A simple method for filtering spatial data

M. SPEKKEN [1], A. A. ANSELMI [2], J. P. MOLIN [1]
[1]*Biosystems Engineering Department, University of São Paulo. Piracicaba, SP – Brazil.*
[2]*Crop Science Department, University of São Paulo. Piracicaba, SP – Brazil.*
mspekken@usp.br

Abstract

Sensors applied on agricultural fields collect large amounts of spatial data needed for intervention and decision making, but this may come with a considerable quantity of defective data.The aim of this study was to develop 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. This method was implemented in an algorithm where case studies were inserted. For filtering yield data,while comparing with filter procedures using upper and lower limits, the proposed method was effective in excluding inconsistent points of their neighbors and identified different types of errors as productivity null, wrong set of platform width, and lag/fill modes in headlands. The filters also showed capable of reducing noise in output maps and show potential to smooth boundaries of cluster areas and retrieve higher uniformity within this. Despite the simplicity of parameters in the method, these must still require some calibration for usage.

Keywords: Filter maps, erroneous data, modeling,

Introduction

The development of positioning technologies along with sensors have narrowed the spatial resolution and increased the amount of data collected from farm-fields, which are usually organized in the form of digital maps. Also more files are being created with the growing number of fields that are being scanned with the use of these sensors in the precision agriculture trend,which allows the generation of maps required for interference with the production system or in strategic decision making.

Many sensors are now attached to farm machinery retrieving spatial data, but, while producing a digital map with the data collected from these sensors, a number of errors may occur. 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.

Removal of these errors has been studied by authors using methods that applied sequences of filters to remove defective data (Ping and Dobermann, 2005., Simbahanet et al., 2004., Menegatti and Molin, 2004., Arslan and Colvin, 2002., Blackmore and Moore, 1999). Some of the filters require prior knowledge of the target factor (like crop yield, vegetation index, soil electrical conductivity, etc) for establishing upper and lower thresholds to identify the erroneous data, but data removed outside these boundaries were the biggest cause of loss of good data (Blackmore and Moore, 1999).

Using boundary parameters to classify an entire heterogeneous dataset will often lead to removal of undesired points. But, in such datasets, homogeneity can be found in local regions which demand local analysis. Initiatives like the software VEPSER 1.5 deals with such limitations allowing user defined neighborhood and prediction-block sizes (Whelan, 2002).

Other approaches have been developed to identify specific errors from yield maps considering the values of neighboring points. Thylén et al. (2000) developed some filters that detects and removes erroneous yield data resulted from reduced cutting widths and rapid speed changes by using the value of a point relative to the average and variance of its neighbors, but still requires upper and lower limits. Noack et al. (2003) used a logical path recognition combined with the moving average productivity to automatically establish the limits acceptable to a certain point evaluated.

With a higher number of filtering procedures and parameters required to clean heterogeneous sets of data, as well as the necessity of distinct human investigation for each, a considerable time and/or energy can be consumed in these process. Additionally it is hard to establish patterns for comparing a series of historical data using the same filtering settings once it is influenced by each map producer.

This work proposes a method capable of identifying a varying point in the middle of a delimited group of points using as statistical parameter the coefficient of variation (CV) of the target-attribute. A minimum user given CV is the threshold required for classifying the data unit, which is flexible to classify any sort of data and.

Material and methods

Overview of the filtering method

From a spatial dataset composed by points, a model is proposed for the removal of points that has low consistency of value towards its neighbors. These value is the target attribute that is submitted to filtering (for e.g. yield, electrical conductivity, NDVI, etc), for which the location is defined by latitude and longitude coordinates.

Together with the dataset, two parameters must be provided as input for the model: the range for points around one *Radius*, and the maximum coefficient of variation (*MaxCV*) 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.

The model assumes that the defined radius doesn’t exceed the spatial dependency of the data in any direction, because the filtering process is isotropic.

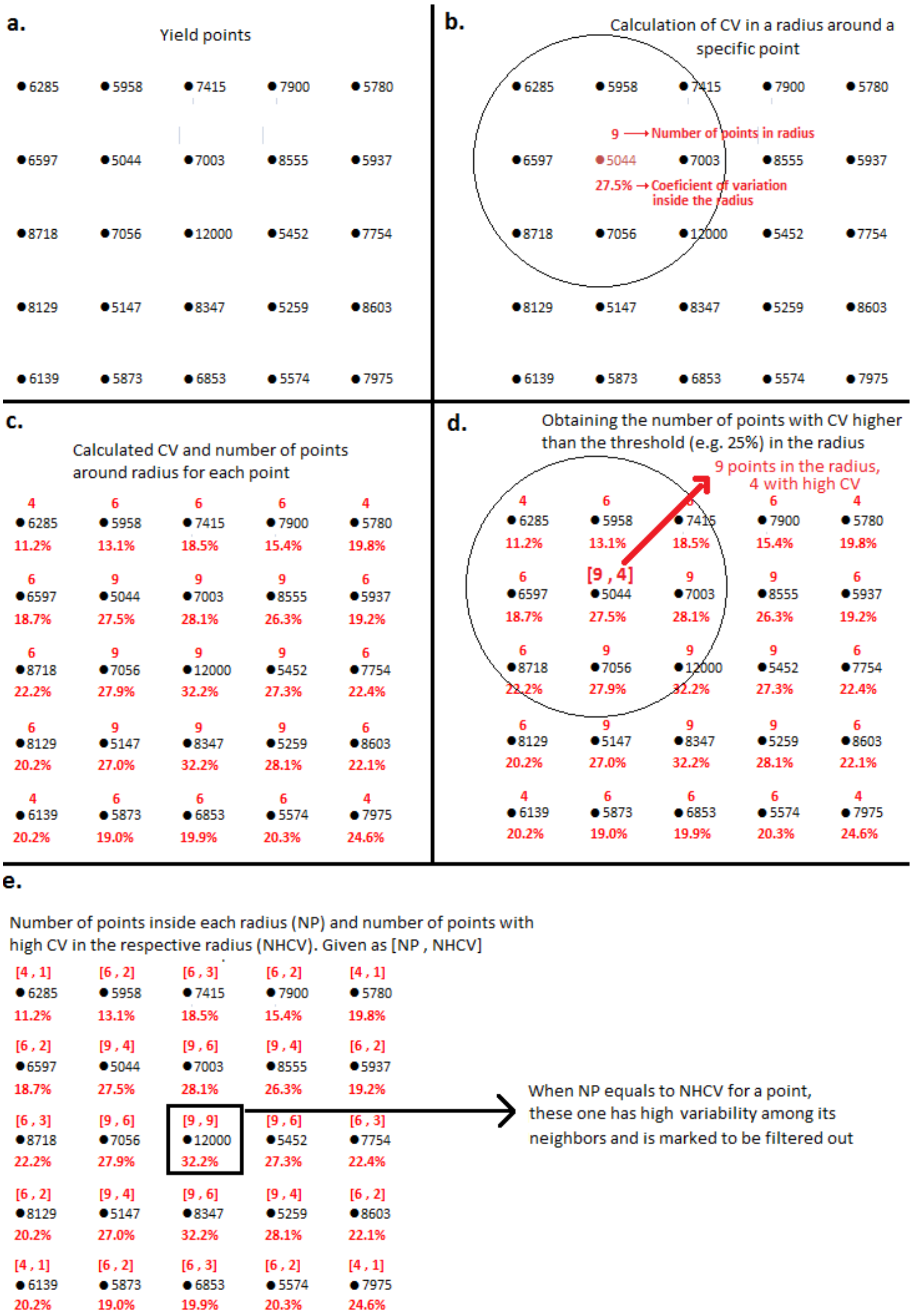
In summary, the method proposes:

- Detecting all points located in a radius range around a point;
- Extracting the CV among these points;
- If the CV value found for these points is higher than the *MaxCV*, a weight is added to all these;
- A next point following point in the dataset is selected and the process repeats;

After this process is finished for all points, the outlier points have a high summed weight and must be therefore discarded.

The filtering methods illustrated in a fictitious dataset in Figure 1, employing a *MaxCV* threshold of 25%. The Figure shows a sequence of 5 steps in the model implementation to identify the outliers; these are described as following:

a. A given spatial dataset is provided with a target attribute (in this example yield in kg). Observe that one point in the center shows a spiked productivity (12.000 kg), which is inconsistent with its neighbors;
b. Considering one point of the dataset as example (illustrated as the red point of value “5044” in Figure 1b), all the surrounding points within the radius-range are identified as neighbors. The number of neighbors (andthe central point itself) is countedand the coefficient of variation (CV) is calculated among them, being both values (the count and the CV) stored as attributes of the central point. The red labels above and under thepoint in Figure 1b illustrate it.
c. The previous procedure is done for all points in the dataset,counting the number of points and calculating the CV (Figure 1c). Observe that, if the filtering would take place at this step, a large number of points would have a CV higher than the *MaxCV*, eliminating a considerable amount of good data.
d. Considering again thepoint of step “b”, the model now countshow many points inside the radius have their stored CV higher than *MaxCV*. This countis also stored as attribute of the central point (in these example, there are 4 points with high CV among 9).
e. The process “d” is repeated for all the points. The point where he and all its neighbors have high CV, is the point that causes the variability in theirmidst, and is marked to be filtered out. Figure 1e display this phenomenon where NP is the number of points in the radius and NHCV is the number of points with high CV inside the same radius, by finding NP equal to NHCV the outlier is identified.


a.
Yield points
b.
Calculation of CV in a radius around a specific point
9 → Number of points in radius
27.5% → Coeficient of variation inside the radius
c.
Calculated CV and number of points around radius for each point
d.
Obtaining the number of points with CV higher than the threshold (e.g. 25%) in the radius
9 points in the radius, 4 with high CV
[9 , 4]
e.
Number of points inside each radius (NP) and number of points with high CV in the respective radius (NHCV). Given as [NP , NHCV]
When NP equals to NHCV for a point, these one has high variability among its neighbors and is marked to be filtered out

[4 , 1] ● 6285 11.2%	[6 , 2] ● 5958 13.1%	[6 , 3] ● 7415 18.5%	[6 , 2] ● 7900 15.4%	[4 , 1] ● 5780 19.8%
[6 , 2] ● 6597 18.7%	[9 , 4] ● 5044 27.5%	[9 , 6] ● 7003 28.1%	[9 , 4] ● 8555 26.3%	[6 , 2] ● 5937 19.2%
[6 , 3] ● 8718 22.2%	[9 , 6] ● 7056 27.9%	[9 , 9] ● 12000 32.2%	[9 , 6] ● 5452 27.3%	[6 , 3] ● 7754 22.4%
[6 , 2] ● 8129 20.2%	[9 , 4] ● 5147 27.0%	[9 , 6] ● 8347 32.2%	[9 , 4] ● 5259 28.1%	[6 , 2] ● 8603 22.1%
[4 , 1] ● 6139 20.2%	[6 , 2] ● 5873 19.0%	[6 , 3] ● 6853 19.9%	[6 , 2] ● 5574 20.3%	[4 , 1] ● 7975 24.6%

Figure 1.Stepwise procedures of the filtering process.

Implementation of the model

An algorithm-application was built in the software Lazarus 1.0 (free Pascal Lazarus Project) implementing the model. It reads text files that must contain at least three numeric attributes being:

- Two attributes containing the latitude and longitude in the WGS 84 datum provided in geographic coordinates (decimal degrees), which is common form for storage of coordinates in agricultural data loggers.
- The target attribute that is subjected to filtering;

The original coordinates are converted into UTM coordinates, allowing the points to be analyzed in a regular metric 2D plane and for calculating distances between them.The points are stored in a matrix of records composed by 6 attributes: UTM easting, UTM northing, target attribute, number of points within radius (NP), coefficient of variation (CV) and number of points with high CV (NHCV) within itsradius. The first three attributes are filled with data from the file.

The user given parameters are the *Radius* (given in meters), and the *MaxCV* (given in percentage).The algorithm loops along all the points finding the ones with distance lower than *Radius* (defining NP) and afterwards calculating the CV among these. Both the NP and the CV are stored in the matrix during this loop.A second loop repeats previous procedure counting,within the radius, those with CV higher than *MaxCV* for each central point, and storing the counted value in matrix as NHCV.

The output of the model is an exported list of points in a file where the NP is lower than the NHCV, and the filtered data can proceed to further processing.

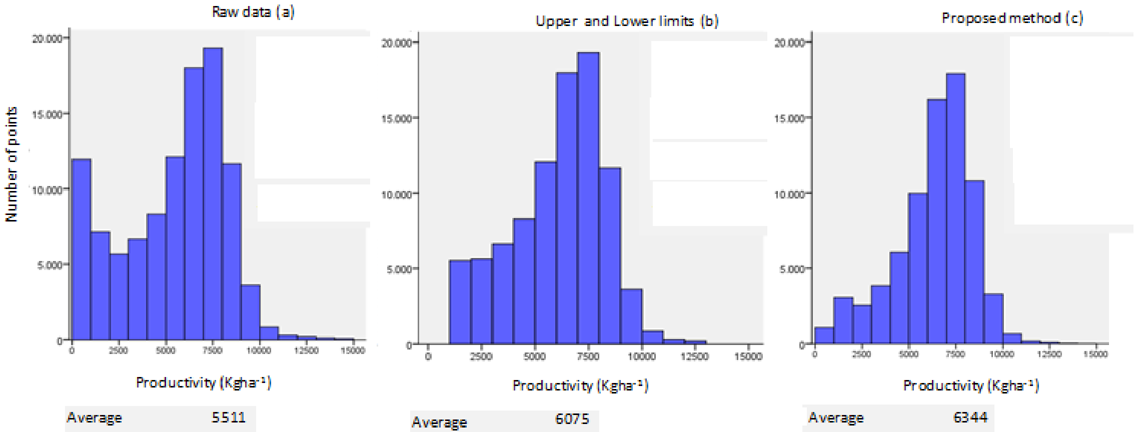
Results and discussion

Yield data case study

A raw dataset of corn yield (106540 points collected at 1 Hz) was applied into the algorithm using a radius of 10m (one and a half times the width of the combine), and a CV of 25%.The average number of points within the moving radius in the whole dataset was 17.

A comparison filtering process was applied removing data from upper and lower limits (respectively 13000 kg ha^{-1} and 1200 kg ha^{-1}). For this latter method steps used were: exclude points outside the field (Blackmore and Moore, 1999., Menegatti and Molin, 2004), points with null moisture (Simbahanet et al., 2004., Menegatti and Molin, 2004) and points with discrepant yield values from upper and lower limits (Blackmore and Moore, 1999., Simbahan et al., 2004,Menegatti and Molin, 2004).

The proposed method eliminated 29% of the original points against 13,6% of the comparison filter. In other studies using steps forfiltering the amount of screened points was between 13-20% of total (Ping and Dobermann, 2005, Simbahan et al., 2004).The resulting average yield increased by 15% and SD decreased by 43% with proposed filter compared to the raw data.The descriptive statistics of the data filtered by these two methods is shown in Figure 2.



Average	5511
Min.	0
Max.	99699
Range	99699
Total n° of points	106540

Average	6075
Min.	1200
Max.	12992
Range	11792
Total n° of points	91970

Average	6344
Min.	59
Max.	17426
Range	17366
Total n° of ponints	75628

Figure 2. Distribution of yield values form: the raw data (a) the comparison filter (b) and the proposed method (c).

The raw data shows a large number of points with null yield. These points are often associated with fill mode and lag time, near the headland and carrier paths (Blackmore and Moore, 1999; Arslan and Colvin, 2002). The proposed model was efficient to identify and delete such points with low productivity. The range of the filtered data of the proposed method was 17366 kg ha^{-1}.The method of filtering limits was also able to identify points with extreme yields, but excluding all points outside that range included also non defective points, at the same time it failed to eliminate erroneous points within the considered interval.

A subset of the data filtered by the different methods is shown in Figure 3.In the ellipses the proposed method (a) was able to keep a significant number of points with very low productivity that would otherwise be selected for elimination in (b). The data in this area represent a low-yield zone that creates difficulty in defining the limits for exclusion.

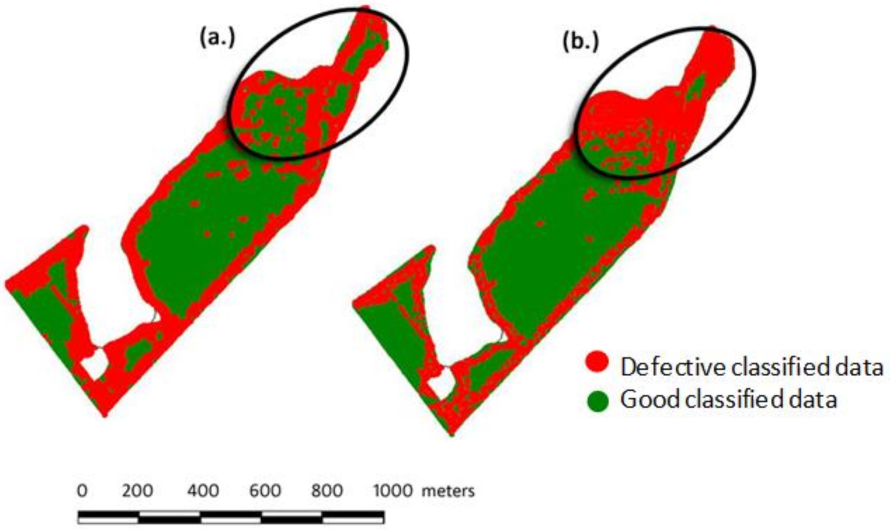


Figure 3. Subset of the case study area showing the methodsto select defective points: filtered by proposed method (a) and filtered by upper and lower limits (b).

Once yield monitors sense changes in grain yield (Arslan and Colvin, 2002) other types of errors that reflect productivity can be identified indirectly through the proposed method, as the crop width entering the header and maneuver points. The first type of error is often characterized by long strips of indicated low yield, running along the length of the field, where the harvester finished working (Blackmore and Moore, 1999), while the maneuver can also be characterized by low yield at the end of each pass. The proposed method identified points which were possibly collected with a width set incorrectly or because of underutilized headers width. Also, the method seemed efficient in identifying erroneous points around carriers.

Interpolated maps (inverse of distance) comparing both methods can be observed in Figure 4. In general, the proposed method was efficient in reducing the noise. Blackmore and Moore (1999) suggested that for the interpretation of the maps and subsequent use for localized management practices it is necessary a certain degree of smoothing of the data.

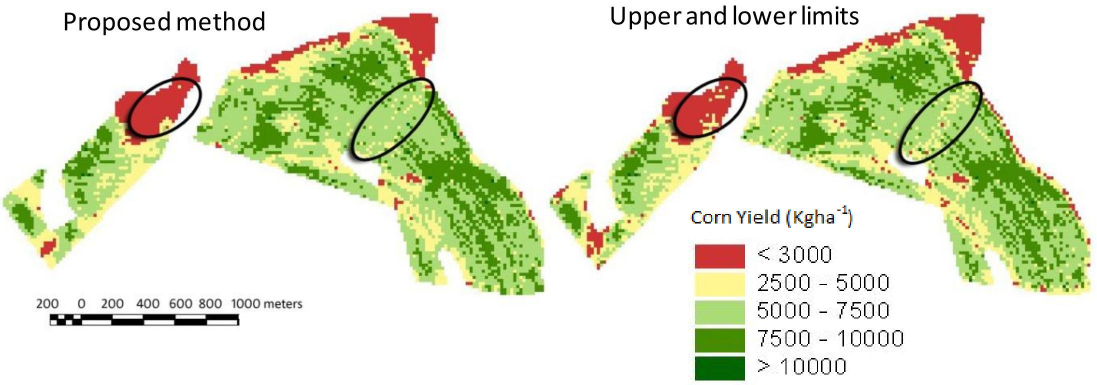


Figure 4.Interpolated maps of the case study area for the two methods of filtering.

The authors also observed that increasing the intensity of filtering (by using low CV base values), that besides removing the errors, it is capable to eliminate small variations on the fields resulting in areas with higher uniformity, suggesting its use to ease the formation of clusters. But these clusters should mainly represent the stable site yield potential delineating large and spatially contiguous areas within a field (Ping and Dobermann, 2005). But such study in beyond the scope of this work.

Considerations about the model

A second case study was inserted in the algorithm to observe the sensitivity of the model to the filtering parameters. A dataset of georeferenced NDVI values containing 46287 points (collected in a frequency of 10Hz) was analyzed under different CV and Radius thresholds.

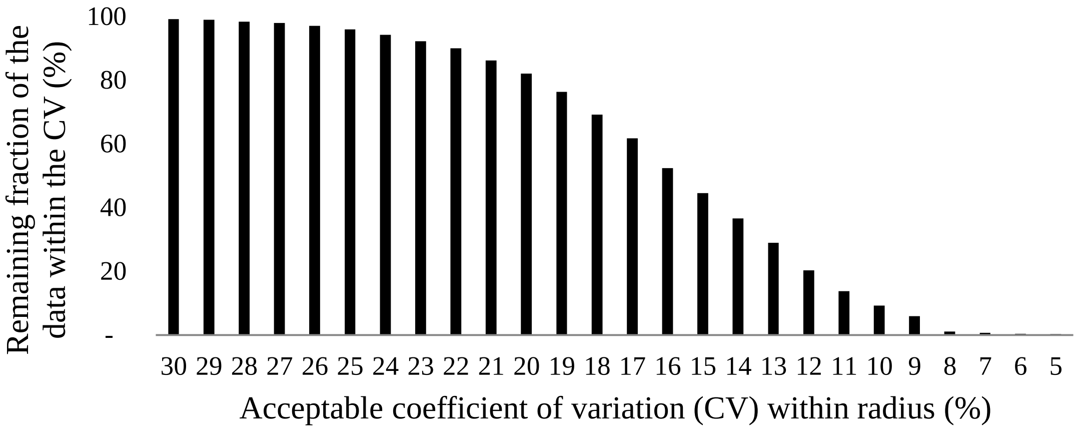


Figure 5.Decreasing quantity of remaining data for tighter CV thresholds.

As shown in Figure 5, defining a suitable CV still poses a problem that requires further study. The current experience with the algorithm suggests that a range of CV between 10 to 25% is capable of eliminating most of the defective data in yield maps. But depending on the purpose the user may desire to have just a few points of high consistency to achieve smoother zones in interpolated maps.

The definition of a radius is also not finally established; it must consider the range of dependency of data. Herein it is assumed that one and a half times the width between passes (as in the yield case study) suffices.

Figure 6 shows the filtering impact of two different radius-ranges in the NDVI dataset.

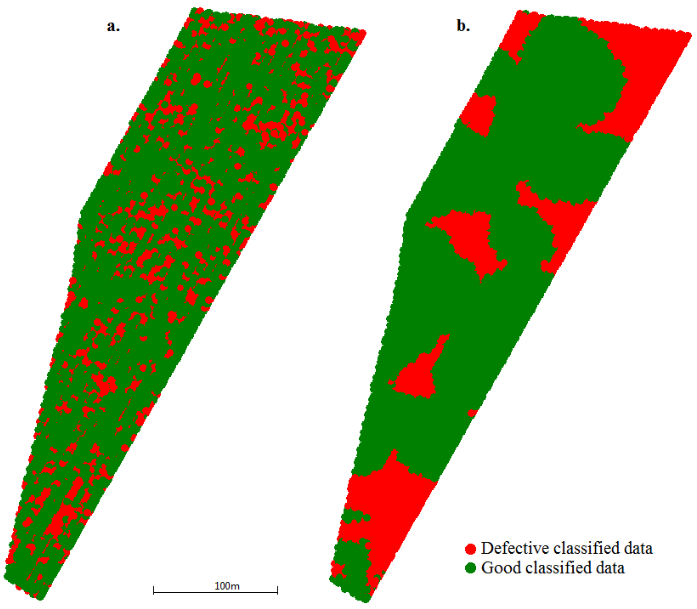


Figure 6. Data classification for two distinct radius-ranges: 2m (a) and 20m (b). Both studies used a CV of 20%.

In Figure 6a the radius was defined by a length slightly higher to the width between the passes (1.5 m), detecting defective data spread along the map. In Figure 6b real good data located in a wider neighborhood containing significant quantity of defective data was aggregated in defective zones, while saving real defective data in good zones. Surprisingly, the number of points classified to be removed was similar (12495 and 12991 respectively in ‘a’ and ‘b’).

Conclusions

A proposed filtering method for spatial data is introduced, which allows identifying and deleting the points causing variation within a given set of neighboring points and preserving points with consistent values showing potential for improving the quality of the map. The method allows many sorts of numeric data associated with a geographic coordinate to be filtered requiring no specific prior knowledge of the target data and using only two input parameters for classification: a Radius that defines a range of points around a location and a maximum acceptable CV of these.

In the case of yield data analysis the method was efficient in identifying and deleting unsuitable points resulting of fill mode and lag time, points near the headlands, carriers and points with erroneous set width, being that all directly impact on the analyzed attribute and represents the majority of errors in yield maps.

The noises on the map interpolated were reduced and there was improvement in smoothing. Therefore, it is suggested that it may facilitate the definition management zones, although further studies are needed for this.

Definitions of specific thresholds and parameters to be used are not yet defined, but some pointed suggestive values are a guide for users to define their own thresholds.

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.

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, Campina Grande, v.8, n.1, p.126-134.

Noack, P. H., Muhr, T., Demmel, M. 2003. An algorithm for automatic detection and elimination of defective yield data. In: Precision agriculture. Proceedings of the 4th European Conference in Precision Agriculture, edited by J. V. Stafford and A. Werner (Wageningen Academic Publishers, Wageningen, The Netherlands), p. 445–450.

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

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

Thylén L.; Algerbo P.A.; Giebel A. 200. An expert filter removing erroneous yield data In: Proceedings of the 5thInternational Conference on Precision Agriculture, edited by P. C. Robert; R. H. Rust and W. E. Larson (ASA, CSSA, and SSSA, Madison, WI).

Whelan, B. M., A. B. McBratney, and B. Minasny. "Vesper 1.5–spatial prediction software for precision agriculture." *Precision Agriculture, Proc. 6th Int. Conf. on Precision Agriculture, ASA/CSSA/SSSA, Madison, WI, USA*. 2002.