A general method to filter out defective spatial observations from yield mapping datasets

Article *in* Precision Agriculture · January 2018

DOI: 10.1007/s11119-017-9555-0

A general method to filter out defective spatial observations from yield mapping datasets

Leroux, Corentin (1-2), Jones, Hazaël (2), Clenet, Anthony (1), Dreux, Benoit (3), Becu, Maxime (3), Tisseyre, Bruno (2)

(1) SMAG, Montpellier, France
(2) UMR ITAP, Montpellier SupAgro, Irstea, France
(3) DEFISOL, Evreux, France
cleroux@smag-group.com

Abstract

Yield maps are recognized as a valuable tool with regard to managing upcoming crop production but can contain a large amount of defective data that might result in misleading decisions. These anomalies must be removed before further processing to ensure the quality of future decisions. This paper proposes a new holistic methodology to filter out defective observations likely to be present in yield datasets. The notion of spatial neighbourhood has been refined to embrace the specific characteristics of such on-the-go vehicle based datasets. Observations are compared with their newly-defined spatial neighbourhood and the most abnormal ones are classified as defective observations based on a density-based clustering algorithm. The approach was conceived to be as non-parametric and automated as far as possible to pre-process a growing number of datasets without supervision. The proposed approach showed promising results on real yield datasets with the detection of well-known sources of errors such as filling and emptying times, speed changes and non-fully used cutting bar.

Keywords: DBSCAN algorithm, filtering, local outliers, on-board sensors, spatial neighbourhood, yield

Introduction

Yield maps have been extensively recognized as a valuable source of information for field decision making (Diker et al. 2004; Florin et al. 2009; Pringle et al. 2003). They effectively provide a global overview of the field spatial variability which makes it interesting to target areas or zones for variable rate management. As a combine harvester passes through a field, yield monitors acquire almost in real-time multiple yield measurements all over the field. At the same time, those data are associated with the GNSS positioning of the machinery which enables precise location of each one of these observations at the within-field level. As such, thousands of yield spatial observations are generated and are ready to be used in the decision-making process. While this considerable volume of data is critical for field management and decision-making, these datasets must be used with great caution. They effectively contain lots of defective observations or technical errors that need to be removed to ensure data quality (Arslan and Colvin, 2002; Blackmore and Moore, 1999). As a consequence, yield datasets are often severely filtered to make sure further analyses are not flawed (Robinson and Metternicht, 2005; Sudduth and Dummond, 2007; Sun et al. 2013). Several authors have described to what extent a yield map could evolve after removing abnormal values (Simbahan et al. 2004; Sudduth and Dummond, 2007). Griffin et al. (2008) have even shown that these latter observations were able to influence field management decisions.

These technical errors or defective observations have been largely documented in the literature. Lyle et al. (2013) have proposed a categorization of those latter errors into four major groups: (i) harvesting dynamics of the combine harvester, (ii) continuous measurements of yield and moisture, (iii) accuracy of the positioning system and, (iv) harvester operator. These technical errors are briefly described hereafter, in the previously defined order, along with methodologies that have been proposed by the scientific community to identify these defective observations.

- The harvesting dynamics of the machine include three different offsets, referred to as the lag time, filling time and emptying time (Blackmore and Moore, 1999). The lag time induces an offset between the actual and the true location in space of a yield observation because the yield is not measured simultaneously with the cutting of the crop. Some attempts have been made to determine this offset through (i) geostatistical methods (Chung et al., 2002), (ii) image processing techniques (Lee et al. 2012) and (iii) signal deconvolution (Arslan, 2008;

Reinke et al. 2011). The filling time at the start of a harvest pass leads to an under-estimation of the yield because the grain flow is increasing and still has not reached a plateau, i.e. the permanent regime. Therefore, yield measurements do not match the expected true yield values. At the end of a harvest pass, some grain might still continue to flow after the last crop was harvested and the lag time has been reached. As a consequence, the latest observations of a harvest pass are generally under-estimated. The methods that have been proposed so far are exclusively visual, i.e. the grain flow is plotted against the travel time or distance of the machine and the data located before or after the plateau are removed (Lyle et al. 2013; Simbahan et al. 2004).

- Continuous measurements relate to yield and moisture observations. So far, studies have focused on thresholds, mostly determined empirically, to identify measurement errors (Sudduth and Drummond, 2007; Taylor et al. 2007). Arslan and Colvin (2002) have reported sensor accuracies varying between 1 and 4% while other authors have found differences up to 10% depending on environmental conditions during data acquisition, e.g. steep slopes (Reitz and Kutzback, 1996). To overcome that issue, a couple of studies have focused on the impact of the combine harvester vibrations on the yield measurement accuracy (Hu et al. 2012; Jingtao and Shuhui, 2010).

- The accuracy of the positioning systems can lead to (i) observations outside field boundaries, (ii) measurements at the same spatial location, i.e. co-located points, or (iii) deviations in space according to a predefined harvest pass (Blackmore and Moore, 1999). The two first types of errors are easily handled by removing the points outside the boundaries of the field or points with similar co-ordinates (Robinson and Metternicht, 2005; Simbahan et al. 2004). Some algorithms have been implemented to reconstruct precisely the harvest passes by studying the angles formed by consecutive points (Lyle et al., 2013). Suspicious points – those the combine harvester is not likely to have gone through – are removed from the dataset.

- Last type of errors has to do with the harvester operator. First, large variations in speed are likely to have a major impact on the yield dataset quality (Arslan and Colvin, 2002; Sudduth and Drumond, 2007). Speed issues are generally processed the same way as yield and moisture, i.e. by setting thresholds to the whole dataset or only to neighbouring data (Lyle et al. 2013). The harvester operator is also likely to overlap consecutive or adjacent harvest passes which may result in yield measurement errors. Some authors have focused on this 'not fully used cutting bar' effect and have come up with vector-based pre-processing methods to take into account these overlaps, mainly by reconstructing harvesting polygons (Drummond et al., 1999). These vector-based methods are heavily dependent on the positioning accuracy of the GNSS device and require a large processing time. Other authors have proposed specific on-board systems, such as those based on ultrasonic sensors (Zhao et al. 2010). Finally, harvest turns and headlands are also responsible for bad yield estimates (Lyle et al. 2013). Studies dedicated to these last sources of errors – though limited in the literature – have focused on finding the points inside harvest turns or headlands by using distance or angle measures between consecutive points. Suspicious points are removed.

On-board sensors such as yield monitors generate an extremely large amount of observations. This considerable volume of observations requires the filtering approaches to be at the same time automated, very general and non-parametric (Simbahan et al. 2004; Spekken et al. 2013). The automation condition is fundamental with regard to the increasing size and number of yield datasets to process. For instance, it would not be conceivable for an operator or advisor to spend time on the correction of hundreds of possible within-field yield maps. General and non-parametric detection methods are also to be preferred because of the diversity of datasets that have to be processed. These datasets are effectively acquired through a variety of acquisition systems – machines, sensors – and on multiple crops, with different operators and under varying conditions of acquisition, e.g. topography or climate. It is therefore important to make sure that the approaches are able to deliver conclusive results whatever the dataset to be analysed. Even though new operating systems exist to improve the quality of yield datasets, e.g. ultrasonic sensors (Zhao et al. 2010), it can be argued that all the actual combine harvesters are far from being equipped with it. General methods are therefore also required to process datasets arising from multiple types of machines, whatever the level of additional equipment installed. It must be kept in mind that agronomic datasets are often included in complex processes of field management and decision-making, and are sometimes used as inputs in agronomic models. Data filtering methods have therefore to be robust enough so that the decision-making process is accurate and not flawed. A limitation of the actual literature is that most of the existing approaches are semi-automatic and rely on expert thresholds and filters. These last aspects might be problematical for the processing of yield maps at a larger scale as filtering settings can be influenced by each map producer and as skilled operators might be required for a considerable amount of time (Spekken et al. 2013).

The principal contribution of this work is to propose a new holistic data-driven method to filter out defective observations from on-the-go yield datasets. To the best of the author's knowledge, very few general or holistic data filtering approaches have been dedicated to within-field yield datasets. The methodology is firstly formalised and described to set all the concepts and definitions related to the removal of defective observations in yield datasets. Then, an implementation of the methodology is proposed with an emphasis on the approach to be as automated and non-parametric as possible. Finally, the approach is tested on real datasets obtained from grain flow sensors mounted on combine harvesters.

On-the-go vehicle based datasets and spatial outlier detection

Acquiring observations with on-board sensors

In agriculture, data acquisition with on-board sensors can be understood as a sequential procedure through time during which a machine acquires information of a variable Z in space. Indeed, the data collection process follows a temporal dynamic, i.e. observations are recorded in a specific order and one at a time as the machine passes through the field (Fig. 1). The machine can simply be modelled by a structuring element that moves through the field, e.g. a rectangle whose dimensions are defined by the characteristics of the machine and the associated on-board sensors. On-the-go measurements are punctual observations, i.e. diverse realisations of Z, and each point synthesizes the response of Z over the corresponding structuring element. The spatial resolution of the sensed variable is controlled by the distance between consecutive records and determined by the distance between adjacent passes of the machine. The spatial distance between consecutive observations is related to the speed of the machine and the sampling frequency. In a given field, this frequency of acquisition is generally stable which means that the distance between consecutive records only relies on the travel speed of the machine. On the other hand, the distance between adjacent passes depends on multiple parameters such as the work of the machine, the crop being sensed, or the cost of data acquisition among others. For instance, when a combine harvester with an on-board grain yield monitor passes through a field, the distance between adjacent passes is related to the width of the cutting bar because the whole field has to be harvested.

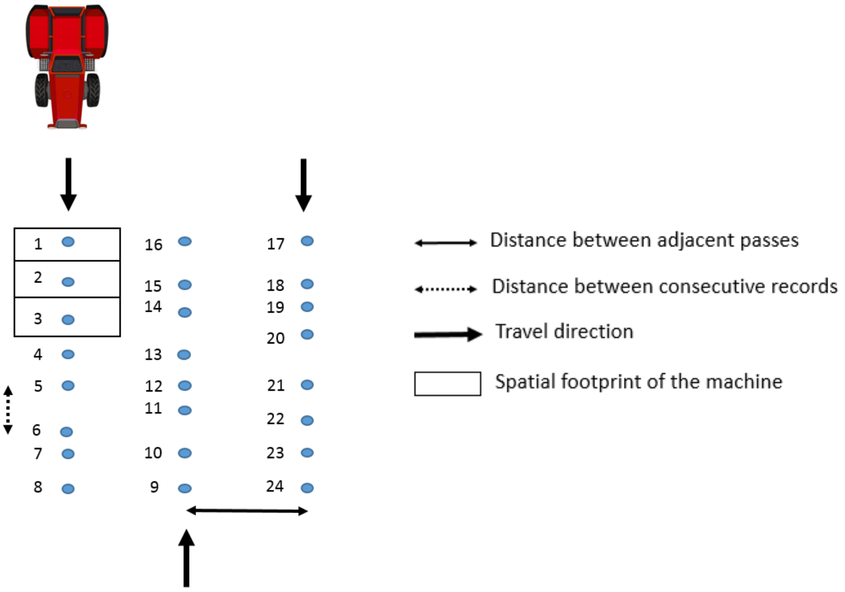
1
2
3
4
5
6
7
8
16
15
14
13
12
11
10
9
17
18
19
20
21
22
23
24
Distance between adjacent passes
Distance between consecutive records
Travel direction
Spatial footprint of the machine

Fig. 1. Principle of data acquisition with on-board sensors.

According to Tobler's first law of geography, everything is related to everything else, but near things are more related than distant things (Tobler, 1970). This concept assumes that there exists some spatial correlation between spatially close observations, to a greater or lesser extent. Multiple studies have shown that this spatial dependency has been clearly exhibited by yield datasets (Pringle et al. 2003; Simbahan et al. 2004). The presence of this spatial correlation is a central feature of the proposed filtering methodology.

Spatial outlier detection

The proposed approach will aim at removing the observations that are the cause of strong local variations of Z(x) which might mask the true spatial correlation between neighbouring points. This approach can therefore be seen as a spatial outlier detection problem. Outlier detection is one of the major areas of investigation of the data mining community and has extended to numerous applications such as fraud detection, traffic networks or military monitoring (Ben-Gal, 2005; Gogoi et al. 2011). Hawkins (1980) has proposed a formal definition of an outlier which states that it can be described as an observation that deviates so much from the rest of the observations as to arouse suspicions that it was generated by a different mechanism. When observations are located in space, their spatial attributes, i.e. co-ordinates, can be used to define a spatial neighbourhood, known as a group of observations that are relatively close in space. A spatial outlier can then be defined as an observation whose non-spatial attributes behave differently to those of other observations in its spatial neighbourhood. From these two definitions arises the distinction between global and local outliers (Chen et al., 2008). Indeed, spatial outliers are only investigated in a spatial neighbourhood, meaning that the non-spatial attributes of outliers do not necessarily deviate from the entire dataset. On the contrary, the definition of an outlier proposed by Hawkins (1980) assumes a specific behaviour of an observation with regard to the whole dataset.

Spatial outlier detection has gained much interest with the increasing amount of spatial observations available. Although many more algorithms have been proposed to deal with traditional outliers, i.e. observations with no reference in space, several methods have been specifically addressed to the detection of spatial outliers (Chen et al. 2008; Filzmoser et al. 2014; Harris et al. 2014; Lu et al. 2003). These approaches generally involve three major steps. First, for each observation x_i, a spatial neighbourhood $N(x_i)$ needs to be associated with each observation. To do so, the user can either define a spatial distance beyond which observations are no longer part of the spatial neighbourhood or select the number of k spatially close observations that belong to the spatial neighbourhood of each observation (k nearest neighbours). The next step in spatial outlier detection is the computation of a metric to quantify the difference between the non-spatial attributes of each observation and those of its spatial neighbourhood. This problem has been well formalized by Lu et al. (2003). Let f_A be an attribute function such that $f_A(x_i)$ is the value of the attribute A of x_i. Let g_A be an attribute function such that $g_A(x_i)$ is a summary statistic of the attribute A of the observations belonging to $N(x_i)$. A comparison function h_A can then be defined as a function of f_A and g_A to measure the 'outlierness' of each observation x_i with regard to $N(x_i)$. The 'outlierness' reports to what extent a given observation can be considered an outlier. A high indicator of 'outlierness' means that an observation is likely to be of low quality and as such can be regarded as a defective observation. As an example, Lu et al. (2003) have proposed a function g_A that returns the median of the attribute A of all the observations inside $N(x_i)$ and h_A was defined as f_A - g_A. Finally, the observations are directly classified as outliers or normal observations (Chen et al., 2008), or at least they are ordered from the most to the least suspicious observation (Filzmoser et al. 2014; Lu et al. 2003). In the last case, a threshold has to be manually selected to separate the outliers from the non-outliers.

The definition of outliers in on-the-go vehicle-based datasets such as yield datasets has not been stated so far and there is a need to be more specific about it. Observations can be considered as outliers if they are significantly different from their neighbouring observations. From a general perspective, outliers are removed from the datasets because they can negatively impact the quality of the entire population of observations. These outliers are often the result of a sensor error or a very particular and isolated phenomenon, e.g. game damage. However, in the case of sensors embedded on mobile machines, some outliers arise from the machine pass in itself, i.e. from the data collection process. These types of observations are different from their neighbouring observations, not because they are abnormal but rather because these observations were acquired under a specific acquisition process. For instance, when the cutting bar of a combine harvester is not fully used during a machine pass, yield observations are under-estimated because the grain flow is weighed over a harvest area that is bigger than it should be (Arslan and Colvin, 2002; Lyle et al. 2013).

Material and methods

A new data filtering algorithm dedicated to on-board sensor measurements

A specific neighbourhood for each observation

Spatial neighbours are observations that are relatively close to each other in the space domain. When acquiring observations with on-board sensors, the data collection process follows the passes of the machine. This means that spatially close observations might have been acquired (i) during a short time interval, i.e. these observations belong at least to the same machine pass, or (ii) at different time periods, i.e. they belong to different passes. Given the varying machine dynamics through the passes, spatially close observations in the same pass do not necessarily have the same characteristics as spatially close observations in adjacent passes. In fact, it is reasonable to assume that the data collection process induces in itself an anisotropic phenomenon in the direction of the machine pass, i.e. between observations that belong to the same pass. This phenomenon should be taken into account separately in the definition of the neighbourhood for each observation. As a consequence, the proposed approach attempts to remove the observations that are the cause of strong local variations of Z which might mask the true correlations between spatially close observations (i) in the same pass of the machine and (ii) in different passes of the machine.

More formally, the spatial neighbourhood *N*(x_i) of an observation x_i can be separated into two different neighbourhoods: a spatio-temporal and a spatio-not-temporal neighbourhood. The spatio-temporal neighbours of x_i are the spatial neighbours that are, at the same time, near in space and time to x_i. An observation x_i and its spatio-temporal neighbours are acquired in a short time interval. Spatio-not-temporal neighbours are near observations in the space domain but not in the time domain. From now on, spatio-temporal and spatio-not-temporal neighbours will be referred to as ST and SNT neighbours. Hence, for each observation x_i, the spatial neighbourhood *N*(x_i) is divided into *ST*(x_i) and *SNT*(x_i). An example is given in Figure 2. Three passes are travelled in opposite directions. Observation 13 has ST neighbours (observations 10, 11, 12, 14 and 15 for example) and SNT neighbours (observations 2 to 7 and 18 to 23, for instance). The number of ST and SNT neighbours depends on the size of the neighbourhood. Note that the use of the two neighbourhoods makes possible a distinction between the specific machine dynamics inside the same pass and those in different passes of the machine.

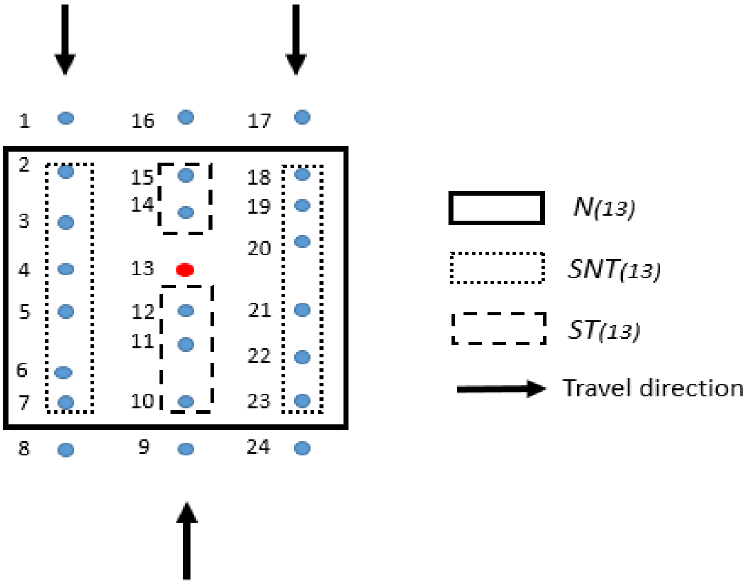

Fig. 2. ST and SNT neighbourhoods of an observation. *Each observation x_i has a ST(x_i) neighbourhood (observations are acquired in a short time interval) and a SNT(x_i) neighbourhood (observations belong to different passes).*

Given the spatial footprint of the machine and the sampling frequency, the spatial distance between x_i and the observations inside *SNT(x_i)* is often larger than that between x_i and the observations inside *ST(x_i)*. If the spatial neighbourhood of x_i is defined according to the *k* nearest neighbours, it may be difficult to control the amount of *ST* and *SNT* neighbours. As a consequence, it was decided to select the observations inside *N*(x_i) via a maximal spatial distance below which observations belong to *N*(x_i), and not to rely on a number of neighbours. This spatial distance was set as a function of the distance between adjacent passes, e.g. the cutting width of the combine harvester. Once observations inside *N*(x_i) were found, they were split between *ST(x_i)* and *SNT(x_i)*. To avoid choosing a specific spatial distance for this neighbourhood research, observations inside *N*(x_i) were selected in three different squared neighbourhoods of size two, three and four cutting widths of the machine. The algorithm was then applied on each of these neighbourhoods and the results were averaged over them. It must be stated that the use of these three spatial neighbourhoods gave three times more importance to the neighbours located at a distance less than two cutting widths of the machines than to those located at a distance of four cutting widths of the machine.

A robust metric to quantify the 'outlierness' of each observation

Now that neighbouring relationships have been defined between observations, the spatial outlier-based methodology can be put into place. Each observation x_i will be compared to the observations belonging to its two different neighbourhoods, i.e. *ST* and *SNT* neighbours, to evaluate the 'outlierness' of x_i. As previously explained, a large 'outlierness' value between an observation x_i and its ST, SNT or both ST and SNT neighbours, indicates that the attribute of x_i is significantly different to the attribute of its neighbours and therefore that x_i might be considered as an outlier. As two neighbourhoods are considered for each observation, the attribute functions f_A, g_A and the comparison function h_A can be computed twice. This leads to two measures of 'outlierness', one between x_i and the observations inside *ST(x_i)*, and the other between x_i and the observations inside *SNT(x_i)*. Given the number of defective observations likely to be present in yield datasets, each observation x_i needs to be compared to the observations inside *ST(x_i)* and *SNT(x_i)* with robust metrics not sensitive to outliers. To lessen the influence of possible outliers inside *ST(x_i)* and *SNT(x_i)*, the attribute function g_A was set to return the median of the observations belonging to *ST(x_i)* and *SNT(x_i)*. This summary statistic was proven to be effective in several studies (Chen et al., 2008, Lu et al., 2003). The 'outlierness' measures are defined in the same way for *ST(x_i)* and *SNT(x_i)* with regard to x_i.. The comparison function h_A, i.e. the 'outlierness' measure, was defined as follows:

$$h_A = f_A - g_A \tag{1}$$

Where f_A and g_A are the attribute functions of the variable A corresponding respectively to observation x_i and the observations inside *ST(xi)* and *SNT(xi)*, h_A is the comparison function between f_A and g_A.

Bivariate plot of 'outlierness'

Each observation x_i is now characterized by two measures of 'outlierness' which can be represented in a bivariate plot of 'outlierness' (Fig. 3). The bivariate plot does no longer contain spatial information, i.e. co-ordinates, which means that the spatial outlier detection has now turned into a traditional outlier detection with a two-dimensional dataset. Hence, from now on, all the notions of distances will only refer to distances between observations in the bivariate plot, in the non-spatial attributes domain. From a general perspective, outliers can be defined as those observations that have a strong disagreement with either ST, SNT or both ST and SNT neighbours. Despite the relatively high number of defective observations that can be found in datasets obtained from on-board sensors, the majority of observations can be considered as non-outliers. These non-outliers, or normal observations, must have similar characteristics to that of their ST and SNT neighbours and should all be found in the central portion of the bivariate plot (Fig. 3). Indeed, normal observations have been given a small 'outlierness' measure in absolute to indicate that their attribute value is really similar to that of their neighbours. Observations with large 'outlierness' values with regard to either ST, SNT or both ST and SNT neighbours should be relatively far from the rest of the observations and should be classified as outliers. All these observations must be now classified as outliers or non-outliers to be able to automatically filter a high quantity of maps.

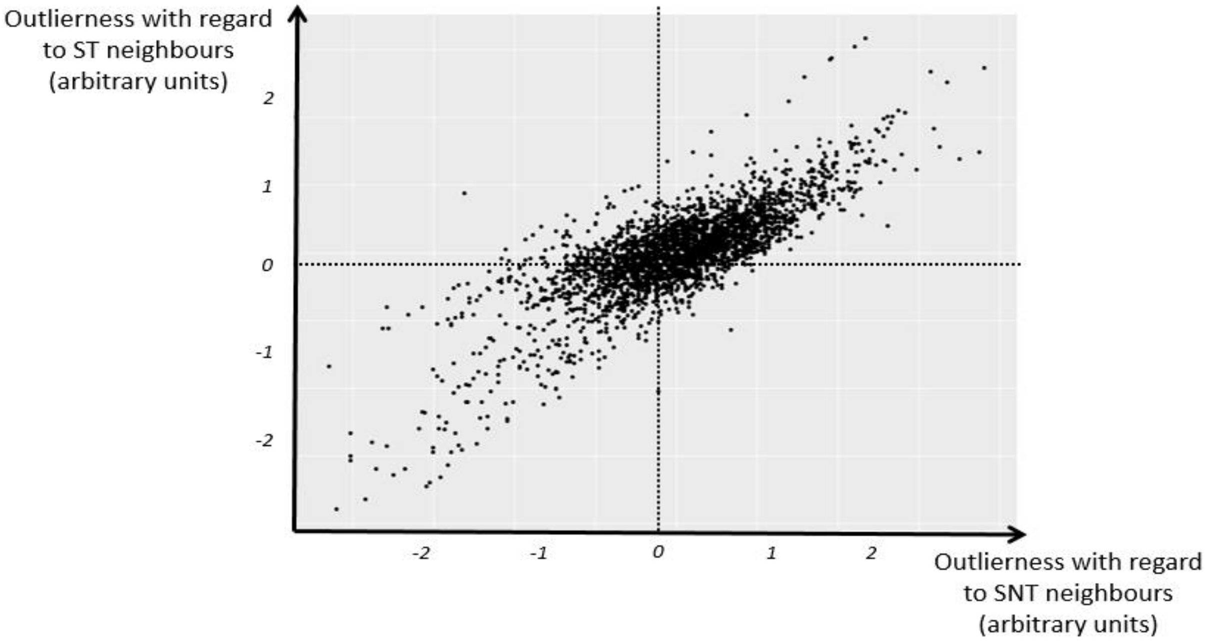

Fig. 3. 'Outlierness' of each observation with its ST and SNT neighbours. *The majority of observations in the centre of the plot have a small 'outlierness' value (relative to zero) with regard to their ST and SNT neighbours which indicates that these observations have a consistent behaviour with their neighbours.*

A density-based clustering algorithm

As the majority of observations are considered non-outliers, the density of observations around normal observations should be much higher than around outliers (Fig. 3). Multiple density-based methods have been proposed in the literature but the threshold to classify observations as outliers or non-outliers is very often selected manually. As a consequence, it was decided to go further and to cluster observations that shared the same density of observations around them. One strong advantage of the clustering-based methods is that they do not give an 'outlierness' score to each observation but rather intend to discover groups of similar observations. On top of that, to automate the outlier identification, a non-parametric, or unsupervised, method should be preferred. Indeed, datasets acquired with on-board sensors are obtained through a variety of conditions, e.g. sensor, crop, operator, field characteristics, conditions of acquisition, and it was considered irrelevant to infer or consider a specific data distribution. A non-parametric method was also needed to deal with any arbitrary shape of the data distribution that was likely to occur. The algorithm DBSCAN (Density-Based Spatial Clustering of Applications with Noise) was selected because of its ability to combine both advantages of density and clustering-based approaches (Ester et al., 1996). This method also fulfils the constraints that were set previously, especially regarding the use of a non-parametric approach to classify the observations. Duan et al. (2007) proposed some improvements of the DBSCAN algorithm but they were not considered very useful in this outlier detection case. Other traditional methods commonly reported, i.e. distribution-based or distance-based (Filzmoser et al. 2014; Harris et al. 2014). might have been used but they were considered difficult to automate and to use in a non-parametric manner. Indeed, distribution-based methods rely on strong statistical assumptions with regard to the distribution of the variable of interest. Distance-based methods often require the variable distribution to be normal so that reliable thresholds can be used to classify observations as outliers (Filzmoser et al., 2014).

DBSCAN requires two parameters to identify clusters: the distance from each observation to its neighbours (ε) and the minimum number of observations inside the neighbourhood given the distance ε (*Minpts*) (Fig. 4). It must be clear that the DBSCAN algorithm is applied on the bivariate plot of 'outlierness' and not on the initial dataset. To avoid any confusion with the neighbourhood $N(x_i)$ previously introduced, this new neighbourhood of an observation x_i will be referred to as $NO(x_i)$, i.e. Neighbourhood with regard to 'Outlierness' values. For a given observation x_i, the algorithm finds its neighbouring observations $NO(x_i)$ given the distance ε and tests whether this *NO* neighbourhood contains at least *Minpts* observations (Fig. 4). When this condition is fulfilled, x_i is set inside the core of a cluster and the algorithm expands the cluster by applying the same method to the observations inside $NO(x_i)$ and their corresponding neighbours until the constraint relative to *Minpts* is no longer respected. For instance, in Figure 4, the triangles have at least five neighbours within an ε distance and therefore are included in the core of the cluster. The square is reachable by one of the triangles, but this square has less than five neighbours. The stars are not reachable by any point inside the core of the cluster and will not be part of the central cluster. If an observation x_j is inside the neighbourhood of an observation x_i but the neighbourhood of x_j contains less than *Minpts* observations, observation x_j is labelled as noise but is still included in the cluster corresponding to x_i, e.g. the square in Fig. 4 (Ester et al., 1996). The insertion of x_j in the cluster related to x_i helps retrieve the global shape of the cluster rather than the core of the cluster only. This method was considered appropriate to build one large cluster retaining all the normal observations while leaving the outliers in other clusters. To obtain a reliable clustering, it was necessary to define the optimal parameters ε and *Minpts*. Some works had already been proposed to determine automatically these criteria but still requires some manual thresholds (Sawant, 2014). The previous work helped to develop a fully-automated approach.

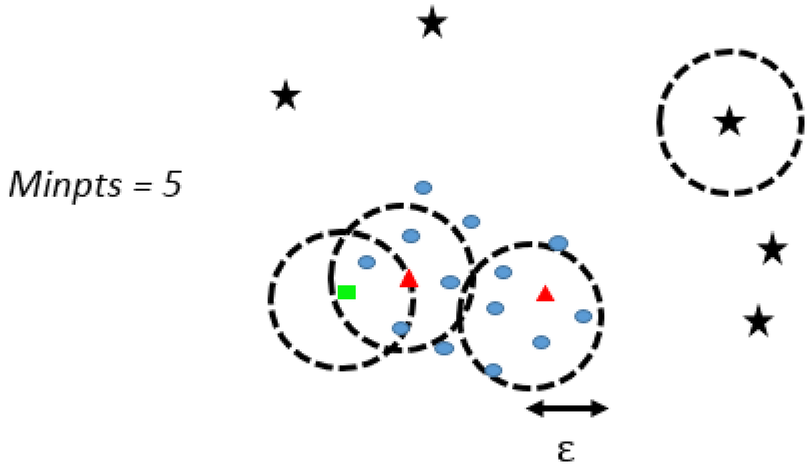

Fig. 4. Application of the DBSCAN algorithm.

The distance ε was defined in the first place as the most frequent distance between two different observations (Fig. 7). In fact, as the majority of observations, i.e. the non-outliers, are expected to be clustered in the same group, the most frequent distance between two different observations should be a characteristic of normal observations. The distances were calculated as euclidean distances between two observations within the bivariate plot of ‘outlierness’ (Fig. 3). The ‘outlierness’ measures were centred and reduced to avoid giving too much influence to one of these two measures of disagreement. Given this optimal ε distance, the number of neighbours inside $NO(x_i)$ was computed for each observation x_i. The distribution of the *NO* neighbours was used to select an optimal value for the *Minpts* parameter (Fig. 7). As the *Minpts* value increases, the size of the clusters diminishes because less and less observations fulfil the requirement regarding the minimum number of observations inside their neighbourhood. This *Minpts* parameter must not be set too high so that the whole shape of the cluster is taken into account. It was stated that a break in the *NO* neighbours’ distribution should reflect an optimal separation between different clusters. This break was chosen to be a local minimum in the distribution of the *NO* neighbours. Indeed, a local minimum corresponding to k neighbours indicates that the observations that have k neighbours within a ε distance are located at the border between two clusters of different density of neighbours. This first local minimum was considered a good indicator of the separation between normal and outlying observations. To optimally select the parameters ε and *Minpts*, the densities of (i) the distance between different observations and (ii) the number of neighbours for an optimal ε distance, were estimated via a kernel density estimation (KDE).

Adjusted filtering for wrongly identified outliers

When the ST and SNT neighbourhoods of an observation x_i contain many defective observations, the function h_A might be sensitive to these outliers, even if robust metrics are used. As a consequence, some observations might be wrongly classified as outliers only because their neighbourhood is outnumbered by outliers. To overcome this limitation, the ‘outlierness’ values attributed to each observation x_i that was previously classified as an outlier has to be re-evaluated. More specifically, each observation must be compared to a neighbourhood that only contains non-outlying observations considering the first iteration of the approach. In this way, the influence of outliers in a spatial neighbourhood is removed. To account for the wrongly identified outliers, a second iteration of the proposed approach was put into place. For each observation x_i, $h_A(x_i)$ was recalculated except that this time, the neighbourhoods of x_i were set free of other outliers. This means that if an observation is definitely an outlier, removing outliers from its neighbourhood will still classify this observation as an outlier. On the other hand, if the observation was wrongly classified as an outlier, removing outliers from its neighbourhood would significantly decrease the ‘outlierness’ values associated and therefore would lead to classifying the observations as a normal one. Once each observation x_i was given new $h_A(x_i)$ values with regard to both ST and SNT neighbours, the classification based on the DBSCAN algorithm was run a second time to identify the real outliers.

Last considerations before using the proposed algorithm

The adjusted spatial outlier detection was not applied directly on the raw dataset. Some corrections were added before applying the proposed algorithm to improve the quality of the results. Among the observations that were likely to affect the efficiency of the proposed algorithm, especially co-located points and global outliers were of great concern and were removed before searching for spatial outliers. Co-located records are observations that are acquired at the same spatial position either due to a stop of the combine or to an error in the GNSS position. In either case, these observations must be filtered out because they exhibit most likely an abnormal value. Global outliers were removed because they could be spotted relatively easily and could have some influence on the detection of the spatial outliers. Global outliers were removed in a non-parametric way following the method of Hubert and Van der Veeken (2008).

To ease the understanding of the proposed approach and knowing that the procedure requires to travel between multiple domains, i.e. spatial, temporal, and attribute, a flowchart of the algorithm is provided (Fig. 5). A step-by-step description of the data filtering process is also presented afterwards.

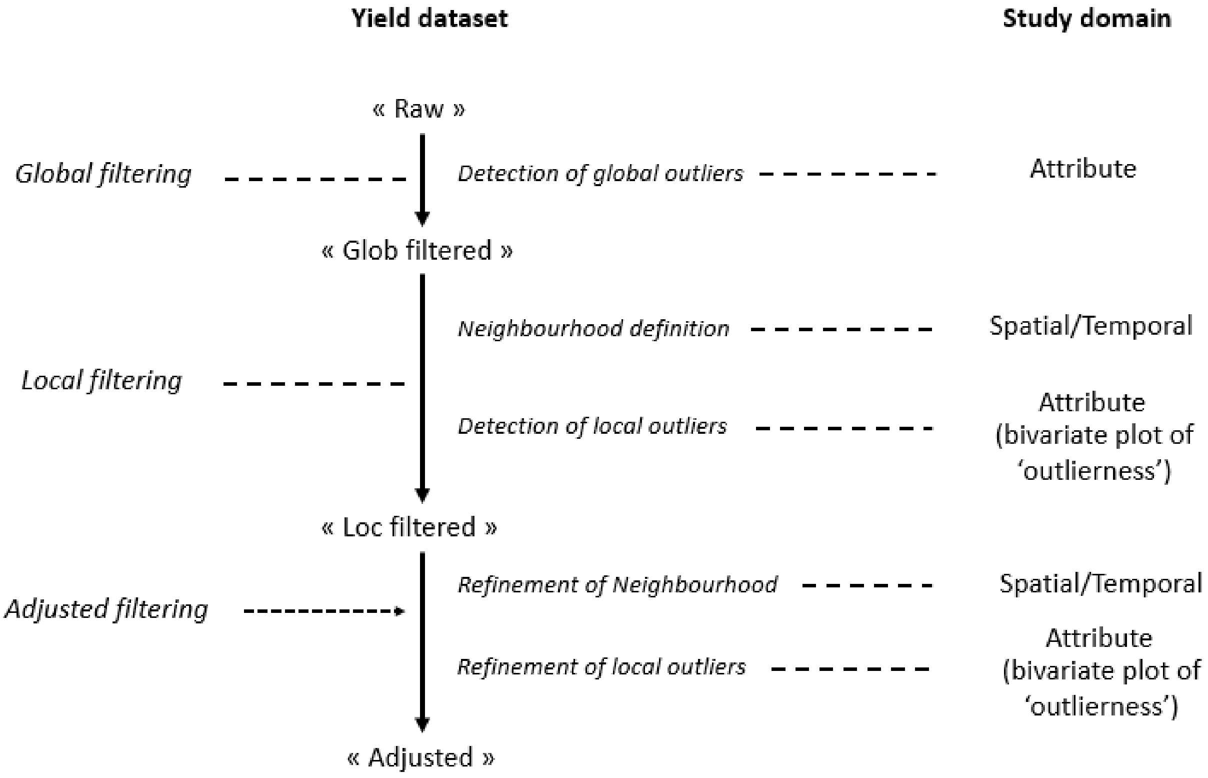
Yield dataset
Study domain
« Raw »
Global filtering
Detection of global outliers
Attribute
« Glob filtered »
Neighbourhood definition
Spatial/Temporal
Local filtering
Detection of local outliers
Attribute (bivariate plot of ‘outlierness’)
« Loc filtered »
Adjusted filtering
Refinement of Neighbourhood
Spatial/Temporal
Refinement of local outliers
Attribute (bivariate plot of ‘outlierness’)
« Adjusted »

Fig. 5. A simple flowchart of the proposed approach

Algorithm (Adjusted Spatial Outlier detection)

```
1. Remove co-located points and global outliers [Global filtering]
2. Remove local outliers [Local filtering]
    a. For each point $x_i$ remaining in the dataset:
        i. Compute the square neighbourhoods based on a radius of two, three and four cutting widths. For each of these neighbourhoods:
            1. Separate the neighbourhood $N(x_i)$ into $ST(x_i)$ and $SNT(x_i)$
            2. Calculate the 'outlierness' value of $x_i$ with regard to $ST(x_i)$ and $SNT(x_i)$ using the function $h_A$
        ii. Average the 'outlierness' values with regard to $ST(x_i)$ and $SNT(x_i)$ over the three square neighbourhoods
    b. Determine optimal parameters $\varepsilon$ and Minpts for the DBSCAN algorithm
    c. Apply DBSCAN to extract the cluster consisting of normal observations
3. Refine detection of local outliers [Adjusted filtering]
4. Extract the final cluster of normal observations.
```

The whole methodology was developed using the R statistical environment (R Core Team, 2013).

Evaluation of the proposed algorithm and datasets used

The proposed algorithm was tested on ten real within-field yield datasets arising from three different farms, i.e. two located near Evreux, in the north-western part of France (Farm 1 - WGS84: E:0.779, N:48.955; Farm 2 - WGS84: E:1.032, N:48.828) and one located close to Peterborough, UK (Farm 3 - WGS84: E:-0.105, N:52.643). Fields were mostly cropped in wheat and harvested with combines of different brands, especially New Holland (Turin, Italy) and Claas (Harsewinkel, Germany) combines. Among these ten datasets, two of them (dataset 1 from Farm 3 and dataset 2 from Farm 2) were selected to provide readers with a deeper analysis of the proposed approach. The two datasets (datasets 1 and 2) were especially chosen for containing different sources of defective observations. Table 1 and 2 respectively report yield statistics for the two (datasets 1 and 2) and ten datasets under consideration.

For the ten yield datasets, the proposed approach was evaluated in the same way as in many previous studies, i.e. by looking at the yield distribution and spatial structure before and after filtering out outliers (Simbahan et al., 2004, Sudduth et al., 2007). This evaluation procedure still has some limits as this validation remains somehow qualitative. Indeed, outliers are not labelled in the yield datasets so one cannot be entirely sure whether an outlier is truly one. However, this procedure was considered sufficient in the first instance. Furthermore, for datasets 1 and 2, the detected outliers were plotted on their corresponding field to better understand their characteristics.

Results and discussion

Improvements in the yield distribution and spatial structure

A specific attention to datasets 1 and 2

Both raw and filtered yield datasets of datasets 1 and 2 are presented via their principal descriptive statistics (Table 1) and semi-variograms (Fig. 6). The pre-filtering step, i.e. *Glob outliers*, consisted in the removal of co-located points and global outliers. Observations with a yield value equal to zero were also discarded because they were likely to mask the presence of some global outliers. Indeed, observations with a zero yield value are definitely not expected and might have been obtained when the cutter bar was lowered while the crop had already been harvested. The removal of zero yield values, co-located points and global outliers substantially changed the summary statistics of yield datasets by lowering the standard deviation by a factor of 2 and increasing the average yield in the fields by almost 5%. More interestingly, these first outliers were completely masking the yield spatial structure in the two fields of interest (Fig. 6). Indeed, the semi-variograms testify of a clear yield spatial structure with well-defined nugget and sill parameters. These results demonstrate to what extent a simple pre-filtering approach such as the removal of global outliers and really unexpected values (zero-yield observations) can improve the characteristics of within-field yield datasets.

Table 1. Yield descriptive statistics (t ha^{-1}) of datasets 1 and 2. *'Raw' stands for the original dataset. 'Glob filtered' is the original dataset after the pre-filtering step (essentially global outliers, co-located points and zero-yield observations. 'Loc filtered' is the dataset after the pre-filtering step and the removal of local outliers. 'Adjusted' is the dataset after adjustment for wrongly identified outliers. SD stands for standard deviation. Nb. observations is the number of observations in the corresponding dataset.*

Dataset	Type	Min	Mean	Median	Max	SD	Nb. observations
1	Raw	0	7.75	8.13	90.41	2.65	6526
	Glob filtered	3.20	8.04	8.20	11.64	1.38	6143
	Loc filtered	4.26	8.26	8.32	11.31	1.04	5333
	Adjusted	4.56	8.26	8.31	11.37	1.06	5400
2	Raw	0	8.65	9.10	40.00	1.99	3279
	Glob filtered	5.80	9.07	9.20	11.20	0.87	3003
	Loc filtered	6.80	9.16	9.20	11.20	0.68	2743
	Adjusted	6.80	9.16	9.20	11.20	0.70	2803

Local outliers were removed from the previously pre-filtered dataset (*Loc filtered*). These outliers have less influence on yield summary statistics compared to the global outliers (Tab. 1). This is essentially due to the fact that these statistics characterize the yield dataset at a global level. However, it can be seen that local statistics are substantially impacted by these local outliers (Fig. 6). The spatial structure appears effectively much more clearly once outliers have been removed. As expected, the final step of the proposed methodology, i.e. *Adjusted*, does not produce major improvements on either the yield distribution or spatial structure. This step can rather be considered like a refinement of the proposed approach and was not aimed to drastically impact the yield characteristics.

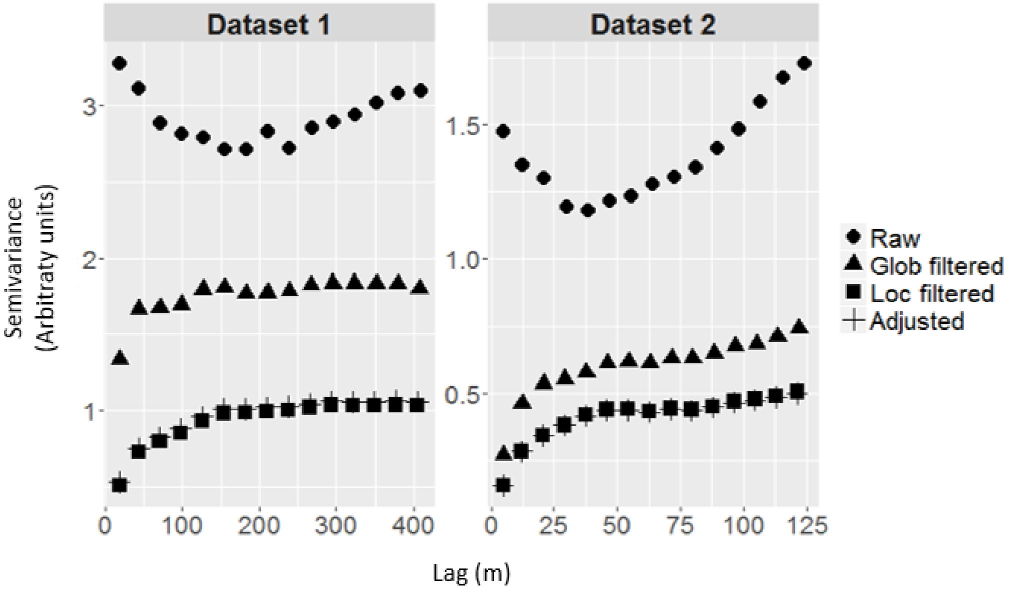

Fig. 6. Spatial structure of yield datasets 1 and 2 with the proposed methodology. '*Raw' stands for the original dataset. 'Glob filtered' is the original dataset after the pre-filtering step (essentially global outliers, co-located points and zero-yielding observations. 'Loc filtered' is the dataset after the pre-filtering step and the removal of local outliers. 'Adjusted' is the dataset after adjustment for wrongly identified outliers.*

Analysis of the ten datasets under study

Table 2 reports descriptive and spatial statistics regarding the ten datasets under consideration. All the raw yield datasets exhibit a large variability, i.e. high coefficient of variation, because of the presence of global and local defective observations. The influence of local outliers on the yield spatial structure is clear for all the ten datasets under study (Table 2). Indeed, nugget to sill ratios are significantly improved, i.e. reduced, once local outliers are filtered out from the yield datasets. Even though the level of autocorrelation remains medium for some datasets after the removal of local outliers, e.g nugget to sill ratio more than 50%, it must be clear that the spatial structure is still stronger than when local outliers were left inside the yield datasets. The proposed methodology removed a relatively high number of observations, i.e. from 19 to 50% of the dataset size (Table 2). These defective observations are, at the same time, global and local outliers, and with different proportions of each type for each dataset. Some datasets effectively contain more global outliers, e.g. because of more measurements when the cutting bar was up, while more local outliers have been filtered in others, e.g. more speed changes. Note also that this number of defective observations substantially varies among the datasets under study which demonstrates that all yield datasets are different and that a general filtering methodology is interesting to consider.

Table 2. Yield statistics for the ten datasets under consideration. *Spatial statistics are presented before (Glob filtered) and after (Loc Filtered) removing yield local outliers. The percentage of points removed during the whole filtering process (from 'raw' to 'adjusted' yield datasets) is also reported. CV stands for the coefficient of variation.*

Dataset	**Descriptive statistics (Raw yield dataset)**			**Spatial statistics**		
	Surface (ha)	Mean (t.ha^{-1})	CV (%)	Nugget/Sill (%) *Glob filtered*	Nugget/Sill (%) *Loc Filtered*	Points removed (%)
1	20.5	7.75	23.0	72	52	19
2	3.5	8.65	34.2	66	49	17
3	13.1	5.1	114	100	50	42
4	28.0	4.9	57.5	100	55	45
5	45.2	6.3	48.5	82	41	33
6	10.5	7.1	67.7	85	33	50
7	25.1	7.5	50.3	92	29	33
8	13.1	7.6	60.7	85	76	32
9	22.2	7.1	73.4	84	40	39
10	30.5	9.4	21.1	32	21	21

Evaluation of the density-based clustering approach

Detection of the DBSCAN parameters

Figure 6 demonstrates the application of the DBSCAN algorithm on the bivariate plot of 'outlierness' for datasets 1 and 2. For both datasets, there is a clear maximum value in the density of distances between different observations which enables a clear detection of the ε distance (Fig. 7, left). As the distance between different observations increases, more and more distant observations are considered. Small distances between different observations characterize essentially the nearest neighbour distances between observations inside the core of the cluster of normal observations which is why the density is relatively low for these distances. The clear peak identifies the distance between two different observations that is the most representative of the cluster of normal observations. After the peak, the larger distances account for very distant observations such as, for instance, a normal observation and an outlier, or two normal observations very far from each other inside the cluster of normal observations. The most frequent distance between two different observations should therefore reliably discriminate the cluster of normal observations.

For both datasets, the first local minimum in the density of the number of neighbours, i.e. corresponding to the parameter *Minpts*, appears relatively clearly. For the optimal ε distance that was previously chosen, as the number of neighbours increases, the density of the number of neighbours starts decreasing relatively quickly then increases smoothly at first, then more abruptly (Fig. 7, left). The first peak and neighbouring values is due to outlying observations that have a few number of neighbours within an ε distance. The last peak and neighbouring values are related to normal observations, i.e. inside the core of the cluster of normal observations, with a very high number of neighbours within an ε distance. Between these two peaks, the first local minimum in the density of the number of neighbours, from the lesser to greater number of neighbours, is considered a good separator between the cluster of normal observations and the outliers. Indeed, it separates a high-density region from low-density regions. The first local minimum is also generally the global minimum in the distribution of the *NO* neighbours. It was therefore selected as a good estimate to separate outliers from normal observations.

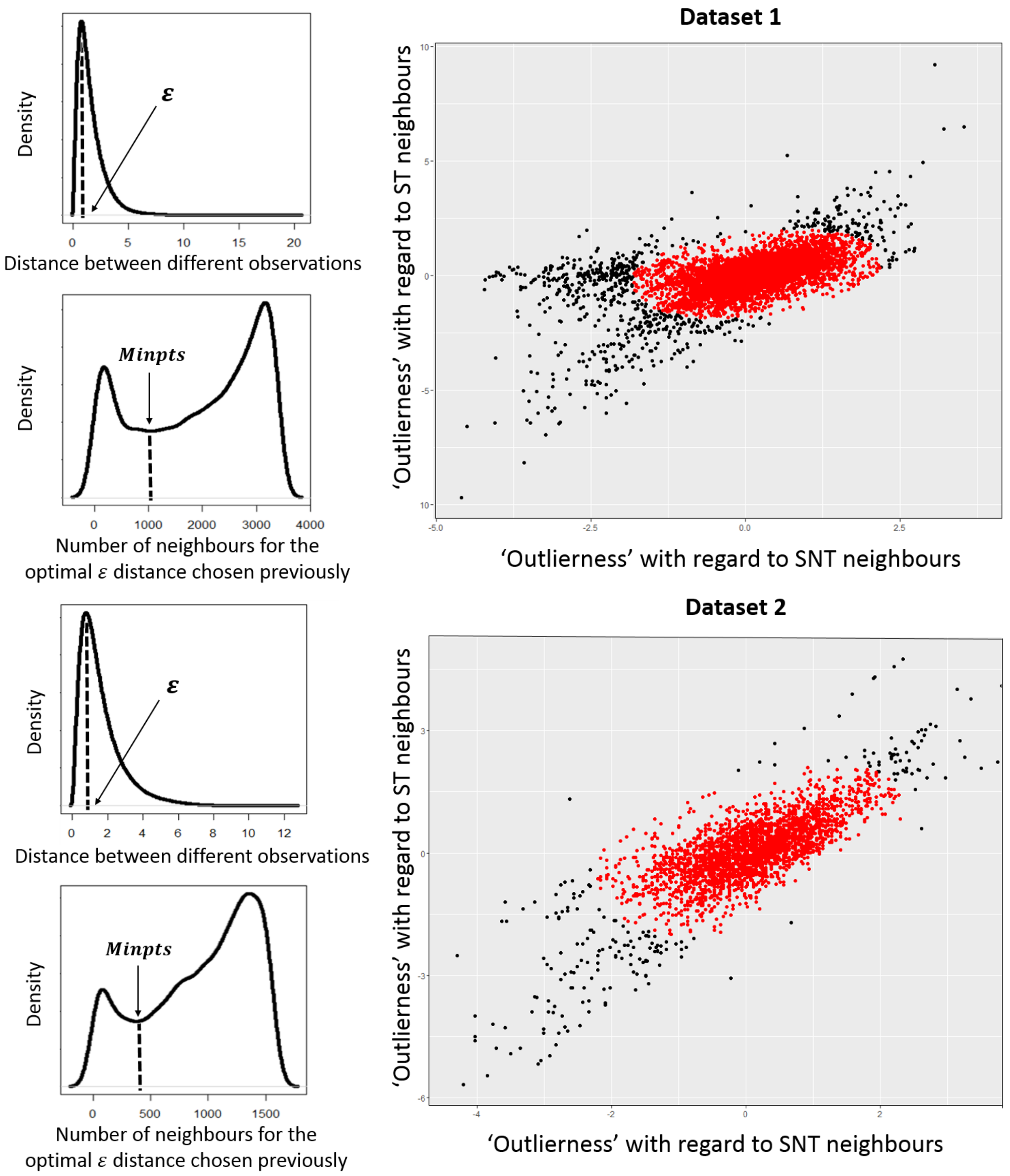

Fig. 7. Optimal selection of DBSCAN parameters and corresponding detection of normal observations. *For each dataset, the plot in the top left-hand corner helps determine the optimal ε distance while that in the bottom left-hand corner enables retrieval of the Minpts parameter of the DBSCAN algorithm. The right plot shows the cluster of normal observations in the centre portion of the plot (red dots in the online version) and the outliers identified by the proposed method.*

Be aware that the identification of the DBSCAN parameters, i.e. ε and *Minpt*, was clear for the ten yield datasets under consideration (data not shown) meaning that the outliers could be separated from the rest of the observations. Depending on the type and number of defective observations, the shape of the density curves did not match perfectly but the overall structure was similar.

Detection of local outliers

Using the parameters previously defined, the DBSCAN algorithm was able to find a large and dense cluster in the centre of the data for both datasets (Fig. 7, right). Regarding dataset 1, some outliers expand towards the left of the plot and exhibit a large 'outlierness' value with regard to their SNT neighbours (value far from 0) and a low 'outlierness' value with regard to their ST neighbours. These outliers are observations that belong to passes harvested with a low cutting width (Fig. 8, left). Indeed, a long tail of low-yielding observations surrounded by observations with a much higher yield value is very often the sign of a non-fully used cutting bar. These observations are consistent with spatially close observations in the same pass because all of them were recorded with a low cutting width. In contrast, adjacent passes were harvested with a full cutting bar which is why these outliers do not share similar characteristics with spatially close observations in adjacent passes.

For the two fields under study, some outliers are located on the diagonal of the plot, i.e. the yield values of these observations are significantly different from those of their ST and SNT neighbours. This characteristic is specific to observations recorded at the start and end of each row (Fig. 8). When the combine harvester enters or leaves a pass, the grain flow can be significantly different from within the pass. The filling time at the start of a harvest pass induces an under-estimation of the yield because the grain flow is increasing and still has not reached a plateau, i.e. within the pass. Therefore, the yield measurement does not match the expected true yield. At the end of a harvest pass, some grain might still continue to flow after the last crop was harvested and the lag time has been reached. As a consequence, these observations have a different behaviour to that of spatially close observations in the same and in adjacent passes. These two known sources of error are the most easily detected on the right-hand side plot because the corresponding observations are relatively clustered on the plot.

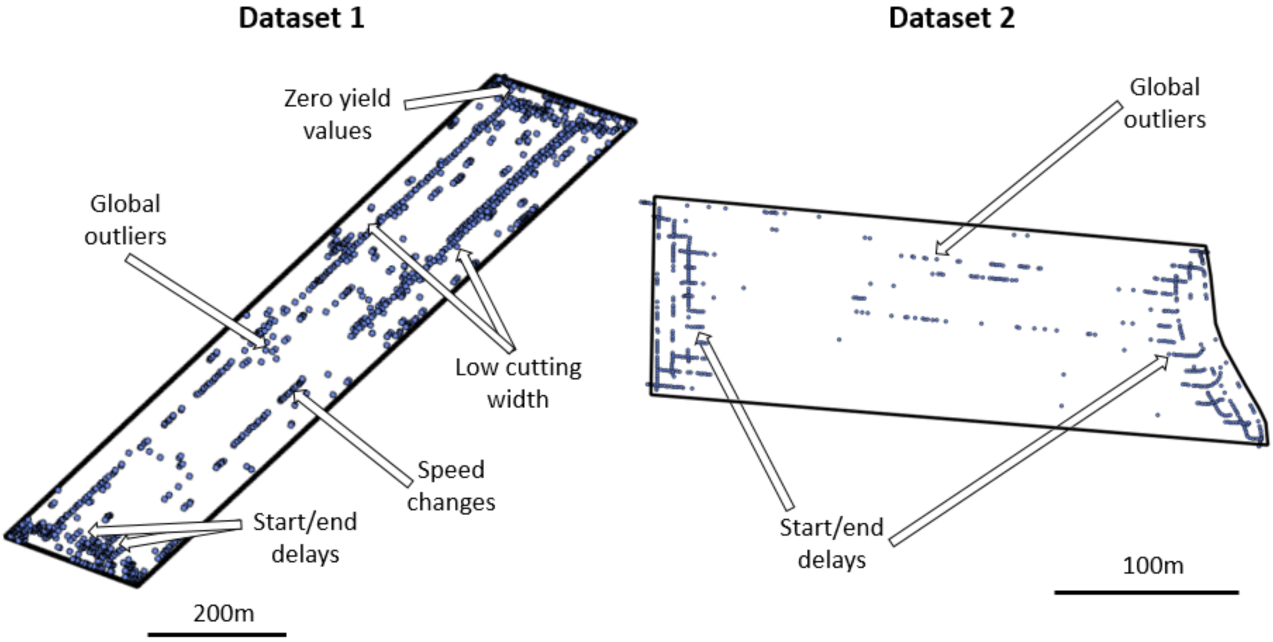

Fig. 8. Location and label of the outliers detected by the proposed methodology within the fields.

In contrast to the two previous sources of errors, some outliers are detected more irregularly in the field and might be the sign of abrupt speed changes or bad moisture/yield records. Some of these other sources of error, e.g. speed changes can also be identified relatively precisely. From a practical point of view, the yield is the ratio of the grain flow to the corresponding harvested area during a fixed time interval. A harvest area can be defined by both the cutting width and the travel speed of the machine. As a consequence, large speed variations during a specific time interval result in large yield variations. Observations acquired during a speed change will likely have different properties than those of spatially close observations in the same and in adjacent passes.

Regarding the ten datasets under study, each bivariate plot of ‘outlierness’ had its proper characteristics depending on the types of outlier present (data not shown). Some features were recurrent, e.g. the outliers located on the diagonal of the plot, because all the yield datasets contained observations related to the filling and emptying time of the machine, to a greater or lesser extent. Others were less present such as the tail of outliers expanding towards the left of the plot because overlaps were very rare within the datasets. From a general perspective, it is clear that each yield dataset has its own properties. This implies that there is a need for filtering procedures to be as flexible and general as possible so that each dataset can be processed accordingly no matter the type or number of outliers.

The proposed methodology has only been applied to the yield attribute in yield datasets. It could be argued that none of the other attributes in yield datasets, such as the speed of the machine or the grain moisture, had been used to detect the possible outliers. It was actually considered that all these attributes were used in the calculation of the yield attribute and therefore that any strong deviations of one of these attributes should have led to a bad yield estimate that would have been spotted as outliers with regard to its ST or SNT neighbours.

In this study, several outliers that were detected by the proposed algorithm were put in relation to some technical errors that can be found within yield datasets (Fig. 8). Nonetheless, it remains relatively difficult to assess the effectiveness of a specific filtering methodology. In fact, as raw yield observations are not labelled within the datasets, one cannot be entirely sure whether an observation identified as an outlier is truly one. Obviously, some errors are clearly visible on the map but for others, it is much more difficult to be sure, even with a skilled operator. To cope with this issue, one possibility could be to generate simulated yield datasets in which the location of outliers is known so as to assess more objectively the interest and reliability of a filtering approach (Leroux et al., 2017). Another improvement of the proposed methodology would be to intend to correct the outliers detected instead of abruptly removing them. Indeed, even if the removal of outliers is not dramatic for the size of the yield datasets as they already contain lots of observations, it could be still interesting to see whether a proper correction is possible. As it was found that multiple sources of error had a specific behaviour in the bivariate plot of ‘outlierness’, it might be conceivable to identify and label these errors so as to propose a correction. For instance, observations belonging to passes harvested with a low cutting width could be extracted and corrected properly by estimating the proportion of the cutting width that was actually used when these observations were acquired.

Conclusion

A new holistic data-driven method was proposed to filter out local outliers from within-field yield datasets. This approach essentially consisted in finding observations whose attribute of interest had the most significant difference with regard to that of the observations inside their spatial neighbourhood. To meet the specificities of within-field yield datasets, a new concept of neighbourhood has been formalised. Outlying observations were then detected by a density-based clustering method. One of the major interests of the approach is that it does not require any manual settings prior to the filtering. All metrics and thresholds are driven by the data themselves. The approach was successfully tested on yield datasets but could be extended to many more spatial datasets from on-the-go sensors. Besides, it must be said that the methodology was applied solely on the yield attribute, i.e. on univariate datasets. The approach could also be extended to datasets of higher dimension. Overall, the proposed algorithm was proven effective at removing unwanted observations from on-the-go vehicle-based yield datasets and should be used as a first step before deeper processing. Despite significant improvements in the distribution and spatial structure of yield datasets, the evaluation of the algorithm was still subjective. Future work will involve the comparison of multiple approaches through the use of simulated datasets to offer much more objective conclusions.

References

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

Arslan, S. (2008). A Grain Flow Model to Simulate Grain Yield Sensor Response. *Sensors,* 8, 952–962.

Ben-Gal, I. (2005). Outlier Detection. *The Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers. Boston:* Kluwer Academic Publishers.

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

Chen, D., Lu, C-T., Kou, Y. & Chen, F. (2008). On Detecting Spatial Outliers. *Geoinformatica*, 12, 455-475

Chung, S. O., Sudduth, K. A., & Drummond, S. T. (2002). Determining Yield Monitoring System Delay Time With Geostatistical and Data Segmentation Approaches. *Transactions of the ASAE*, 45, 915-926.

Diker, K., D.F. Heerman, & M.K. Brodahl. (2004). Frequency analysis of yield for delineating yield response zones. *Precision Agriculture*, 5, 435–444.

Drummond, S. T., Fraisse, C. W., & Sudduth, K. A. (1999). Combine Harvest Area Determination by Vector Processing of GPS Position Data. *Transactions of the ASAE*, 42, 1221–1227.

Duan, L., Xu, L., Guo, F., Lee, J., & Yan, B. (2007). A local-density based spatial clustering algorithm with noise. *Information Systems*, 32, 978–986

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In E. Simoudis, J. Han, and U. Fayyad (Eds.), *Proceedings of Second International Conference on Knowledge Discovery and Data Mining*, Palo Alto, CA, USA: AAAI Press, pp 226–231.

Filzmoser, P., Ruiz-Gazen, A. & Thomas-Agnan, C. (2014). Identification of local multivariate outliers. *Statistical Papers*, 55, 29-47.

Florin, M.J., McBratney, A.B., Whelan, B.M. (2009). Quantification and comparison of wheat yield variation across space and time. *European Journal of Agronomy*, 30, 212-219.

Gogoi, P, Bhattacharyya D, Borah B, & Kalita JK (2011). A survey of outlier detection methods in network anomaly identification. *Computer Journal*, 54, 570–88.

Griffin, T., Dobbins, C., Vyn, T., Florax, R., & Lowenberg-DeBoer, J. (2008). Spatial analysis of yield monitor data: case studies of on-farm trials and farm management decision making. *Precision Agriculture*, 9, 269–283

Harris, P., Brunsdon, C., Charlton, M., Juggins, S., & Clarke, A. (2014). Multivariate Spatial Outlier Detection Using Robust Geographically Weighted Methods. *Math Geosciences*, 1–31.

Hawkins, D. (1980). *Identification of Outliers*, London, UK: Chapman & Hall

Hu, J., Gong C., & Zhang Z. (2012) Dynamic Compensation for Impact-Based Grain Flow Sensor. *In: Li D., Chen Y. (eds) Computer and Computing Technologies in Agriculture V. CCTA 2011. IFIP Advances in Information and Communication Technology*, vol 370, 210-216, Berlin, Heidelberg, Germany: Springer

Hubert, M., & Van der Veeken, S. (2008). Outlier detection for skewed data. *Journal of Chemometrics*, 22, 235–246

Jingtao, Q., & Shuhui, Z. (2010). Experiment research of impact-based sensor to monitor corn ear yield. *International Conference on Computer Application and System Modeling*, IEEE, 101, 187–192.

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, 743–752.

Leroux, C., Jones, H., Clenet, A., Dreux, B., Becu, M., Tisseyre, B. (2017). Simulating yield datasets: an opportunity to improve data filtering algorithms. In J A Taylor, D Cammarano, A Preashar, A Hamilton (Eds.), Proceedings of the 11th European Conference on Precision Agriculture, Precision Agriculture ’17 (*Advances in Animal Biosciences 8 (2) 600-605).*

Lu, C.-T., Chen, D., & Kou, Y. (2003). Algorithms for spatial outlier detection. In X.Wu, A. Tuzhilin, and J. Shavlik (Eds.) *Proceedings of the Third IEEE International Conference on Data Mining*, Los Alamitos, CA, USA: IEEE Press, pp 597-600.

Lyle, G., Bryan, B., & Ostendorf, B. (2013). Post-processing methods to eliminate erroneous grain yield measurements: review and directions for future development. *Precision Agriculture*, 15, 377-402.

Pringle, M. J., McBratney, A. B., Whelan, B. M., & Taylor, J. A. (2003). A preliminary approach to assessing the opportunity for site-specific crop management in a field, using a yield monitor. *Agricultural Systems*, 76, 273–292.

R Core Team (2013). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria.

Reinke, R., Dankowicz, H., Phelan, J., & Kang, W. (2011). A dynamic grain flow model for a mass flow yield sensor on a combine. *Precision Agriculture*, 12, 732–749

Reitz, P., & H. D. Kutzbach (1996). Investigations on a particular yield mapping system for combine harvesters. *Computers and Electronics in Agriculture*, 14, 137–150.

Robinson, T. P., & Metternicht, G. (2005). Comparing the performance of techniques to improve the quality of yield maps. *Agricultural Systems*, 85, 19–41

Sawant, K. (2014). Adaptive methods for determining DBSCAN parameters. *International Journal of Innovative Science, Engineering & Technology*, 1, 330-334

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

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. The Netherlands:* Wageningen Academic Publishers, pp 259-266.

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.

Taylor, J. A., Mcbratney, A. B., & Whelan, B. M. (2007). Establishing Management Classes for Broadacre Agricultural Production. *Agronomy Journal*, 99, 1366–1376.

Tobler W. (1970) A computer movie simulating urban growth in the Detroit region. *Economic Geography*, 46, 234-240

Zhao, C., Huang, W., Chen, L., Meng, Z., Wang, Y., & Xu, F. (2010). A harvest area measurement system based on ultrasonic sensors and DGPS for yield map correction. *Precision Agriculture*, 11, 163-180.