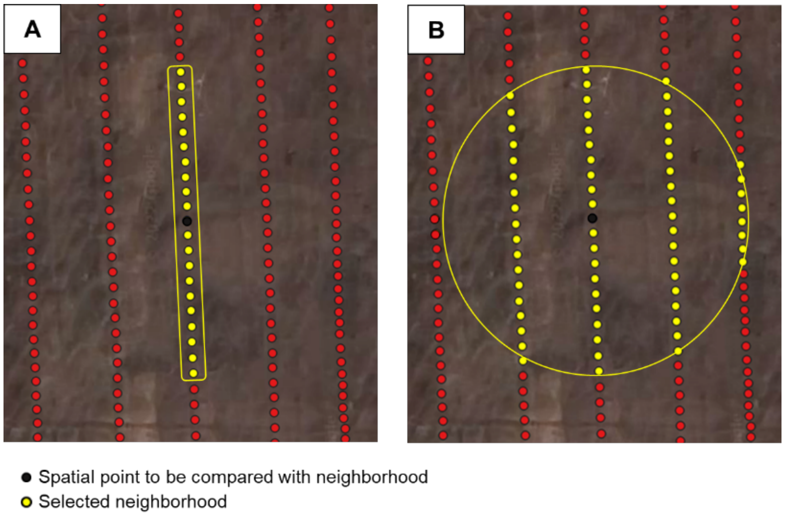

Fig. 1 Illustration of the definitions of unidirectional (within-row) neighborhood (A) and omnidirectional (within- and between-rows) neighborhood (B)

The second methodology that was reviewed in greater detail was proposed by Maldaner et al. (2022) and used a simpler approach than the first while requiring more user inputs. For the global and local filters, upper and lower bounds are defined as the median plus and minus a coefficient times the median. The coefficient can be different for the global and local filters. At the local filtering step, the median is calculated for the observations in the neighborhood, in a way similar to Step 2 for the first methodology but by using a radius equal to 2.5 times the pass width. An observation value found to be below the lower bound or above the upper bound is flagged as an outlier. This methodology removes outliers following the sequence: global, local unidirectional, and local omnidirectional.

Both methodologies use the distance between two adjacent passes (pass width) to determine the neighborhood for local filtering. This distance can easily be obtained using a Geographic Information System (GIS). However, this requires another user input. Considering that the position offset methodology of Lee et al. (2012) calculates a pixel size that corresponds to 50 to 80% of the pass width, we propose 2 times the pixel size estimated at the position offset step to be used as an estimate of the pass width.

Interpolation

Finally, in the process of using PSS data as a decision-making tool or for any other application in precision agriculture, point data must be converted to a map format. Interpolation is used for this means. Many interpolation methods are available and can be applied. The most common for agricultural data are: ordinary kriging (OK), universal kriging (UK), inverse distance weighting (IDW), and nearest neighbors (NN). After selection of the most appropriate filtering, the filtered PSS data were submitted to these four interpolation procedures.

Even though normality is not an assumption about the data distribution for geostatistical analysis, a common practice is to test for normality and if rejected, to apply a data transformation. This can help improve the results of the OK and UK procedures, as they are based, in different ways, on mean values (Isaaks and Srivastava, 1989; Schossler et al., 2019). We, therefore, recommend that for interpolation based on kriging, the data distribution be tested for normality, and when rejected, a Box-Cox transformation (Box and Cox, 1964) be used.

The IDW interpolation requires a maximum number of neighbors and a weight applied to distances (Shepard, 1968). Either parameter can be user-defined or automatically optimized. A common approach is to divide the dataset into 'test and train' subsets and evaluate different options and combinations of parameter values in relation to prediction errors. The combination of parameter values providing the smallest prediction error is selected (Barbulescu et al., 2020). This procedure can be computationally costly and time-consuming. Thus, for datasets with >1000 points (size of most PSS datasets), a threshold distance is recommended to reduce the computational cost for IDW interpolation (Hengl, 2009). We also propose the use of a Limited Memory Algorithm for Bound Constrained Optimization through an L-BFGS-B algorithm (Byrd et al., 1995), to obtain optimized parameters for IDW interpolation.

Evaluation

Data from four PSS instruments were used to evaluate the steps and different methodologies described above. Table 1 presents the different sensors, characteristics of the data collection (pass swath and distance between consecutive points), data density, and variables evaluated. The data from galvanic contact apparent electrical conductivity (GC) and ground penetrating radar (GPR) are from a 43-ha field, while the γ-ray and electromagnetic induction apparent electrical conductivity (EMI) come from an 82-ha field.

Raw data was standardized to zero mean and unit variance to facilitate the interpretation of results. The evaluation of position offset and filtering steps (global and local) took place by comparing the changes in the estimated parameters of fitted variogram models and the average root mean squared error (RMSE) from a 10-fold cross-validation using OK interpolation.

To compare interpolation methods, data distribution normality was tested first, and a Box-Cox transformation was applied as required. Transformed data was used for each interpolation method, including IDW and NN in addition to OK and UK. No Box-Cox transformation was applied before interpolation because, by definition, a Box-Cox transformation modifies the data distribution, which would affect the comparison among methodologies in their efficiency to remove global and local outliers.

For each interpolation method, a 10-fold cross-validation was performed, and the RMSE and the coefficient of determination (R^2) between predicted and observed values were reported. For any kriging procedure in this study, three models (exponential, spherical, and Gaussian) were fitted by weighted least squares to the empirical variogram. The best variogram model was selected based on the error sum of squares. Evaluation of frameworks and methodologies was performed with customized scripts and functions written in the R language (R Core Team, 2021).

Table 1. Information for the different proximal soil sensors used for the evaluation of processing methodologies

Sensor Model	Manufacturer Provider	Technique	Swath (m)	Point Distance (m)	Samples .ha^{-1}	Nb. Obs.	Variables
EM38-MK2	Geonics Limited (Mississauga, Ontario-CA)	EC_a - Electromagnetic Induction (EMI)	25	4.1	108	8852	EC_a Deep (0-1.5 m)
SIR-4000 (400 MHz Antenna)	GSSI (Nashua, New Hampshire-USA)	Ground Penetrating Radar (GPR)	34	5	63	2714	Instantaneous Amplitude (0-0.1 m)
SoilOptix	SoilOptix (Tavistock, Ontario-CA)	Passive Gamma-Ray (γ-ray)	12	12	61.5	5048	Count Rate
Veris 3100	Veris Technologies (Salinas, Kansas-USA)	EC_a - Galvanic Contact (GC)	18	3.3	147	6331	EC_a Shallow (0-0.3 m)

Results and Discussion

Table 2. Geostatistical analysis results (Nugget, Sill: variogram model parameter estimates) and root mean squared error (RMSE) from ordinary kriging cross-validation for position offset and filtering procedures for each proximal soil sensor. Data is standardized to zero mean and unit variance. Bold and red-colored numbers represent the lowest values, and dashes indicate that the procedure was not applied or no point was removed.

Process	EMI[a] – 0-1.5 m			GPR[b] – 0-0.1 m			γ-Ray – Count Rate			GC[c] – 0-0.3 m		
	Nugget	Sill	RMSE	Nugget	Sill	RMSE	Nugget	Sill	RMSE	Nugget	Sill	RMSE
Raw	0.050	0.802	0.190	0.658	0.949	0.817	0.896	0.970	0.975	0.141	1.167	0.246
Offset Applied	0.047	0.802	0.192	0.633	0.942	0.814	-	-	-	-	-	-
Operational	0.034	0.517	0.203	0.347	0.680	0.666	0.855	0.989	0.982	0.102	1.150	0.233
Global Leroux et al. 2018	-	-	-	-	-	-	0.857	0.983	0.980	-	-	-
Local Leroux et al. 2018	0.004	**0.138**	**0.092**	0.122	0.314	0.424	0.575	0.688	0.797	0.015	**0.534**	0.134
Global Maldaner et al. 2022	0.012	0.444	0.179	0.262	0.444	0.565	0.748	0.831	0.904	0.056	0.665	0.203
Local Maldaner et al. 2022	**0.000**	0.282	0.129	**0.061**	**0.226**	**0.295**	**0.225**	**0.382**	**0.538**	**0.000**	0.600	**0.102**

[a] EMI – electromagnetic induction apparent electrical conductivity, [b]GPR – Ground Penetrating Radar, [c]GC – galvanic contact apparent electrical conductivity

Position Offset

Offset values ranging from -25 to 25 m in 0.5 m steps were used for the phase correlation analysis for each PSS dataset. It was found that a negative shift of -3 m and -5 m needed to be applied to the EMI and GPR data, respectively; for GC, the highest phase correlation coefficient was obtained at 0 m, so no offset should be applied. For these three sensors, a well-determined distribution for the phase correlation coefficient was obtained with a peak at the estimated offset value. For γ-ray, high variability in the phase correlation coefficient was observed, which resulted in a lack of convergence to a specific offset. A visual analysis of the raw data from the γ-ray sensor showed a high variability at short distances, that is, a weaker spatial correlation. This can also be seen in the high values of the Nugget relative to the Sill (Nugget representing 92.3% of Sill) in

Table 2. The position offset methodology under evaluation assumes the presence of spatial correlation in the data, so the results obtained for the γ-ray sensor data could be related to its weaker spatial correlation. Dashes for GC and γ-ray in Table 2 indicate that no offset was applied to the data, and the geostatistical analysis results and RMSE values are the same as for the Raw dataset, that is, when the data is standardized to zero mean and unit variance (without Box-Cox transformation).

Due to the low computational cost of the position offset methodology, multiple interactions can be assessed, together with an estimation of the variability. When a high variability in the interactions is observed, a warning can be issued to the user, or an automatic decision based on a threshold can be taken on whether or not to use the estimated offset (Lee et al., 2012; Sudduth et al., 2012).

One may also be intrigued by the need to apply a negative offset to the data from EMI and GPR. The data from these two sensors were collected by contractors, who used proprietary software to combine the GNSS data with the sensor readings. Small inconsistencies in this combination could have caused a shift in the data, resulting in the need for a negative offset. For GPR, the negative offset could also be associated with a pass-to-pass error because the data was collected with a push-cart (>15 min between the beginning and end of a pass), and data was not collected using a Real-Time Kinematic (RTK) level receiver.

Moreover, the offsets estimated for the EMI and GPR sensors are close to their average distance between consecutive points (Table 1), so the results could be related to the data collection density. Comparing geostatistical analysis results before and after the position offset (Table 2), a slight decrease in the Nugget effect for both sensors can be observed, indicating a small increase in the capability of the model to capture variability at shorter distances, even though there is a small increase in the interpolation RMSE.

To further assess the uncertainty about the effectiveness of this methodology in determining the need for a position offset in the data, a simple complementary analysis was performed by simulating a negative shift of -7 m in the GC data (between the GNSS antenna placement and the sensor). As a result, the phase correlation analysis indicated the need for a positive offset of 7 m. Accordingly, this methodology shows potential for the identification of a need for a position offset in PSS data, while being very suitable for batch processing because of its computational efficiency and autonomy. Future work will focus on simulating more datasets with a variety of offsets and evaluating the performance of the methodology to identify and correct them.

Operational, Global, and Local Filtering

Operational

The operational filtering methodology removed about 30%, 17%, 22%, and 18% of the observations from EMI, GPR, γ-ray, and GC, respectively. In view of geostatistical analysis results in Table 2, RMSE decreased for GPR and GC, and slightly increased for EMI and γ-ray. The Nugget effect presented to be smaller for all the PSS, which shows that the fitted variogram model captures more spatial dependence from the data. The blue dots presented in the Removal Step maps in Figure 2 represent the points removed by the Operational Filter, mostly observations made at the borders of the field, in maneuvers or turns, and points too close or too distant from its consecutives were removed, which represents a good performance of the filter. Similar results can be observed in the maps for the other PSS (Figures 3-5).

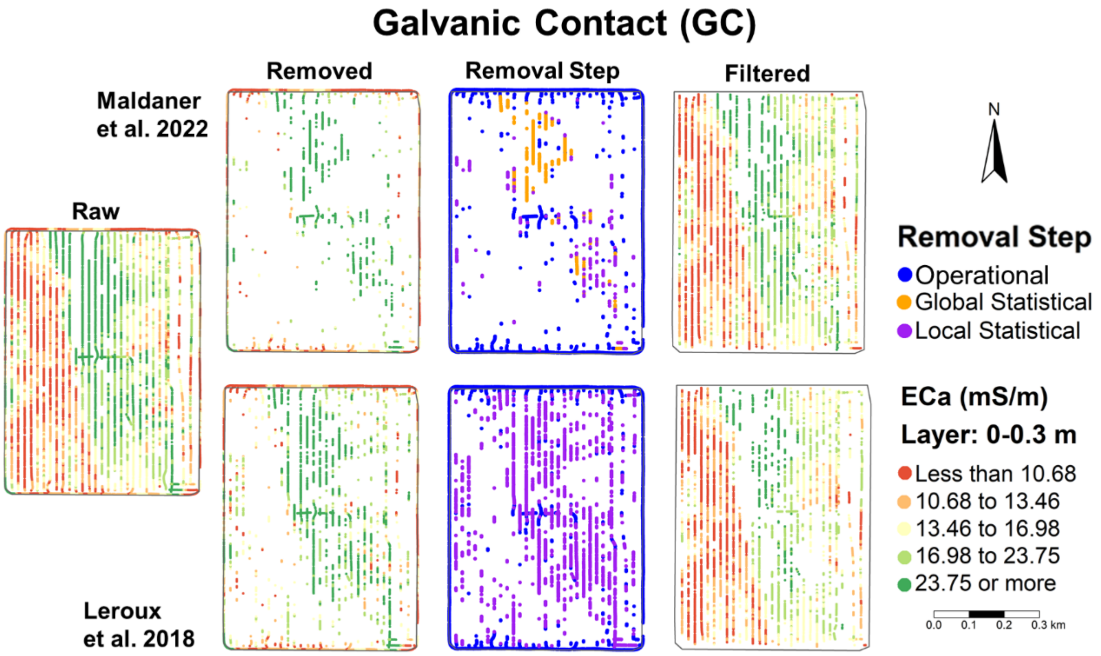

Fig. 2 Results of global and local filtering methodologies for galvanic contact (GC) sensor apparent electrical conductivity (EC$_a$). Results are presented for the methods of Maldaner et al. (2022), top, and Leroux et al. (2018), bottom. Points removed at each filtering step are classified at the Removal Step maps

Global Statistical

The methodology applied by Leroux et al. (2018) for the removal of global statistical outliers did not identify any extreme values for three of the four PSS datasets. This same methodology only removed 3 points from the γ-ray data and dashes in Table 2 indicate that no analysis was performed for this step because the results would be the same as for the Operational step. However, the methodology proposed by Maldaner et al. (2022) removed about 0.4%, 2%, 2.3%, and 4% of the observations from EMI, GPR, γ-ray, and GC, respectively, in addition to the percentages for Operational filter. The removed points in this step for the GC sensor can be observed in Figure 2 under Global Statistical (orange dots) in the Removal Step maps. Since no observations were removed by this step when using Leroux et al.'s (2018) methodology, no points for this class can be observed on the map in this case.

Maldaner et al.'s (2022) methodology requires a user-defined coefficient for this filtering step. Considering that our study focuses on the batch processing of PSS data, we tested multiple values for this coefficient on each of the sensors. We found that the filtering strategy used by this method is sensitive to the value of the coefficient, which tends to be proportional to the variability in the data. To obtain the results above, the coefficients used for EMI, GPR, γ-ray, and GC were 0.2, 1.6, 0.2, and 1.6, respectively. Spekken et al. (2013), who had formerly proposed a methodology very similar to Maldaner et al. (2022), also reported such behavior when selecting the threshold to determine the presence of spatial outliers. Despite this disadvantage, there was an overall reduction of the data variance (Sill), Nugget effect, and interpolation RMSE for all four PSS, indicating an increase in the data quality.

Local Statistical

The Local Statistical filter is the last step towards the improvement of the data quality. At this step, it is possible to fully assess the differences between the filtering methodologies under evaluation. Table 2 shows that the lowest values for Nugget, Sill, and RMSE are for Maldaner et al.'s (2022) methodology for GPR and γ-ray, while Leroux et al.'s (2018) methodology provides the lowest Sill for GC and the lowest Sill and RMSE for EMI. Through the local filtering, Leroux et al. (2018) and Maldaner et al. (2022) methodologies respectively removed an additional 15%, 7.7%, 6.7%, and

25% of the observations, and 3%, 21%, 28%, and 3% of the observations, for EMI, GPR, γ-ray, and GC.

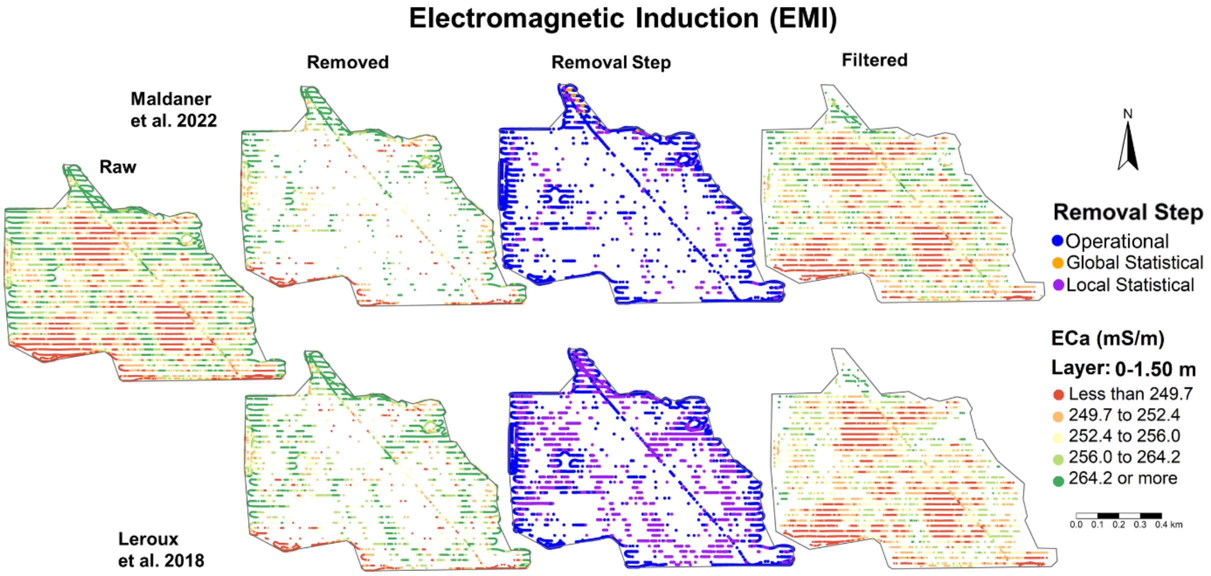

Fig. 3 Results of global and local filtering methodologies for electromagnetic induction (EMI) sensor apparent electrical conductivity (EC_a). Results are presented for the methods of Maldaner et al. (2022), top, and Leroux et al. (2018), bottom. Points removed at each filtering step are classified at the Removal Step maps

Although the geostatistical analysis results support that the methodology of Leroux et al. (2018) performs better filtering for two of the four sensors (GC and EMI), a visual inspection of the removed and filtered maps (Figures 2 and A1 – maps for Leroux et al., 2018) suggests possible excessive removal of observations during the Local Statistical step. By comparing the Raw and Removed data points, the removal of some true patterns in the field can be observed when the Leroux et al. (2018) methodology was used (*e.g.*, excessive removal of high EC_a values for the EMI sensor in Figure 3).

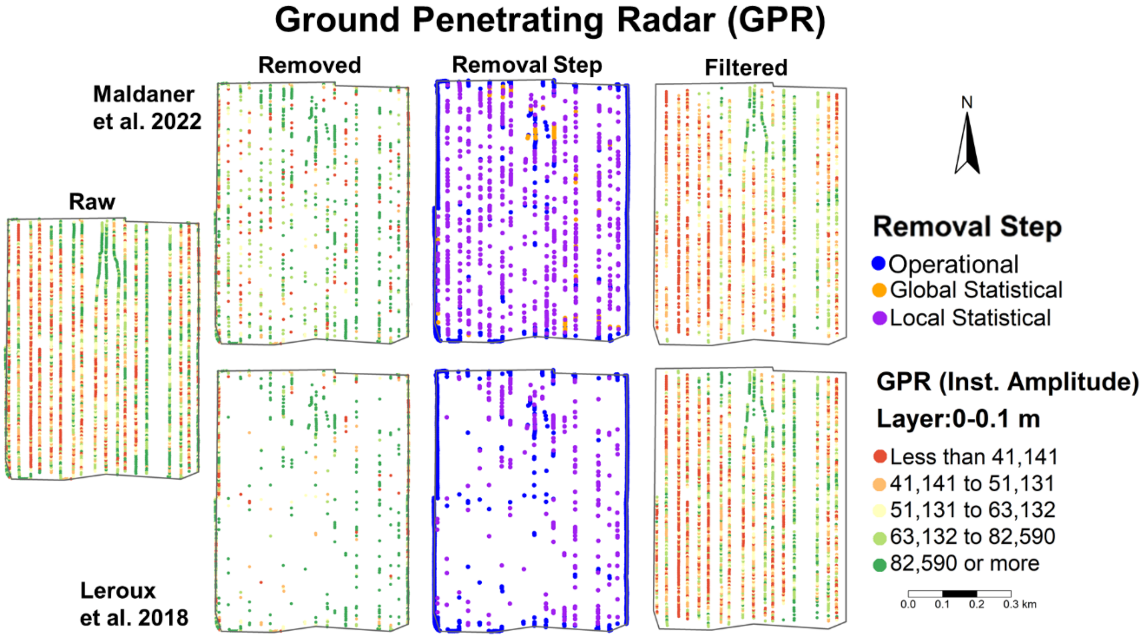

Fig. 4 Results of global and local filtering methodologies for ground penetrating radar (GPR) Instantaneous Amplitude at 0-0.1m depth. Results are presented for the methods of Maldaner et al. (2022), top, and Leroux et al. (2018), bottom. Points removed at each filtering step are classified at the Removal Step maps

The issue raised in the previous paragraph might cause the removal of important patches in the field which could affect further analysis using the filtered data. In addition, the results presented for GPR and γ-ray for Leroux et al. (2018) (Figures 4 and 5) suggest insufficient removal of spatial outliers. The maps obtained for these sensors when using this methodology still present some noise and do not reveal a well-delimited spatial distribution (Figures 4 and 5). On a separate note, the identification of spatial outliers based on the non-parametric methodology of Leroux et al. (2018) requires a few more extra steps than Maldaner et al. (2022) methodology, and therefore, is more computationally costly.

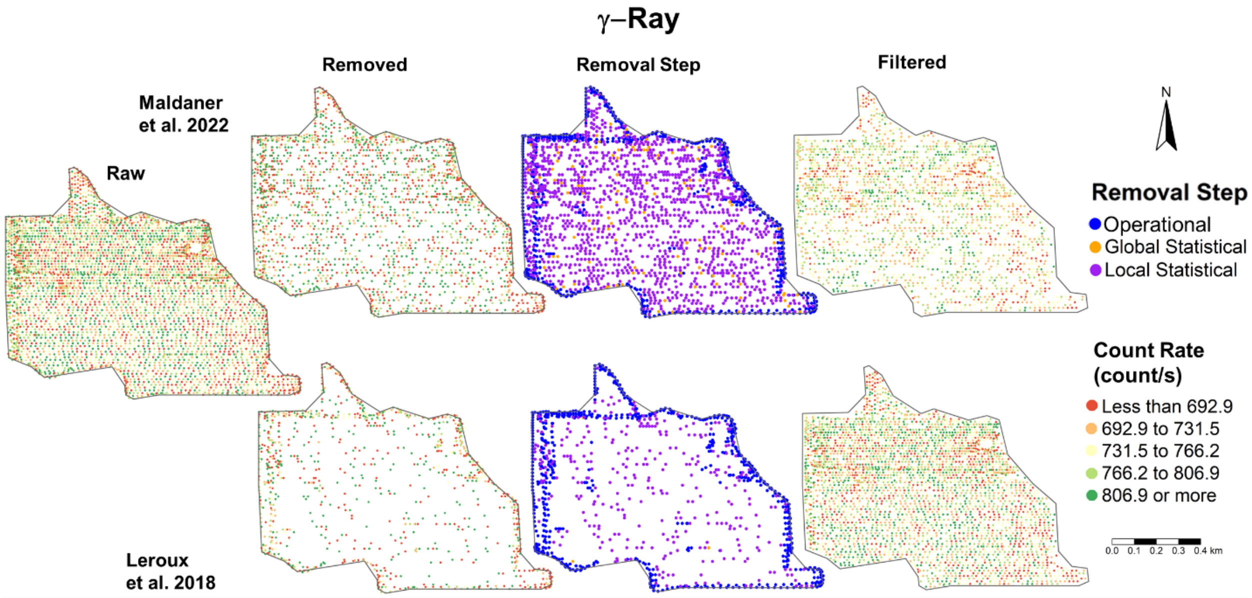

Fig. 5 Results of global and local filtering methodologies for γ-ray count rate. Results are presented for the methods of Maldaner et al. (2022), top, and Leroux et al. (2018), bottom. Points removed at each filtering step are classified at the Removal Step maps.

Based on the above analyses and results, Maldaner et al. (2022) methodology appears to be the most suitable to process the particular dataset used in the present study. Accordingly, its results were used for interpolation. However, as briefly mentioned during the presentation of the Global Statistical step, a disadvantage of this methodology is its sensitivity to a user-defined coefficient. When considering data batch processing, the need for a user-defined coefficient affects the flow and efficiency of the process. For the Local Statistical step, the same sensitivity as for the Global Statistical step was observed, and after trying a number of values, 0.05, 0.4, 0.1, and 0.5 were selected as the best coefficients for EMI, GPR, γ-ray, and GC, respectively.

Overall, the two methodologies presented advantages and disadvantages. The proposed filtering steps (operational, global, and local statistical) improved the data quality; nevertheless, there is still a need for the development and evaluation of more efficient, non-parametric methodologies for global and local statistical filtering, in order to improve the batch processing of these steps.

Interpolation

The four PSS datasets, after applying all of the above steps, had the normality of their distributions tested (α=0.05). It was rejected for all variables, so each set of data was submitted to a Box-Cox transformation. The four interpolation methods were applied to the transformed data. Results are reported in Table 3, where the values for OK may differ from those in Table 2 since no data transformation was performed at the evaluation step.

Table 3. Interpolation methods root mean squared error (RMSE) and coefficient of determination (R^2) resulted from 10-fold cross-validation for each proximal soil sensor.

Interpolation Method	EMI[a] – 0-1.5 m		GPR[b] – 0-0.1 m		γ-Ray – Count Rate		GC[c] – 0-0.3 m	
	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2
NN[d]	0.320	0.899	0.855	0.406	1.149	0.120	0.200	0.960
IDW[e]	0.258	0.934	0.699	0.516	0.875	0.235	0.159	0.975
OK[f]	0.228	0.949	0.665	0.558	0.858	0.263	0.203	0.960
UK[g]	0.231	0.947	0.662	0.563	0.859	0.262	0.143	0.980

[a] EMI – electromagnetic induction apparent electrical conductivity, [b]GPR – Ground Penetrating Radar, [c]GC – galvanic contact apparent electrical conductivity, [d]NN – Nearest Neighbor, [e]IDW – Inverse Distance Weighting, [f]OK – Ordinary Kriging, [g]UK – Universal Kriging

Overall, kriging presented the best performance for all the four different PSS data, while NN was the worst. Even though OK and UK performed better than the other two, automatic calculation and model fitting to a variogram is still challenging (Oliver and Webster, 2014). Thus, to properly apply geostatistics to specific data, there would be a need for some user analysis of the variogram and model. For example, analyzing the data from GC and its distribution on the X and Y coordinate, a trend was observed. In this case, UK is the most suitable kriging method instead of OK. However, an analysis of the data distribution and the variogram had to be performed to identify the trend and model it.

On the other hand, by automatically optimizing the IDW parameters similar results to kriging were obtained. For GC data, where a trend was observed, IDW even outperforms OK. In this scenario, when focusing on batch processing, IDW might be a good option. However, there are still some other interpolation options that could be evaluated, such as spatial random forest (Sekulić et al., 2020), support vector machines, and their combination with IDW (Willam et al., 2022). Although, these also need the optimization of their hyper-parameters, a time-consuming and computationally costly procedure.

For future work, this framework should be tested with simulated datasets, while implementing and/or developing more robust methodologies for the spatial outlier detection step. Newly developed interpolation methodologies, using machine learning and their combination with traditional interpolation methods, should also be tested.

Conclusion

A framework for batch processing of high-density anisotropic data from proximal soil sensors (PSS) was proposed and tested. It was found that data quality improved after applying the framework which used different methodologies at each step. However, the development of more computationally efficient and autonomous methods is required, as the methods available in the literature require expert thresholds or they are not robust enough for use with data with such variability as the PSS data collected in agricultural fields.

Acknowledgments

We would like to thank all Olds College crew involved in the collection of the proximal soil sensor. SoilOptix Inc. (Tavistock, ON, Canada) is also thanked for providing the complete gamma ray analysis. This research is part of the project "Agricultural Multi-Layer Data Fusion to Support Cloud-Based Agricultural Advisory Services" supported by Mitacs through the Mitacs Accelerate program.

References

Adamchuk, V. I., Hummel, J. W., Morgan, M. T., & Upadhyaya, S. K. (2004). On-the-go soil sensors for precision agriculture. *Computers and Electronics in Agriculture*, *44*(1), 71–91. https://doi.org/10.1016/j.compag.2004.03.002

Arslan, S., & Colvin, T. S. (2002). Grain yield mapping: Yield sensing, yield reconstruction, and errors. *Precision Agriculture*, *3*(2), 135–154. https://doi.org/10.1023/A:1013819502827

Barbulescu, A., Bautu, A., & Bautu, E. (2020). Optimizing inverse distanceweighting with particle swarm optimization. *Applied Sciences (Switzerland)*, *10*(6). https://doi.org/10.3390/app10062054

Blackmore, S., & Moore, M. (1999). Remedial Correction of Yield Map Data. *Precision Agriculture*, *1*(1), 53–66. https://doi.org/10.1023/A:1009969601387

Box, G. E. P., & Cox, D. R. (1964). An Analysis of Transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, *26*(2), 211–243. https://doi.org/10.1111/j.2517-6161.1964.tb00553.x

Byrd, R. H., Lu, P., Nocedal, J., & Zhu, C. (1995). A Limited Memory Algorithm for Bound Constrained Optimization. *SIAM Journal on Scientific Computing*, *16*(5), 1190–1208. https://doi.org/10.1137/0916069

Chung, S. O., Sudduth, K. A., & Drummond, S. T. (2002). Determining Yield Monitoring System Delay Time with Geostatiscal and Data Segmentation Approaches. *Transactions of the ASAE*, *45*(4), 915–926. https://doi.org/10.13031/2013.9938

Colaço, A. F., Richetti, J., Bramley, R. G. V., & Lawes, R. A. (2021). How will the next-generation of sensor-based decision systems look in the context of intelligent agriculture? A case-study. *Field Crops Research*, *270*(February), 108205. https://doi.org/10.1016/j.fcr.2021.108205

Ester, M., Kriegel, H.-P., Sander, J., & Xiaowei, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In E. Simoudis, J. Han, & U. Fayyad (Eds.), *Second International Conference on Knowledge Discovery and Data Mining* (pp. 226–231). Menlo Park, California: American Association for Artificial Intelligence (AAAI).

Fulton, J. P., & Port, K. (2018). Precision Agriculture Data Management. In *Precision Agriculture Basics* (pp. 169–187). https://doi.org/10.2134/precisionagbasics.2016.0095

Hengl, T. (2009). *A Practical Guide to Geostatistical Mapping*. http://spatial-analyst.net/book/

Hubert, M., & Van Der Veeken, S. (2008). Outlier detection for skewed data. *Journal of Chemometrics*, *22*(3–4), 235–246. https://doi.org/10.1002/cem.1123

Isaaks, E. H., & Srivastava, R. M. (1989). *Applied Geostatistics*. NY: Oxford Univeristy Press, Inc.

ISO. (2010). *ISO 12188-1:2010 Tractors and machinery for agriculture and forestry — Test procedures for positioning and guidance systems in agriculture — Part 1: Dynamic testing of satellite-based positioning devices*.

Ji, W., Biswas, A., Adamchuk, V., Perron, I., Cambouris, A., & Zebarth, B. (2018). Proximal soil sensing-led management zone delineation for potato fields. *14th International Conference on Precision Agriculture: 24 - 27 June*, 1–14.

Lee, D. H., Sudduth, K. A., Drummond, S. T., Chung, S. O., & Myers, D. B. (2012). Automated Yield Map Delay Identification Using Phase Correlation Methodology. *Transactions of the ASABE*, *55*(3), 743–752. https://doi.org/10.13031/2013.41506

Leroux, C., Jones, H., Clenet, A., Dreux, B., Becu, M., & Tisseyre, B. (2018). A general method to filter out defective spatial observations from yield mapping datasets. *Precision Agriculture*, *19*(5), 789–808. https://doi.org/10.1007/s11119-017-9555-0

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. https://doi.org/10.1007/s11119-013-9336-3

Maldaner, L. F., Molin, J. P., & Spekken, M. (2022). Methodology to filter out outliers in high spatial density data to improve maps reliability. *Scientia Agricola*, *79*(1), 1–7. https://doi.org/10.1590/1678-992x-2020-0178

Menegatti, L. A. A., & Molin, J. P. (2004). Removal of errors in yield maps through raw data filtering. *Revista Brasileira de Engenharia Agrícola e Ambiental*, *8*(1), 126–134. https://doi.org/10.1590/s1415-43662004000100019

Oliver, M. A., & Webster, R. (2014). A tutorial guide to geostatistics: Computing and modelling variograms and kriging. *CATENA*, *113*, 56–69. https://doi.org/10.1016/j.catena.2013.09.006

Ping, J. L., & Dobermann, A. (2005). Processing of yield map data. *Precision Agriculture*, *6*(2), 193–212. https://doi.org/10.1007/s11119-005-1035-2

R Core Team. (2021). R: A Language and Environment for Statistical Computing. Vienna, Austria. https://www.r-project.org/

Roberton, S. D., Lobsey, C. R., & Bennett, J. M. L. (2021). A Bayesian approach toward the use of qualitative information to inform on-farm decision making: The example of soil compaction. *Geoderma*, *382*(August 2020), 114705. https://doi.org/10.1016/j.geoderma.2020.114705

Rodrigues, F. A., Bramley, R. G. V., & Gobbett, D. L. (2015). Proximal soil sensing for Precision Agriculture: Simultaneous use of electromagnetic induction and gamma radiometrics in contrasting soils. *Geoderma*, *243–244*, 183–195. https://doi.org/10.1016/j.geoderma.2015.01.004

Saifuzzaman, M., Adamchuk, V., Buelvas, R., Biswas, A., Prasher, S., Rabe, N., et al. (2019). Clustering tools for integration of satellite remote sensing imagery and proximal soil sensing data. *Remote Sensing*, *11*(9). https://doi.org/10.3390/rs11091036

Schepers, A. R., Shanahan, J. F., Liebig, M. A., Schepers, J. S., Johnson, S. H., & Luchiari, A. (2004). Appropriateness of Management Zones for Characterizing Spatial Variability of Soil Properties and Irrigated Corn Yields across Years, 195–203.

Schossler, T. R., Mantovanelli, B. C., de Almeida, B. G., Freire, F. J., da Silva, M. M., de Almeida, C. D. G. C., & Freire, M. B. G. dos S. (2019). Geospatial variation of physical attributes and sugarcane productivity in cohesive soils. *Precision Agriculture*, *20*(6), 1274–1291. https://doi.org/10.1007/s11119-019-09652-y

Sekulić, A., Kilibarda, M., Heuvelink, G. B. M., Nikolić, M., & Bajat, B. (2020). Random forest spatial interpolation. *Remote Sensing*, *12*(10), 1–29. https://doi.org/10.3390/rs12101687

Shekhar, S., Lu, C. T., & Zhang, P. (2003). A unified approach to detecting spatial outliers. *GeoInformatica*, *7*(2), 139–166. https://doi.org/10.1023/A:1023455925009

Shepard, D. (1968). A two- dimensional interpolation function for irregularly-spaced data. In R. B. S. Blue & A. M. (Eds. . Rosenberg (Eds.), *Proceedings of the 1968 ACM National Conference* (pp. 517–524). New York: ACM Press.

Simbahan, G. C., Dobermann, A., & Ping, J. L. (2004). Screening Yield Monitor Data Improves Grain Yield Maps. *Agronomy Journal*, *96*(4), 1091–1102. https://doi.org/10.2134/agronj2004.1091

Singh, A. K., & Lalitha, S. (2018). A novel spatial outlier detection technique. *Communications in Statistics - Theory and Methods*, *47*(1), 247–257. https://doi.org/10.1080/03610926.2017.1301477

Spekken, M., Anselmi, A. A., & Molin, J. P. (2013). A simple method for filtering spatial data. In J. V. Stafford (Ed.), *Precision agriculture'13: Proceedings of the 9th European conference on precision agriculture* (pp. 259–266). Wageningen: The Netherlands: Wageningen Academic Publishers.

Sudduth, K. A., & Drummond, S. T. (2007). Yield Editor: Software for Removing Errors from Crop

Yield Maps. *Agronomy Journal*, *99*(6), 1471–1482. https://doi.org/10.2134/agronj2006.0326

Sudduth, K. A., Drummond, S. T., & Myers, D. B. (2012). Yield editor 2.0: Software for automated removal of yield map errors. *American Society of Agricultural and Biological Engineers Annual International Meeting 2012, ASABE 2012*, *4*(12), 3378–3391. https://doi.org/10.13031/2013.41893

Sun, W., Whelan, B., McBratney, A. B., & Minasny, B. (2013). An integrated framework for software to provide yield data cleaning and estimation of an opportunity index for site-specific crop management. *Precision Agriculture*, *14*(4), 376–391. https://doi.org/10.1007/s11119-012-9300-7

Vega, A., Córdoba, M., Castro-Franco, M., & Balzarini, M. (2019). Protocol for automating error removal from yield maps. *Precision Agriculture*, *20*(5), 1030–1044. https://doi.org/10.1007/s11119-018-09632-8

Willam, G., Domingos, P., Magalhães, S., Pereira, G. W., Valente, D. S. M., de Queiroz, D. M., et al. (2022). Soil mapping for precision agriculture using support vector machines combined with inverse distance weighting. *Precision Agriculture*, (0123456789). https://doi.org/10.1007/s11119-022-09880-9