DOI: http://doi.org/10.1590/1678-992X-2020-0178

ISSN 1678-992X

SCIENTIA AGRICOLA

Research Article

Methodology to filter out outliers in high spatial density data to improve maps reliability

Leonardo Felipe Maldaner*, José Paulo Molin, Mark Spekken

Universidade de São Paulo/ESALQ – Depto. de Engenharia de Biossistemas – Lab. de Agricultura de Precisão, C.P. 09 – 13418-900 – Piracicaba, SP – Brasil.
*Corresponding author <leonardofm@usp.br>

Edited by: Thiago Libório Romanelli

Received June 18, 2020
Accepted September 11, 2020

ABSTRACT: The considerable volume of data generated by sensors in the field presents systematic errors; thus, it is extremely important to exclude these errors to ensure mapping quality. The objective of this research was to develop and test a methodology to identify and exclude outliers in high-density spatial data sets, determine whether the developed filter process could help decrease the nugget effect and improve the spatial variability characterization of high sampling data. We created a filter composed of a global, anisotropic, and an anisotropic local analysis of data, which considered the respective neighborhood values. For that purpose, we used the median to classify a given spatial point into the data set as the main statistical parameter and took into account its neighbors within a radius. The filter was tested using raw data sets of corn yield, soil electrical conductivity (ECa), and the sensor vegetation index (SVI) in sugarcane. The results showed an improvement in accuracy of spatial variability within the data sets. The methodology reduced RMSE by 85 %, 97 %, and 79 % in corn yield, soil ECa, and SVI respectively, compared to interpolation errors of raw data sets. The filter excluded the local outliers, which considerably reduced the nugget effects, reducing estimation error of the interpolated data. The methodology proposed in this work had a better performance in removing outlier data when compared to two other methodologies from the literature.
Keywords: precision agriculture, local analysis, map accuracy

Introduction

Sensors in agricultural fields collect large amounts of spatial data needed for site-specific management; however, this may come with a considerable quantity of defective data that need to be excluded to provide quality to maps (Spekken et al., 2013; Lyle et al., 2014). In maps, outliers are spatially referenced objects whose non-spatial attribute values are significantly different from the corresponding values in their respective spatial neighborhoods (Shekhar et al., 2003). They can be observed in local regions that demand specific analysis, making them difficult to exclude (Singh and Lalitha, 2017).

Authors have applied sequences of filters to remove defective data errors (Ping and Dobermann, 2005; Simbahan et al., 2004; Menegatti and Molin, 2004; Arslan and Colvin, 2002). Some filters require prior knowledge of the target factor to establish upper and lower thresholds to identify outliers; however, data removed outside these boundaries were the major causes of losses of good data (Spekken et al., 2013).

To filter a large amount of PA data, Leroux et al. (2018) created a data-filtering algorithm dedicated to data generated by the onboard sensor. The most abnormal data points are classified as defective observations based on a density-clustering algorithm. Vega et al. (2019) created a script for the software R to automate error removal from yield maps. First, the data were screened by filtering null and edge yield values, as well as global outliers. Second, spatial outliers or local defective observations were deleted by using the Local Moran index of spatial autocorrelation and the Moran plot.

To create a user-friendly tool, Spekken et al. (2013) developed a generic software capable of identifying and filtering erroneous data points that are inconsistent with their neighboring points. Although this software is easy to use, removal of erroneous data could also eliminate relevant data. According to Leroux et al. (2018), data filtering methods have to be robust enough to ensure accuracy to the decision-making process. The objective of this research was to develop and test a filter to identify and exclude spatial outliers in the high-density spatial data set. We also investigated whether the filter could help decrease the sampling error and improve the characterization of spatial variability in high sampling spatial data.

Materials and Methods

Data sets

We processed data sets generated by sensors in high spatial resolution for agricultural applications. The methodology was tested using raw data sets of corn yield, soil ECa, and SVI. Yield is the main information for site-specific field management. Data on corn yield were generated by a yield monitor, composed of sensors that measure grain flow inside the harvester elevator thus estimating the number of grains harvested (Molin et al., 2015). Soil ECa was generated from a four-point system composed of a metal structure with six cutting disks serving as electrodes that, in contact with the soil, measure the electric current (Rabelo et al., 2014). According to Molin and Rabelo (2011), ECa data are generally used to estimate soil texture, because in the absence of salinity in the soil, ECa correlates with the water content of the soil, consequently

correlating with the texture or relief of the field. The sugarcane vegetation index was generated by an active canopy optical sensor that measures canopy reflectance of plants commonly used for the variable rate of N in the site-specific application in sugarcane (Amaral et al., 2018).

Each raw data set was recorded with different frequencies and different widths between the travel paths (Figure 1). They were organized by rows in a text file. The text files must contain at least three numeric attributes: two attributes with latitude and longitude, and the target attribute that was subjected to the filter. The first row of the file must contain a header (denomination) of attributes. Coordinates must be in either WGS 84 datum provided in geographic coordinates (decimal degrees), commonly used for storage of coordinates in agricultural data loggers, or the metric form (UTM). The headers were named with initials "Lat" and "Long" or "X" and "Y" for automatic identification, while the user must inform the column to filter the attribute variable. The original coordinates in the geographic format (decimal degrees) were then converted into UTM coordinates, allowing the points to be analyzed in a regular metric 2D plane and to calculate distances between them.

Filtering data

According to Vega et al. (2019), the global filter avoids variance inflation in the local analysis due to very low or very high data values; therefore, a global filter method was added to precede the local filter. In the global filter, the median value of the attribute points values under analysis is used to calculate the upper (Eq. 1) and lower (Eq. 2) cut-off limits of discrepant values (Maldaner and Molin, 2020).

$$LimS = M_k + M_k \times v \quad (1)$$

$$LimI = M_k - M_k \times v \quad (2)$$

where: $LimS$ is the upper limit; $LimI$ is the lower limit; M_k is the median of all values located in the data set; v is the maximum variation accepted for the median. A global outlier in the data set is a point with a value greater or smaller than upper and lower cut-off limits, respectively.

The local filter was divided into two steps: anisotropic and isotropic local filters. The anisotropic filter created by Maldaner and Molin (2020) was used to filter sugarcane yield data. The filter detected all points located in a radius range (R) around a point x_i within a single direction (Figure 1). The point x_i is compared with previous and subsequent k neighbors. The k is the number of neighbors whose Euclidean distance is less than or equal to the radius R (blue line in Figure 2). The median of these k neighbors was calculated, and Eq. 1 and Eq. 2 were applied to point x_i. If the value of the point x_i was greater or smaller than the upper and lower cut-off limits, it was considered a local outlier and then excluded from the data set.

In the isotropic filter, the methodology created by Spekken et al. (2013) was adapted to identify outliers in the data set. The isotropic filter, by contrast, detected all k points neighbors located in an R around a point x_i in any direction (Figure 2). Then, the median of these k neighbors was calculated, and Eq. 1 and Eq. 2 was applied to point x_i. While Spekken et al. (2013) added weight to the points with values outside the cut-off limits, our filter methodology excludes the point x_i with a value greater or smaller than the upper and lower cut-

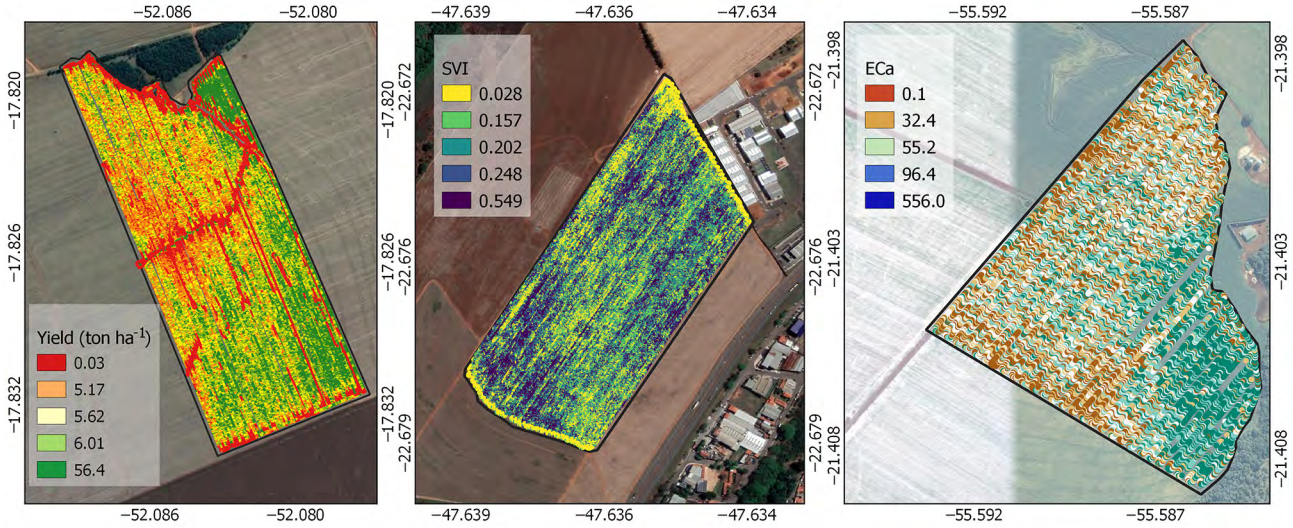

Figure 1 – Raw data sets of corn yield, sugarcane vegetation index (SVI), and soil electrical conductivity (ECa).

off limits. Finally, the points not excluded by the filter were saved in a text file.

Software development

We built an algorithm-application with the methodology to remove spatial outliers in the software NetBeans IDE 8.1, which is a free-integrated development environment and is an open-source for the development of desktop applications when using the Java platform. The application was structured on a single interface (Figure 3) with three user inputs. The user input variable was the maximum acceptable variation of the median used to calculate Eq. 1 and Eq. 2 for the global and local filter. The value of the radius R was used to identify the neighboring spatial points in the local filter. The filtering

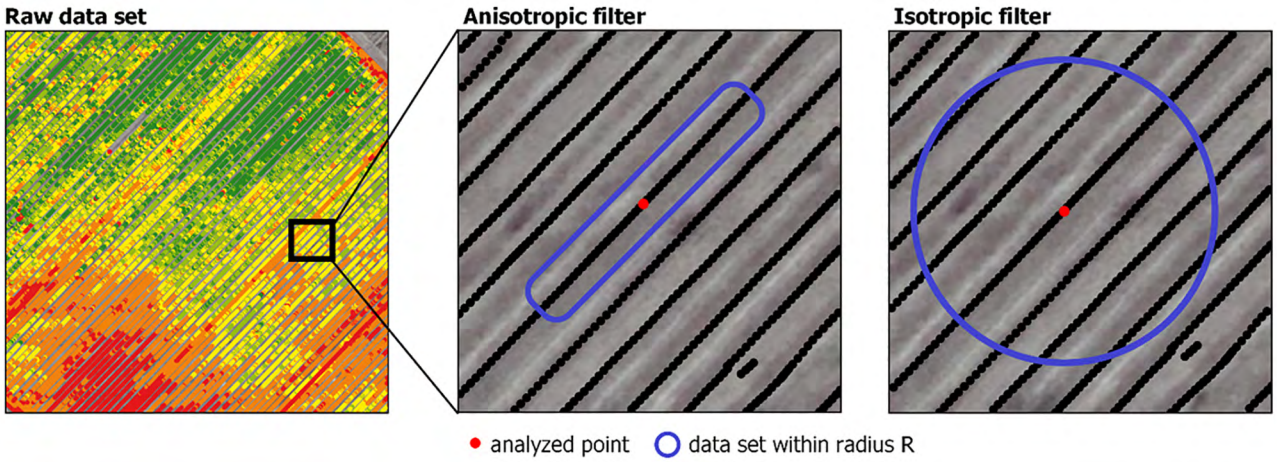

Figure 2 – Identification of neighboring points in the anisotropic and isotropic filter.

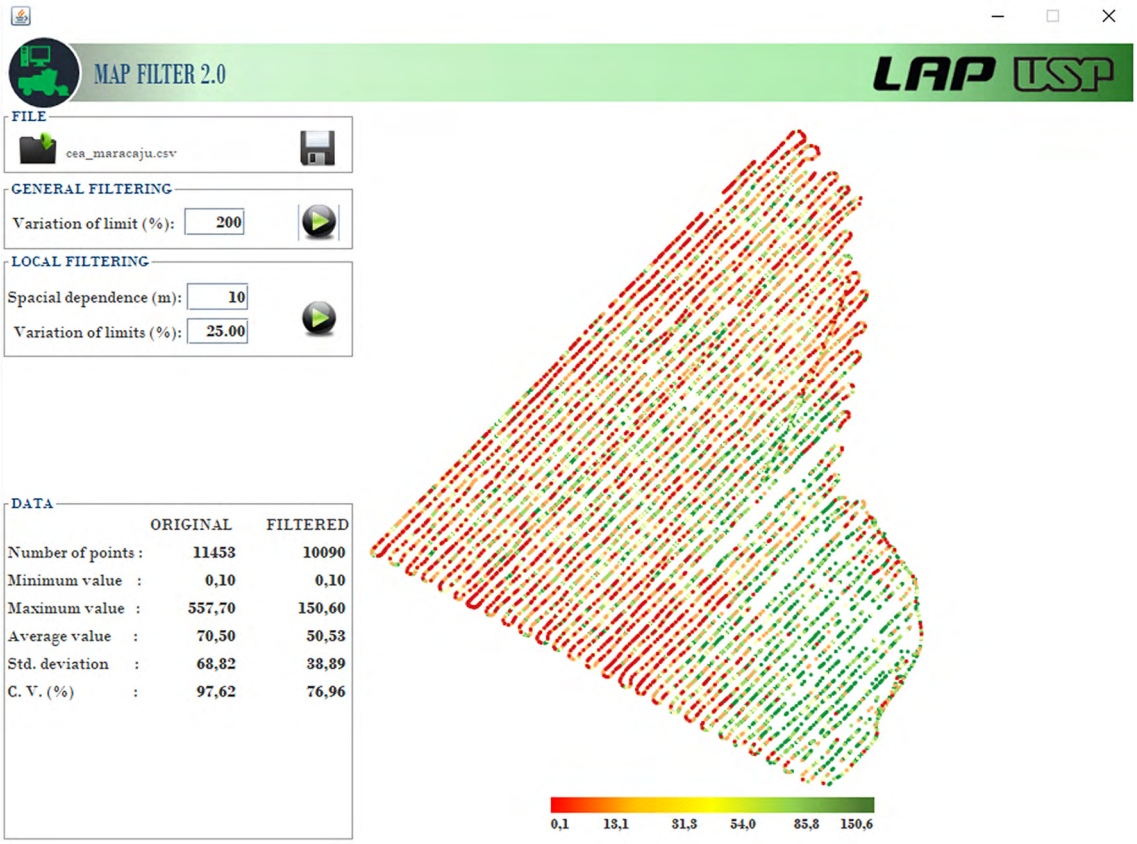

Figure 3 – The interface of the algorithm-application.

data were plotted on the display and the application calculates the descriptive statistic before and after the cleaning process. Therefore, the user could perform a visual analysis of the data and reapply the global and local filtering with other input values.

Analysis

A process with the different variables R and *v* was performed to analyze the influence of these variables on the identification of spatial outliers. Each data set was filtered using an initial R value 1.5 times the width of the travel paths, and at the end of each processing step, the radius (R) was increased by adding a path width.

The spatial dependence of the raw data generated by the semivariogram was the maximum R value (Table 1). For each R, the raw data processed with *v* between 5 % and 50 % were tested. Table 1 presents the numbers of the filter process for each raw data.

The impact of the proposed filter to exclude spatial outliers on raw data was quantified by evaluating changes of semivariogram parameters and the kriging prediction accuracy between raw and cleaned data (Vega et al., 2019). To choose the output with the lower root mean square error (RMSE) of cross-validation result (Isaaks and Srivastava, 1989), semivariograms were individually modeled to test the spherical, exponential, and Gaussian models. These models were fitted to each raw dataset and after applying the filter.

The data filtered were compared by the methods of Spekken et al. (2013) and Vega et al. (2019). All statistical analyses were performed using R software (R Core Team, version 3.6.0) using the gstat library (Pebesma, 2004) to study the evolution of semivariogram parameters and evaluate the kriging prediction accuracy.

Results and Discussion

The global filter excluded points with values below 0.10 mS m^{-1}, 3.06 Mg ha^{-1}, and 0.09, and above 150.60 mS m^{-1}, 8.09 Mg ha^{-1}, 0.31 for the data set of soil ECa, corn yield, and SVI respectively (Figure 4). This accounted for the removal of 12 %, 15 %, and 14 % of raw data points, respectively. Removal of global outliers substantially decreased the mean of the soil ECa and SVI by almost 29 % and 2 %. This is because the high values points have more influence than the lower values, unlike grain yield data filtering in which most errors in yield values are below average or close to zero (Vega et al., 2019; Leroux et al., 2018; Spekken et al., 2013; Sudduth and Drummond, 2007; Menegatti and Molin, 2004). There was an increase of 5 % in the corn yield mean after the removal of the global outliers (Figure 4). The filtering processes were compared by removing data outside the mean ± 4 SD (Vega et al., 2019). However, it was not possible to exclude all global outliers (Figure 4). Despite the use of the mean ± 3 SD criteria, as suggested by Vega et al. (2019), there was additional removal of

Table 1 – Raw data set characterization.

Data set	Min	Max	Mean	SD	CV (%)	Freq.	Width	n	n ha^{-1}	Range	*nP*
						Hz	m			m	
ECa	0.1	557.7	70.5	68.8	102.4	1.0	15.0	11453	204.9	480.0	320
CY	0.0	99.3	5.4	4.2	129.6	1.0	5.7	58186	502.4	319.0	560
SVI	0.0	0.6	0.2	0.1	257.2	5.0	1.5	423040	18393.0	19.2	320

ECa = Apparent soil electricity conductivity (mS m^{-1}); CY = Corn yield (Mg ha^{-1}); SVI = Sensor vegetation index in sugarcane; SD = standard deviation; CV = coefficient of variation; Freq. = Data collection frequency; Width = travel path width; n = number of points in raw data set; Range = a range of the spatial dependence calculated by the semivariogram; *nP* = The numbers of the process for each raw data.

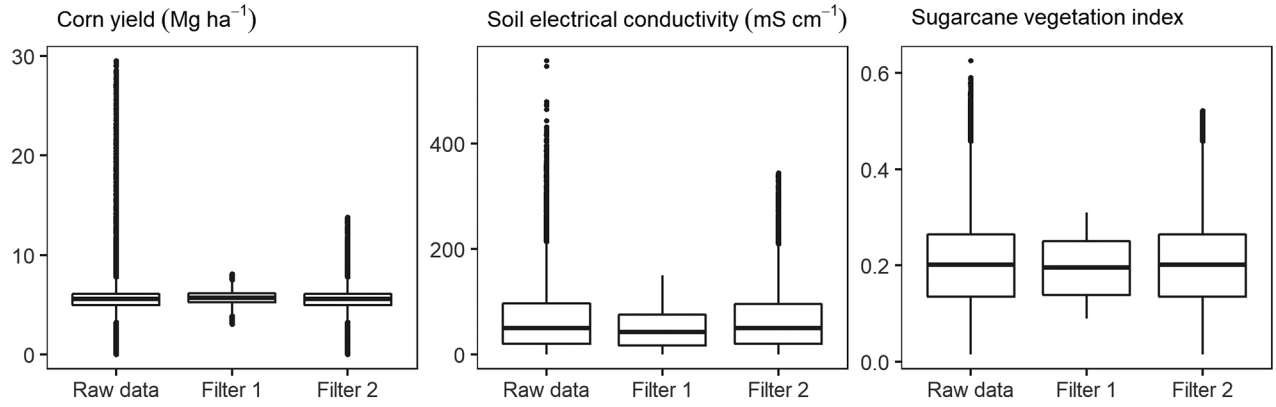

Figure 4 – Boxplot of raw data set before filtering, after the global filter of the current study (Filter 1), and the methodology of Vega et al. (2019) (Filter 2).

valid points, which eliminated data with low and high spots within the field, possibly because extreme values influence the methodology that uses these parameters (mean and SD). The mean is used for normal numerical distributions with a small number of discrepant values (Hubert and Van der Veeken, 2008). This is in contrast to data collected by sensors in agricultural fields, which present a large number discrepant values and, notably, mostly in a non-symmetric distribution (Menegatti and Molin, 2004; Vega et al., 2019; Leroux et al., 2018). This does not happen with the median because it returns the central tendency for distorted numerical distribution. The use of the upper and lower cut-off limits calculated from the median allowed removing more points (Figure 4) without loss of data that characterized the spatial variability, keeping data that showed the low and high spots within the field.

When using the local filter, the lower median variation within the radius resulted in a greater number of points excluded. However, the R values do not influence the number of points excluded. By using 5 % variation and radius equal to the semivariogram range (Table 1), 71 %, 34 %, and 57 % of the points were excluded in the data set of corn yield, soil ECa, and SVI, respectively. The amount of data excluded from corn yield data set was larger than what is normally observed done by other authors (Menegatti and Molin, 2004; Simbahan et al., 2004; Sudduth and Drummond, 2007; Sun et al., 2013; Leroux et al., 2018; Vega et al., 2019). This indicates that the use of 5 % may exclude valid data, leading to loss of data information on small scales. By using the coefficient of variation (CV) in the local filter, Spekken et al. (2013) suggested that a range between 10 and 25 % of CV is capable of eliminating most spatial outliers in yield maps.

Global and local filtering reduced errors between the actual and predicted values by the semivariogram in the cross-validation when compared to the raw data set (Table 2). There was a reduction of 85 %, 97 %, and 79 % in the RMSE for the corn yield, soil ECa, and SVI data sets, respectively. The smaller RMSE in the filtered data sets presented filter cut-off limits with a median variation of 5 %. Therefore, the lower values of median variation resulted in smaller interpolation error (smaller RMSE). There was a decrease in the correlation coefficient values during the cross-validation, dropping from 0.8 to 0.5 (corn yield), 0.9 to 0.8 (soil ECa), and 0.9 to 0.5 (SVI), while the median variation increased. According to Simbahan et al. (2004), the prediction error decreases with the increased detection of local outliers in the protocol for error. These results are smaller than the RMSE of data filtered by the methodologies of Spekken et al. (2013) and Vega et al. (2019). These two methodologies also presented lower values than the raw data. However, the methodology using the Local Moran index (Vega et al., 2019) could not process the SVI data due to the high density of sampled points per area.

All raw data sets present greater nugget effects (Leroux et al., 2018) and the presence of the global and local outliers influenced these values. There was a considerable reduction in the nugget effects after the exclusion of spatial outliers. The exclusion of outliers when using the local filter with a median variation of 10 % was more efficient than the methodologies of Spekken et al. (2013) and Vega et al. (2019), which presented a reduction of 99 % in the corn yield, 80 % in the soil ECa, and 82 % in the SVI compared to their respective raw data sets. Other studies that compared the methods to filter grain yield data also showed a decrease in the nugget effects after data filtering (Leroux et al., 2018; Menegatti and Molin, 2004; Sudduth and Drummond, 2007). The primary aim in filtering spatial errors is to improve the interpolation by kriging, and a reduction in nuggets usually indicates improvement in data quality.

Table 2 – Geostatistical analysis of the data sets before and after applying the filter.

	Corn yield[1]			Soil electrical conductivity			Sensor vegetation index		
Median variation[2]	Nugget	R2[3]	RMSE	Nugget	R2	RMSE	Nugget	R2	RMSE
5	0.177	**0.844**	**0.157**	91.998	**0.981**	**2.453**	0.00080	**0.915**	**0.008**
10	**0.177**[5]	0.683	0.278	**89.630**	0.965	3.660	**0.00078**	0.843	0.014
15	0.178	0.595	0.355	92.350	0.942	5.141	0.00079	0.763	0.020
20	0.177	0.547	0.402	89.980	0.917	6.403	0.00078	0.684	0.025
25	0.178	0.523	0.431	92.176	0.891	7.548	0.00080	0.621	0.029
30	0.177	0.514	0.446	90.208	0.866	8.606	0.00078	0.577	0.031
35	0.179	0.509	0.453	93.680	0.848	9.613	0.00082	0.555	0.033
40	0.179	0.507	0.456	90.737	0.821	10.616	0.00081	0.540	0.034
45	0.180	0.506	0.457	96.517	0.811	11.722	0.00082	0.533	0.035
50	0.179	0.506	0.457	92.994	0.805	12.560	0.00080	0.527	0.035
Spekken et al. (2013)	0.780	0.589	0.850	106.309	0.823	7.891	0.00070	0.698	0.027
Vega et al. (2019)	0.578	0.573	0.989	110.967	0.885	7.832	-[4]	-	-
Raw data set	17.810	0.539	1.057	261.400	0.658	105.2	0.00441	0.510	0.038

[1]The results represent the average values of all files processed according to Table 1. [2]Median variations used in the Eq. 1 and Eq. 2 (cut-off limits). [3]The correlation coefficient between actual and predicted by the semivariogram in the cross-validation. [4]The methodology by Vega et al. (2019) was unable to filter the data set of the sugarcane vegetation index. [5]Lowest value for each parameter.

However, even after the reduction of nuggets by the methodologies of Spekken et al. (2013) and Vega et al. (2019), it is still possible to identify some spatial outliers in the yield map (Figure 5). These methodologies were able to exclude data with only high variation in relation to their neighbors, such as the width platform error in the corn yield data set. Additionally, all points close to field edges were excluded, which should be included as valid data. On the other hand, the filter proposed was capable of excluding erroneous data, such as harvester feed and fill time errors, while keeping the valid data points.

The method proposed (Method 3 in Figure 5) was efficient in filtering points whose values were inconsistent with the neighboring points, which represented most spatial outliers in the maps. The filtering of local spatial outliers resulted in noise reduction within the field, smoothing the variation in values. A certain degree of data smoothing in the field is necessary for interpretation of maps and their use in site-specific management practices (Blackmore and Moore, 1999). Both methodologies of Spekken et al. (2013) and Vega et al. (2019) kept small variations within the field. This was the case for the soil ECa, where both methodologies kept small variations in small distances. The density of points is significantly high, even after the removal of local outlier points. Furthermore, high variability in the SVI values was expected even after filtering. This was due to the large density of points collected by the on-the-go sensor, especially because sugar cane has high biomass variability at short distances (Amaral et al., 2018). The filter proposed characterized the regions (stains) of high and low SVI values within the field.

The algorithm-application developed demonstrated the potential practical use of filtering spatial data by end-users and it was granted the Brazilian patent n° BR512019002014-6.

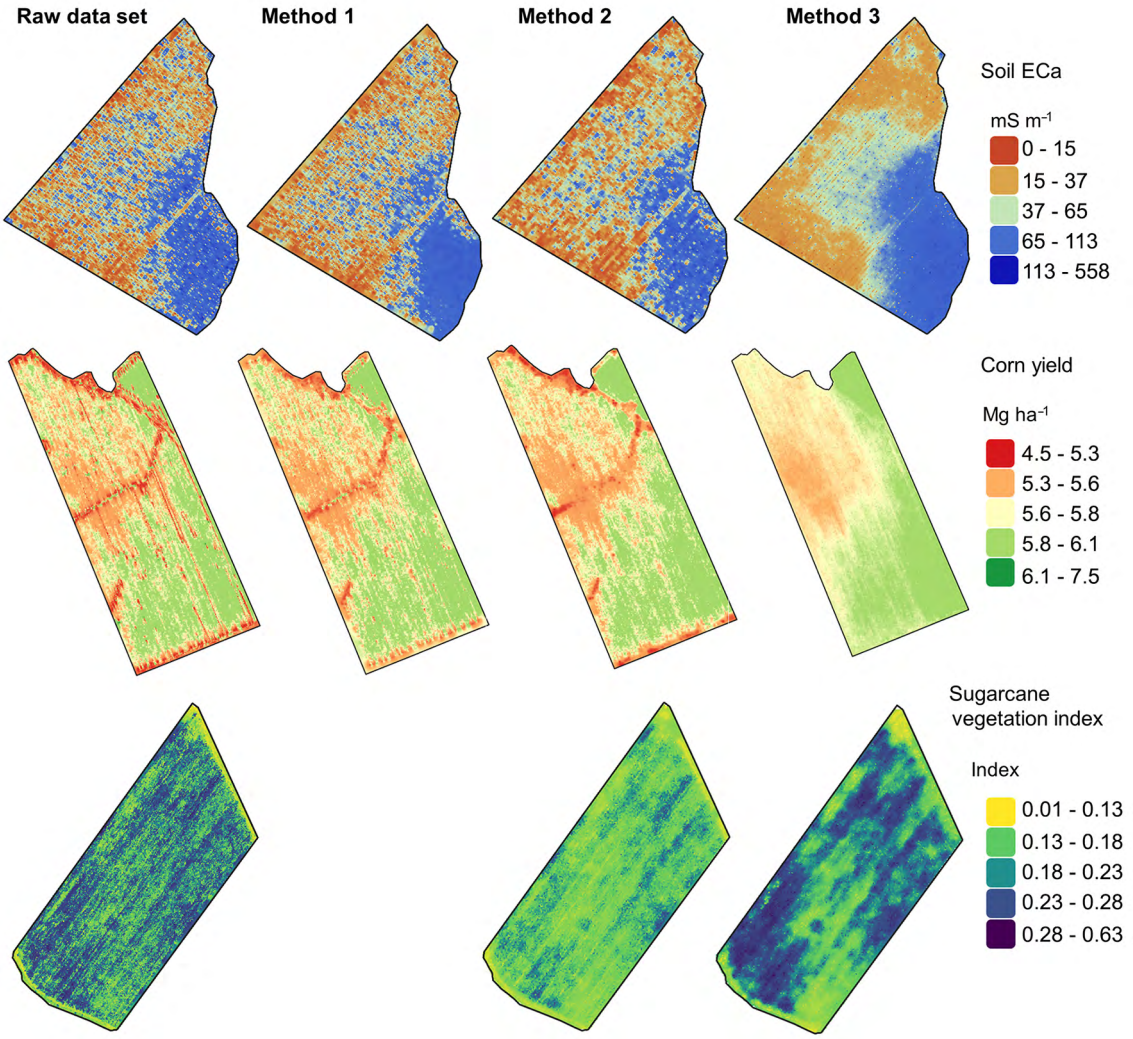

Figure 5 – Maps of the data sets after the filtering process. Method 1: Vega et al. (2019); Method 2: Spekken et al. (2013); Method 3: proposed in this study.

Conclusion

According to the results, the methodology proposed in this work had a better performance in removing outlier data when compared to the other two methodologies. The algorithm-application is a simple tool that could be used in different high-density data sets collected by sensors in agriculture. The results, based on different data sets, showed that the filter improved accuracy of maps. The filter reduced the RMSE compared to interpolation errors of the raw data. The filter excluded the local outliers, which considerably reduced the nugget effects. The global and local filter smoothed the data, characterizing the regions of greater and lower values of the attributes used in this study within the field. The algorithm-application developed has the potential for daily use by end-users as it is practical for filtering spatial data.

Acknowledgments

The authors would like to thank the Brazilian National Council for Scientific and Technological Development (CNPq) for the Ph.D. scholarship granted to the first author (Process 168643/2017-0). We would also like to thank Michael James Stablein of the University of Illinois Urbana-Champaign for his translation services and review of this work.

Authors' Contributions

Conceptualization: Molin, J.P.; Spekken, M.; Maldaner, L.F. **Design of methodology:** Molin, J.P.; Spekken, M.; Maldaner, L.F. **Software development:** Spekken, M.; Maldaner, L.F. **Writing and editing:** Maldaner, L.F.; Molin, J.P.

References

Amaral, L.R.; Trevisan, R.G.; Molin, J.P. 2018. Canopy sensor placement for variable-rate nitrogen application in sugarcane fields. Precision Agriculture 19: 147-160.

Arslan, S.; Colvin, T.S. 2002. Grain yield mapping: yield sensing, yield reconstruction, and errors. Precision Agriculture 3: 135-154.

Blackmore, S.; Moore, M. 1999. Remedial correction of yield map data. Precision Agriculture 1: 53-66.

Hubert, M.; Van der Veeken, S. 2008. Outlier detection for skewed data. Journal of Chemometrics 22: 235-246.

Isaaks, E.H.; Srivastava, R.M. 1989. An Introduction to Applied Geostatistics. Oxford University Press, New York, NY, USA.

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 data sets. Precision Agriculture 19: 789-808.

Lyle, G.; Bryan, B.; Ostendorf, B. 2014. Post-processing methods to eliminate erroneous grain yield measurements: review and directions for future development. Precision Agriculture 15: 377-402.

Maldaner, L.F.; Molin, J.P. 2020. Data processing within rows for sugarcane yield mapping. Scientia Agricola 77: e20180391.

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: 126-134 (in Portuguese, with abstract in English).

Molin, J.P.; Rabello, L.M. 2011. Studies about soil electrical conductivity measurements. Engenharia Agrícola 31: 90-101 (in Portuguese, with abstract in English).

Molin, J.P.; Amaral, L.R.; Colaço, A. 2015. Precision Agriculture = Agricultura de Precisão. Oficina de Textos, São Paulo, SP, Brazil (in Portuguese).

Pebesma, E.J. 2004. Multivariable geostatistics in S: the gstat package. Computers & Geosciences 30: 683-691.

Ping, J.L.; Dobermann, A. 2005. Processing of yield map data. Precision Agriculture 6: 193-212.

Rabello, L.M.; Bernardi, A.C.C.; Inamasu, R.Y. 2014. Soil Electric Conductivity Aparent. p. 48-57. In: Bernardi, A.C.C.; Naime, J.M.; Resende, A.V.; Bassoi, L.H.; Inamasu, R.Y., eds. Precision farming: results from a new look = Agricultura de precisão: resultados de um novo olhar. Embrapa, Brasília, DF, Brazil (in Portuguese, with abstract in English).

Shekhar, S.; Lu, C.T.; Zhang, P.S. 2003. A unified approach to detecting spatial outliers. Geoinformática 7: 139-166.

Simbahan, G.C.; Dobermann, A.; Ping, J.L. 2004. Screening yield monitor data improves grain yield maps. Agronomy Journal 96: 1091-1102.

Singh, A.K.; Lalitha, S. 2017. A novel spatial outlier detection technique. Communications in Statistics-Theory and Methods 47: 247-257.

Spekken, M.; Anselmi, A.A.; Molin, J.P. 2013. A simple method for filtering spatial data. p. 259-266. In: Stafford, J.V., ed. Precision agriculture. Wageningen Academic Publishers, Wageningen, The Netherlands.

Sudduth, K.; Drummond, S.T. 2007. Yield editor: software for removing errors from crop yield maps. Agronomy Journal 99: 1471.

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: 376-391.

Vega, A.; Córdoba, M.; Castro-Franco, M.; Balzarini, M. 2019. Protocol for automating error removal from yield maps. Precision Agriculture 21: 1-15.