Identifying and filtering out outliers in spatial datasets

Leonardo F. Maldaner, Lucas P. Corrêdo, Tiago R. Tavares, Luiz G. Mendez, Cassio Duarte, José P. Molin

Precision Agriculture Laboratory, Biosystems Engineering Department, University of São Paulo, Piracicaba-SP, Brazil.

A paper from the Proceedings of the
14th International Conference on Precision Agriculture
June 24 – June 27, 2018
Montreal, Quebec, Canada

Abstract. *Outliers present in the dataset is harmful to the information quality contained in the map and may lead to wrong interpretations, even if the number of outliers to the total data collected is small. Thus, before any analysis, it is extremely important to remove these errors. This work proposes a sequential process model capable of identifying outlier data when compared their neighbors using statistical parameters. First, limits are determined based on the median range of the values of all the points contained in the dataset. Second, the neighbors are located within the range of the point under analysis. In the anisotropic process, neighbors are defined in a single direction, and then the calculation of median is with the values of the neighboring points located within the radius range next to the point under analysis. Finally, an isotropic process is conducted, where the neighbors are defined and located within the radius range, and the median value is identified. Outliers are data that deviate above or below a given percentage of a set median value. Statistical and geostatistical analysis of the data before and after this process was performed, indicating it was effective in eliminating outliers in the spatial datasets evaluated. The median limits eliminated most of the points with discrepant values from the processed datasets. The anisotropic and isotropic processes eliminated outliers in relation to their neighbors at small distances, reducing the previous nugget values and improving the characterization of the spatial dependence of the datasets.*

Keywords. *Erroneous data, filter maps, processing*

Introduction

Precision farming is dependent on reliable information about the production process and its environmental and physico-chemical parameters (Reiser et al., 2017). The development of positioning technologies along with sensors have narrowed the spatial resolution and increased the amount of data collected from farm-fields (Spekken et al., 2013). While this considerable volume of data is critical for field management and decision-making, these datasets must be used with great caution (Leroux et al., 2018). Data generated by sensors and collected automatically present systematic errors insert in the data set and a post-processing is necessary to eliminate these errors.

Spatial yield data are the most important information for conducting site-specific management of the soil and crop. Outliers present in the spatial dataset is harmful to the information quality contained in the map and may lead to wrong interpretations. A spatial outlier is a spatially referenced object whose non-spatial attribute values are significantly different from those of other spatially referenced objects in its spatial neighborhood (Shekhar and Chawla, 2002).

Blackmore and Moore (1999) reviewed the errors related to yield maps, and for this kind of data, it is necessary to take into account: errors of sensor yield and moisture measurement, harvester fill mode error in headlands, GNSS positioning errors, driver errors, harvester emptying mode error and file write errors. Different methods that applied sequences of filters, which classify, identify and remove spatial outlier have been developed (Leroux et al., 2018; Ping and Dobermann, 2005; Simbahanet et al., 2004; Menegatti and Molin, 2004; Arslan and Colvin, 2002; Blackmore and Moore, 1999).

However, the use of these methodologies becomes restricted only to crop yield data. General methods are required to process datasets from multiple machine types, regardless of the level of additional equipment installed (Leroux et al., 2018). Spekken et al. (2013) developed a generic method able to identify and filter out erroneous data points that are inconsistent with its neighboring points. The method identifies groups of points within a range of one point and retrieves the variation of a target value associated to these, and a variation threshold defines the suitability of the point.

Although there are a large number of methods for identifying and removing spatial discrepant data from a heterogeneous dataset, it is difficult to establish standards for comparing a series of data using the same filtering configurations once it is influenced by each map producer (Spekken et al., 2013). Even though good results have been observed using the method developed by Spekken et al. (2013), it was observed that a pre-processing of the data increases the efficiency of the methodology in the identification and removal of the spatial outlier data. In this sense, the objective of this study was to propose a sequential process model capable of identifying outlier data, when compared to its neighbors, using simple statistical parameters.

Material and Methods

We take three step to remove spatial outlier data: global filter, anisotropic local filter and isotropic local filter. In this procedure, the filtering makes use of the modified methodology of Sudduth et al. (2012), which proposes global filtering of yield data based on the removal of extreme data from a normal distribution of values. The limits for exclusion of outliers are calculated through eq. 1 and eq. 2:

$$LowLim = \text{Median} - \text{Median} \times \text{variation} \quad (1)$$

$$UpLim = \text{Median} + \text{Median} \times \text{variation} \quad (2)$$

where, the variation of median values influence on the values of the upper limit (*UpLim*) and lower limit (*LowLim*). The median value was used because it is not influenced by the extreme values.

After the global filter, a model is proposed for the removal of points with low consistency of value towards its neighbors. Together with the dataset, two parameters must be provided as input for the model: the range for points around one radius, and the maximum median variation (variation of eq. 1 and eq. 2) acceptable for a grouped range of points. The first is used to define the neighbors located at the radius-range of a point, while the latter is the threshold that determines how much a point is allowed to vary in relation to its neighbors. Anisotropic local filter, neighbor data sets located at a constant radius-range within a single row (red polygon Fig. 1). Based on the assumption that machines run through areas in the boustrophedon pattern (Jin & Tag, 2010), the separation of passes will be done by the modified method of Menegatti and Molin (2004) by identifying the extremes of a path. All data that deviate above or below from the *UpLim* and of the *LowLim* is considered as spatial outlier data and is identified for elimination.

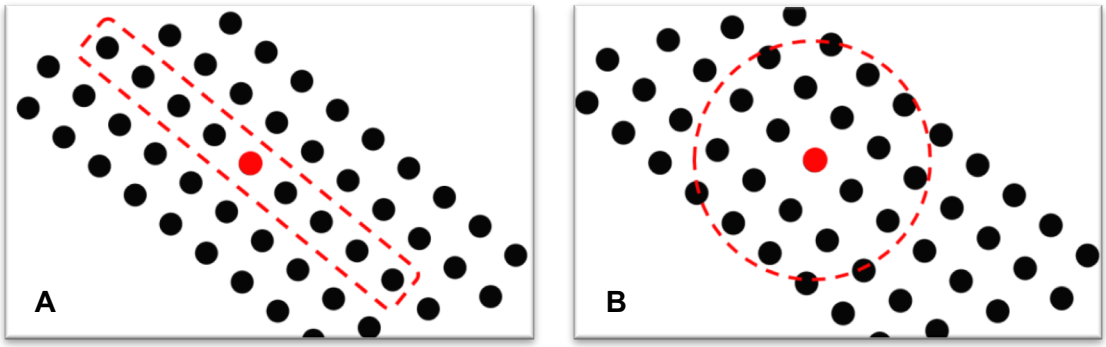

Fig 1. Neighbors located at the radius-range of a point in a single direction (A) and neighbors located by the Spekken et al. (2013) methodology (B).

Finally, spatial data filtering was performed using the methodology described by Spekken et al. (2013). The model assumes that the defined radius does not exceed the spatial dependency of the data in any direction, because the filtering process is isotropic. All process of identifying and removing spatial outlier data was performed using an algorithm created in Java language using NetBeans IDE 8.1 software.

Case study

For the assessment of the three methods we used three datasets from different areas in the western central region of the state of São Paulo, Brazil (22°68'S – 48°40'W). Each area contains soybeans yield data collected at 1.0 Hz. Different types of errors were observed: points with null moisture, harvester fill mode error in headlands, harvester emptying mode error, not fully used cutting bar and points with discrepant yield values (Blackmore and Moore, 1999; Simbahan et al., 2004; Menegatti and Molin, 2004).

Table 1 reports yield statistics for the original data and after filtering data of datasets under consideration. The lower limit of 2.7, 3.7 and 3.3 ton ha^{-1} and the upper limit of 4.9, 5.5 and 4.9 ton ha^{-1} have been applied to fields 1, 2 and 3 respectively. In the anisotropic local filter and in the isotropic local filter we using a radius of 35 m (three and a half times the width of the combine), and a variation of median of the 25%. In the original data set, the median values deviate from the arithmetic mean values, and the minimum and maximum values reinforce the observation of high variability, with high CV.

Table 1. Descriptive statistics and geostatistical analysis of the original soybean yield data (ton ha^{-1}) and after the filtration.

Field	Dataset	Count	Min	Mean	Median	Max	SD	CV	C_0	A	Nugget/Still
					ton ha^{-1}			%		m	
1	Original	71708	0.34	3.85	3.98	13.59	0.95	24.69	0.98	5000.0	0.95
	Filter	48201 (67.2 %)	2.84	3.94	3.95	4.90	0.35	8.86	0.04	138.1	0.54
2	Original	55713	3.50	4.26	4.20	5.60	0.46	10.69	1.05	195.8	0.86
	Filter	44115 (79.2 %)	3.50	4.18	4.17	5.50	0.33	7.86	0.05	59.9	0.58
3	Original	90602	0.34	3.95	4.02	15.78	0.98	24.69	0.08	711.0	0.88
	Filter	51396 (56.7 %)	3.30	4.00	4.00	4.90	0.31	7.76	0.04	338.6	0.56

SD – Standard Deviation; CV – coefficient of variation; C_0 – nugget effect; A – distance.

Process of outlier data removal decreased standard deviation value, increased mean yield in fields 1 and 3, and reduced mean yield in field 2. The proposed model was efficient to identify and exclude points with low and high yield, as can be observed in the change of the original maximum and minimum values and after the data filtering procedure (Table 1). In addition, it is possible observe in parentheses, in the count column, the remained total data after the procedure. There was a considerable reduction in yield amplitude in all fields. A considerable number of data, 32.8%, 20.8% and 43.3%, in fields 1, 2 and 3 respectively, was excluded by applying the three filtering processes. The filtering was able to eliminate all errors found in the three analyzed datasets (Fig 2). Points with null moisture, harvester fill mode error in headlands, harvester emptying mode error, not fully used cutting bar and points with discrepant yield values was excluded.

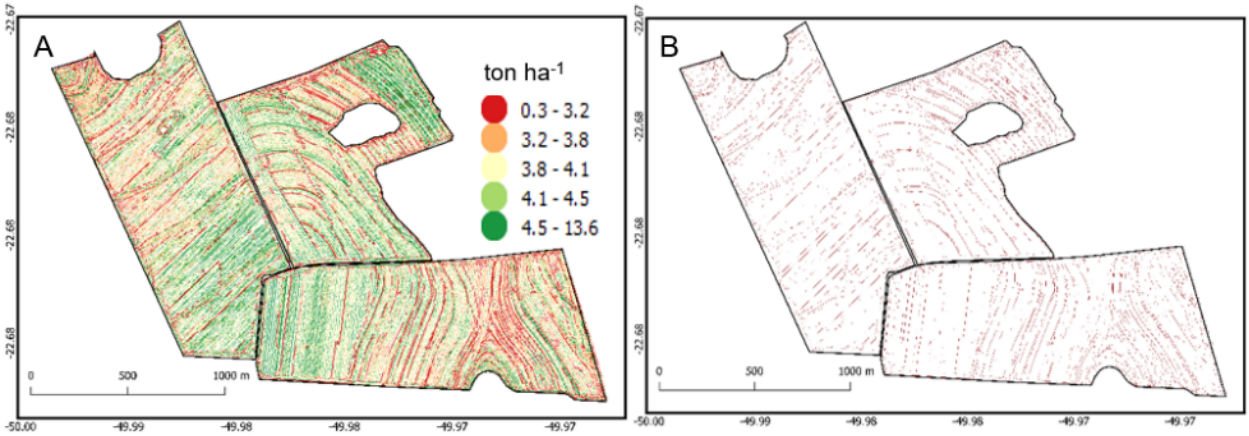

Fig 2. Original soybean yield data (A) and points eliminated by the filtering (B).

These outliers were completely masking the spatial structure of yield. Indeed, semivariograms have demonstrated a moderate spatial dependence of yield with well-defined parameters of nugget and sill, which demonstrate a simple filtering, with the removal of outliers values, can improve the characteristics of yield data sets within the field.

Conclusion

The proposed filtering method for spatial data has increased the efficiency of local filtering by identifying and deleting spatial outlier data and preserving data with consistent values. A considerable number of data was excluded by applying the three filtering processes. In the case of yield data, the method was efficient in identifying and deleting unsuitable spatial data. The filtering was able to eliminate all errors found in the three analyzed datasets.

References

Arslan, S., Colvin, T. S. (2002). Grain yield mapping: yield sensing, yield reconstruction, and errors. *Precision Agriculture* 3, 135–154.

Blackmore, B. S., Moore, M. (1999). Remedial correction of yield map data. *Precision Agriculture.* (1), 53–66.

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*, 1-20.

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.

Ping, J. L., Dobermann, A. (2005). Processing of Yield Map Data. *Precision Agriculture*, 6, 193–212.

Reiser, D., Paraforos, D. S., Khan, M. T., Griepentrog, H. W., & Vázquez-Arellano, M. (2017). Autonomous field navigation, data acquisition and node location in wireless sensor networks. *Precision Agriculture*, 18(3), 279-292.

Shekhar, S. C. Lu, and Zhang, P. (2002). Detecting GraphBased Spatial Outlier. *Intelligent Data Analysis: An International Journal,* 6(5):451–468.

Simbahan, G. C., Dobermann, A., Ping, J. L. (2004). Screening yield monitor data improves grain yield maps. *Agronomy Journal,* 96, 1091–1102.

Spekken, M. A. R. K., Anselmi, A. A., & Molin, J. P. (2013). A simple method for filtering spatial data. *In Precision agriculture'13* (pp. 259-266). Wageningen Academic Publishers, Wageningen.