Knowledge discovery and unsupervised detection of within-field yield defective observations

C. Leroux, H. Jones, A. Clenet, Bruno Tisseyre

Author-produced version of the article published in Computers and Electronics in Agriculture, 2019, N°156, p. 645-659.
The original publication is available at https://www.sciencedirect.com
Doi: 10.1016/j.compag.2018.12.024

Knowledge discovery and unsupervised detection of within-field yield defective observations

Leroux, Corentin (1-2), Jones, Hazaël (2), Clenet, Anthony (1), Tisseyre, Bruno (2)

(1) SMAG, Montpellier, France
(2) ITAP, Montpellier SupAgro, Irstea, Univ Montpellier, Montpellier, France

cleroux@smag-group.com

Abstract

Suspicious observations, or the so-called outliers, are always present, to a greater or lesser extent, in agronomical and environmental datasets. Within field yield datasets are no exception. While most filtering approaches use expert thresholds and dedicated filters to remove these defective observations, more general and unsupervised methods will be required to process a growing number of yield maps. However, by using these last approaches, outliers would be solely identified and would remain unlabeled. This study proposes a methodology to provide a label to these defective observations so that users can better characterize the harvest process, e.g. functioning of the machine, driving of the operator, and provide guidelines for future improvements of equipment and operations processes. Here, it is assumed that outliers have already been detected by a non-parametric and unsupervised published approach. Clusters of outliers are first identified in the data to gather outliers with similar yield outlying characteristics. Once detected, these clusters are given a first-order label which describes the general yield outlying characteristics of the observations that belong to these clusters. Then, within each cluster, each outlier is given a second-order label to provide more information on the origin of the defective observation. Yield simulated datasets with known characteristics and labelled outliers were used to test the methodology. The proposed approach was then applied on real yield datasets with unlabeled outliers. This study shows that it might be conceivable to label outliers detected with an unsupervised approach but that some labels are more accurate than others, especially those related to an unknown cutting width of the harvester or to narrow finishes within the fields. Outlying observations behaved similarly between simulated and real datasets which made it possible to infer more precisely the label of defective observations. By labelling outlying observations, it was possible to provide an appropriate correction to one of the real yield dataset and to restore almost 15% of the outlying observations instead of removing them. This study is a first attempt to provide a label to yield outliers detected from an unsupervised manner.

Keywords: Intentional knowledge, knowledge discovery, outliers clustering, outliers labelling, yield

1. Introduction

The agricultural sector faces an impressive and still increasing flow of data arising from multiple platforms, i.e. satellites, UAV, drones, or embedded and in-situ sensors (Baluja et al., 2012; Debuisson et al., 2010; Oliver, 2010; Santesteban et al., 2013). All these data are very helpful for the decision-making process but come along with varying degrees of quality or reliability. More specifically, defective observations, i.e. the so-called outliers, are likely to be present within these data (Simbahan et al., 2004; Sudduth et al., 2007). Those suspicious observations must be carefully considered before involving the datasets in complex agronomic processes or decisions. This is particularly the case for within-field yield datasets which are a valuable tool to highlight the within-field spatial variability and understand the underlying factors affecting this variability (Pringle et al., 2003). Yield datasets are negatively impacted by a noticeable amount of defective observations widely reported in the literature, e.g. filling and emptying time, speed changes, unknown cutting width when entering the crop, GNSS positioning, harvest turns and narrow finishes (Arslan, 2002; Lyle et al. 2013). It must be clear that these defective observations are not erroneous measurements from the yield monitors. These defective observations are problematical because they do not correspond to the yield that should be observed in the field. They are rather biased by the fact that a combine harvester passes through the field. In the case of within-field yield monitor data, Griffin et al. (2008) have shown that in half of their experiments, the quality of the filtering procedure would have supported different field management recommendations.

Author-produced version of the article published in Computers and Electronics in Agriculture, 2019, N°156, p. 645-659.
The original publication is available at https://www.sciencedirect.com
Doi: 10.1016/j.compag.2018.12.024

For the past twenty years, several approaches have been proposed in the literature to tackle the issue of yield defective observations (Blackmore and Moore, 1999; Leroux et al. 2017; Simbahan et al. 2013; Sudduth et al. 2007; Sun et al. 2013). All these methodologies have come up with one single objective, which is to remove all the outliers from the datasets. This way of thinking is legitimate because (i) these suspicious observations influence the overall quality of the data, and (ii) yield datasets contain lots of yield records which means that these datasets can handle a loss of data. Among the multiple approaches that were published in the literature to filter yield datasets, most of them rely on manual expertise and/or dedicated expert thresholds and filters. With these approaches, the labelling of outlying observations, i.e. the fact of attaching information with respect to the origin of the outlier, is directly provided as each empirical or semi-automatic threshold/filter is specific to a type of defective observation. However, with the growing number of yield maps that will need to be processed in the near future, non-parametric and automatic methodologies might be preferred (Leroux et al., 2017; Spekken et al. 2013). In this latter case, as the filtering is thought from a holistic perspective, the labelling of each outlier is not known when defective observations are identified. There is effectively no information or description attached to the outlier, i.e. the origin of this outlying information, e.g. speed change, filling and emptying time, is not known.

The labelling of outlying yield observations is especially relevant since there exists a lot of expert knowledge on (i) the types of defective observations and on (ii) the attributes associated to the yield records to help explain the origin of the errors (Arslan, 2002; Blackmore and Moore, 1999; Lyle et al. 2013). From a more general perspective, the labelling of observations has multiple interests such as the possibility to (i) explain what is causing these outliers, (iii) characterize the working of a machine or the driving of an operator, (iii) correct outlying observations instead of removing them or (iv) provide guidelines for future improvements of equipment and operations processes (Colaço et al., 2014). Once outliers are detected inside yield datasets, it seems therefore possible to provide a detailed description or at least a labelling of the suspicious observations. However, even though an expertise is available, it can sometimes be quite difficult to assess with a strong confidence whether a detected outlier is truly one. By performing a visual inspection on the field, it can be argued that some outliers are clearly visible, but this is not always the case. Moreover, such a visual inspection is cumbersome and may remain subjective when dealing with large amounts of data to analyze. To improve the identification and labelling process, one solution could be to use simulated datasets in which each observation would be labelled either as a normal or defective observation (Leroux et al. 2018). As the location and labelling of outliers would be known in advance, it would be much easier to validate a proposed procedure.

Assuming that a person's noise is another person's signal, several studies, though much less than those related to outlier detection, have intended to provide a label to outliers so that users can better understand their characteristics and origin (Anguilli et al. 2012; Ertoz et al. 2004; Knorr and Ng, 1999; Marques et al. 2015; Micenková et al. 2013). These studies have been either dedicated to categorical (Anguilli et al. 2009; Ertoz et al. 2004) or numerical data (Knorr and Ng, 1999; Micenková et al. 2013). Given that within a dataset, an observation is characterized by a set of m attributes, most of these works seek to provide a subset of k attributes ($k \leq m$) that best explain the 'outlierness' of each defective observation, i.e. the attributes which make the query observation most outlying. Outliers are generally given a score of 'outlierness' in each possible subset of attributes to record how much these suspicious points deviate from the rest of the data (Duan et al. 2015; Micenková et al. 2013; Vinh et al. 2016). For a given outlier o, the subset of attributes for which the outlying score of o is the highest is generally chosen to be the best descriptor of o. As suggested by Micenková et al. (2013), a reliable and valuable subset of attributes should highlight the 'outlierness' of the defective observations but at the same time be minimal in the number of attributes.

The main contribution of this study is to propose a framework to label outlying within-field yield observations. It is considered that these outliers have already been detected by an unsupervised filtering approach, but they are still missing a label. To the authors' best knowledge, very few unsupervised approaches have been dedicated to outlier detection in within-field yield datasets and none of them have been further extended to give a label to these defective observations once detected. Here, a procedure is proposed to provide outlying observations with a label so that users can extract and gain knowledge with regard to their data. The approach is first validated on simulated yield datasets with known labelled outliers and then tested on real yield dataset with unlabeled outliers.

Author-produced version of the article published in Computers and Electronics in Agriculture, 2019, N°156, p. 645-659.
The original publication is available at https://www.sciencedirect.com
Doi: 10.1016/j.compag.2018.12.024

2. Material and methods

2.1 Theoretical considerations

An important pre-requisite of this study is that outliers are already detected within the yield datasets. The aim is not to provide a way to find outliers but rather to help qualify and describe these defective observations. In this work, it is considered that yield outliers have been identified by a holistic and unsupervised filtering methodology proposed by Leroux et al. (2018). As stated in the introduction section, most of the existing filtering approaches provide a direct labelling of the outlying observations as empirical filters and expert thresholds are involved in the detection process (Simbahan et al., 2004; Sudduth et al., 2007). If the filtering process was to be made from a general, non-parametric and automation perspective, outlying observations would be identified but not labelled. These pre-requisites are becoming essential as more and more yield maps will need to be processed in the future. The objective here is to intend to provide a label to these outlying observations once they are spotted in the datasets. A brief summary of the approach of Leroux et al. (2018) is provided in the next section.

2.1.1 Detection of spatial defective observations using a density-based clustering algorithm

This approach is based on a spatial outlier detection problem in which the authors consider that an observation is defective if this latter is inconsistent with the observations in its neighbourhood. The methodology is divided into three major steps. Firstly, each observation x_i is given two different neighbourhoods. (Fig. 1). The first one is a spatio-temporal neighbourhood (ST), which regroups the spatial observations near in space to x_i and which belong to the same harvest row as that of x_i (Fig. 1). The other is a spatio-not-temporal neighbourhood (SNT), which gathers the spatial observations near in space to x_i and which belong to adjacent harvest rows to that of x_i.

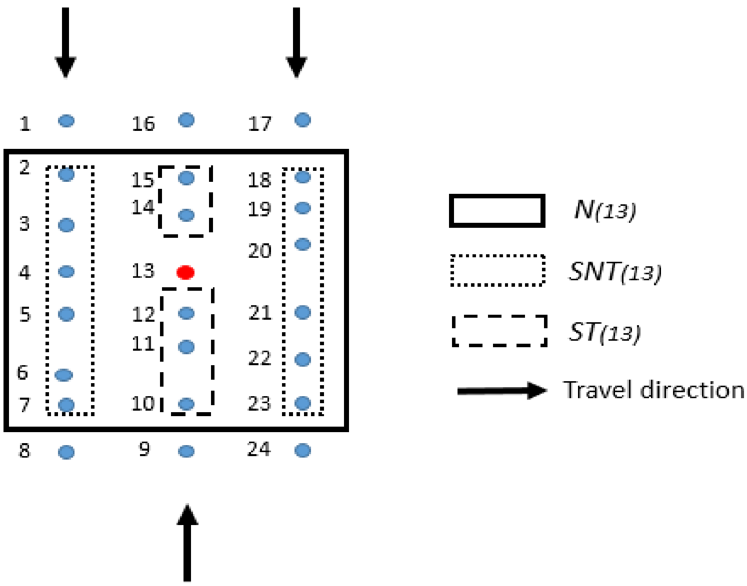

Fig. 1. ST and SNT neighbourhoods of an observation. *Each observation x_i has a ST(x_i) neighbourhood (observations are acquired in a short time interval) and a SNT(x_i) neighbourhood (observations belong to different passes). Source: Leroux et al., 2018)*

Secondly, a robust metric of 'outlierness' which evaluates the degree of inconsistency between the yield of x_i and that of the observations in both its ST and SNT neighbourhoods is computed. This step enables to create a bivariate plot of 'outlierness' which reports, on the x-axis, the 'outlierness' of each observation with regard to its SNT neighbours and, on the y-axis, the 'outlierness' of each observation with regard to its ST neighbours (Fig. 2, left). For instance, an observation in the top-right hand corner of the plot has a higher yield value than both its ST and SNT neighbours. Similarly, an observation in the bottom-left hand corner of the plot has a lower yield value than both its ST and SNT neighbours. Finally, a density-based clustering algorithm, i.e. DBSCAN, is used to identify outlying observations in the bivariate plot of 'outlierness' according to an automatic thresholding (Fig. 2, right).

Author-produced version of the article published in Computers and Electronics in Agriculture, 2019, N°156, p. 645-659.
The original publication is available at https://www.sciencedirect.com
Doi: 10.1016/j.compag.2018.12.024

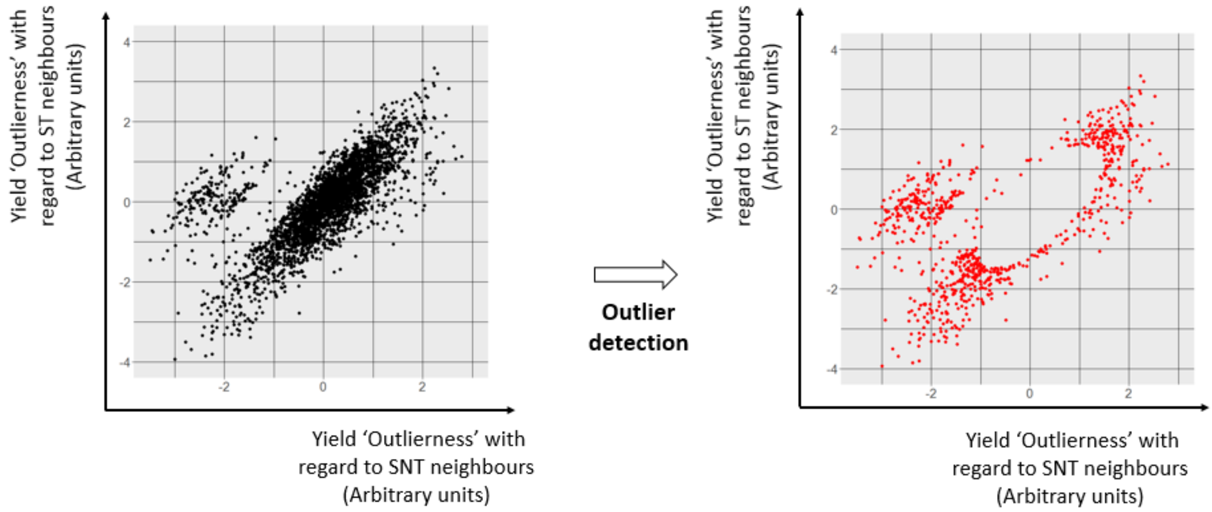

Figure 2. Left – An example of bivariate plot of 'outlierness' with all the observations (black dots on the online version). Right – An example of bivariate plot of 'outlierness' with solely defective observations identified by the method of Leroux et al. (2018) (red dots on the online version).

2.1.2 Making value of the available expertise on yield defective observations

For the past twenty years, there has been a considerable amount of work towards the understanding of the sources of defective observations in yield datasets (Arslan, 2002; Lyle et al. 2013). The latter authors have provided users with a categorization of yield technical errors into four major groups: (i) harvesting dynamics of the combine harvester, e.g. lag time, filling and emptying times, (ii) continuous measurements of yield and moisture, e.g. global/local yield and moisture outliers, (iii) accuracy of the positioning system, e.g. loss of signal, observations outside the field boundaries and, (iv) harvester operator, e.g. speed changes, unknown cutting width when entering the crop, harvest turns, narrow finishes (Lyle et al. 2013). All these errors, except those related to the positioning system, originate changes in the yield value of each defective observation. Given that the approach of Leroux et al. (2018) evaluates the yield outlying characteristics of each observation with respect to its spatial neighbours (ST and SNT) and that each type of error originates specific yield variations, these errors should theoretically have a specific location within the bivariate plot of 'outlierness'.

Given the available knowledge with respect to these defective observations, let us infer the location of these main yield technical errors within the bivariate plot of 'outlierness' (Fig. 3). Filling and emptying times induce a yield underestimation at the beginning and end of each harvest row either because the grain flow has not reached a plateau or because the grain still continues to flow while the header is up. It can be therefore considered that the yield of an observation acquired during these periods of time should not be consistent with that of both ST and SNT neighbours. Filling and empting times should mainly lead to observations located on the bottom left-hand corner of the plot, i.e. bottom and left-hand because this observation should have a lower yield value than both ST and SNT neighbours (Fig. 3). However, it must be said that at the end of the filling time or at the beginning of the emptying time, the grain flow is still relatively close to the permanent regime of the machine. This aspect means that some of these outlying observations might have a higher yield than that of the outlying observations at the beginning of the filling time or at the end of the emptying time. As such, it might be possible to also find (in a relatively small proportion though) outliers related to filling and emptying time in the top right-hand corner of the plot (Fig. 3). Another specification could be added. It has been shown that the underestimation was stronger at the beginning than at the end of the row (Simbahan et al. 2004). As such, observations collected at the end of a harvest row should be closer to the centre of the plot than observations collected at the beginning of the row.

The accuracy of yield and moisture sensors along with local harvest circumstances can influence the accuracy of yield measurements (Lyle et al. 2013). It might happen that yield records are effectively much higher or lower than expected and consequently that they significantly vary from those of their ST and SNT neighbours. Abnormal higher values should therefore be located on the top right-hand side of the bivariate plot of 'outlierness' while abnormal lower values should appear on the bottom left-hand side of the plot (Fig. 3).

Author-produced version of the article published in Computers and Electronics in Agriculture, 2019, N°156, p. 645-659.
The original publication is available at https://www.sciencedirect.com
Doi: 10.1016/j.compag.2018.12.024

Speed changes originate yield under or overestimates depending on if the speed of the harvester increases or decreases. In fact, during a speed change, the considered harvested area is flawed which impacts the quality of the resulting yield records and creates yield biases with respect to their ST and SNT neighbours. Accelerating would cause the observations to be located on the bottom left-hand part of the plot (a similar grain flow for a larger harvested area originates a decrease in yield) while a speed reduction should lead to observations appearing on the top-right hand side of the bivariate plot of 'outlierness' (Fig. 3).

Unknown cutting width, harvest turns and narrow finishes lead to strong yield underestimates because the harvested area is much lower than actually considered. However, in that case, the underestimation is propagated throughout the whole section of the row harvested under these conditions. In other words, it means that yield records are lower than those of their SNT neighbours but are consistent with their ST neighbours. All these observations should therefore be located in the left-hand portion of the plot but relatively close to the horizontal axis (Fig. 3).

It must be understood that Figure 3 is theoretical and has been created with the available knowledge on the main yield technical errors. The location of these errors will be validated later on with simulated and real datasets. Note that this figure could be complemented with other sources of defective observations and might help see interesting trends in the data.

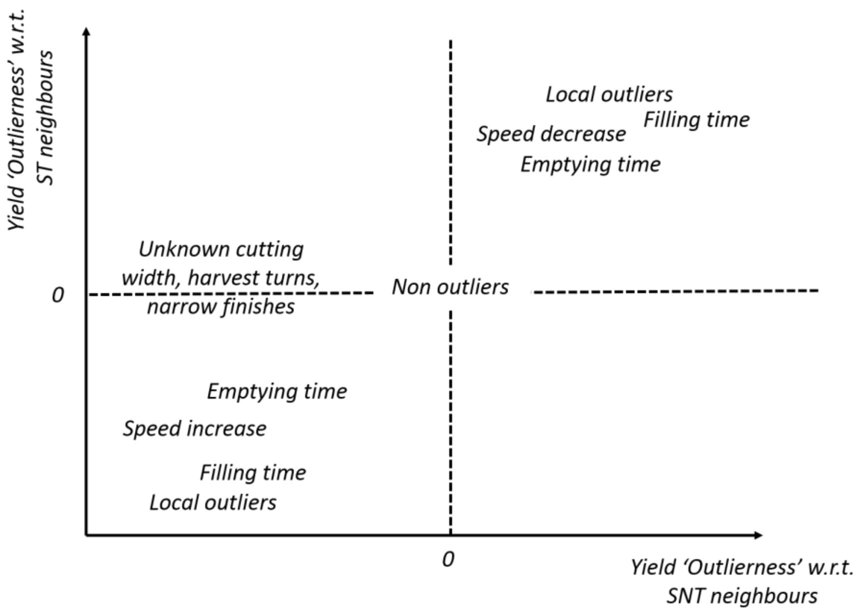

Figure 3. Theoretical location of the main sources of yield technical errors on the bivariate plot of 'outlierness' of Leroux et al. (2018).

2.2 Finding knowledge in outliers

The objective of the present study is to intend to explain why the outliers diverge from the rest of the population so that users can decide what to do with these defective observations. In this study, it is proposed to deal with these outliers using a two-step process: (i) the clustering of outliers so that defective observations that behave similarly are gathered, (ii) the categorization of outliers which aims at providing firstly a label to the clusters of outliers and secondly a label to the outliers within each considered cluster. These steps are described in the two following sections. A flowchart of the proposed methodology is proposed in Figure 4.

Author-produced version of the article published in Computers and Electronics in Agriculture, 2019, N°156, p. 645-659.
The original publication is available at https://www.sciencedirect.com
Doi: 10.1016/j.compag.2018.12.024

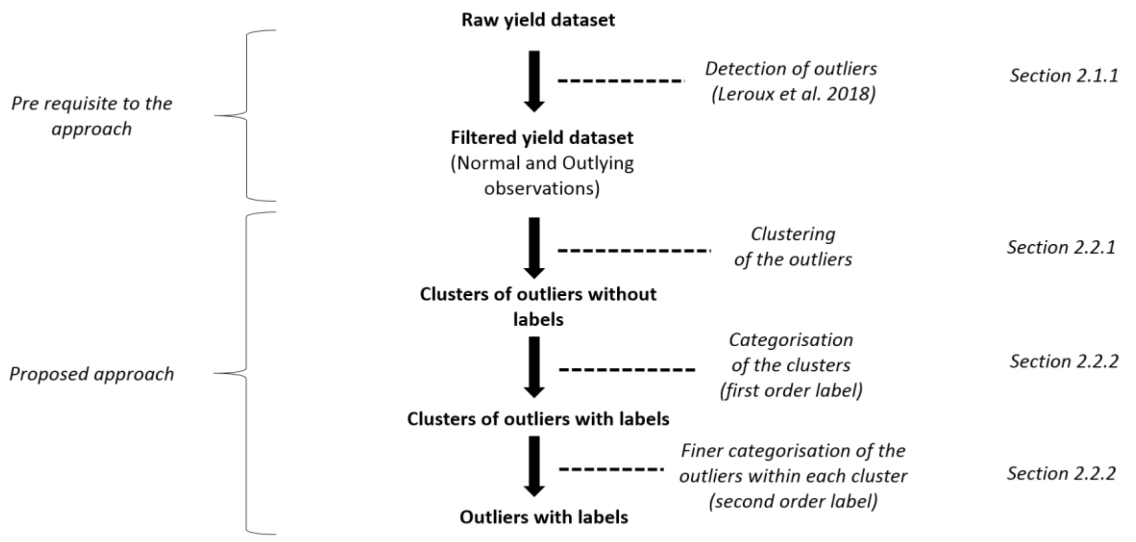

Figure 4. Flowchart of the methodology.

2.2.1 Automatic clustering of outliers

In the bivariate plot of 'outlierness', yield defective observations are clustered in specific portions of the plot (Fig. 2). In this study, an automatic clustering of observations is proposed because it is considered beneficial for the future labelling of observations. Indeed, within each cluster, observations share the same yield outlying characteristics with respect to their spatial neighbours. Grouping observations might help depict general trends or behaviours in these data. To automate the clustering of outliers, an angle-based methodology was put into place. For each outlying observation x_i, the angle that is formed between the horizontal axis and the vector $\overrightarrow{Ox_i}$ was computed; O being the point of coordinates (0,0) in this plot (Fig. 5, left).

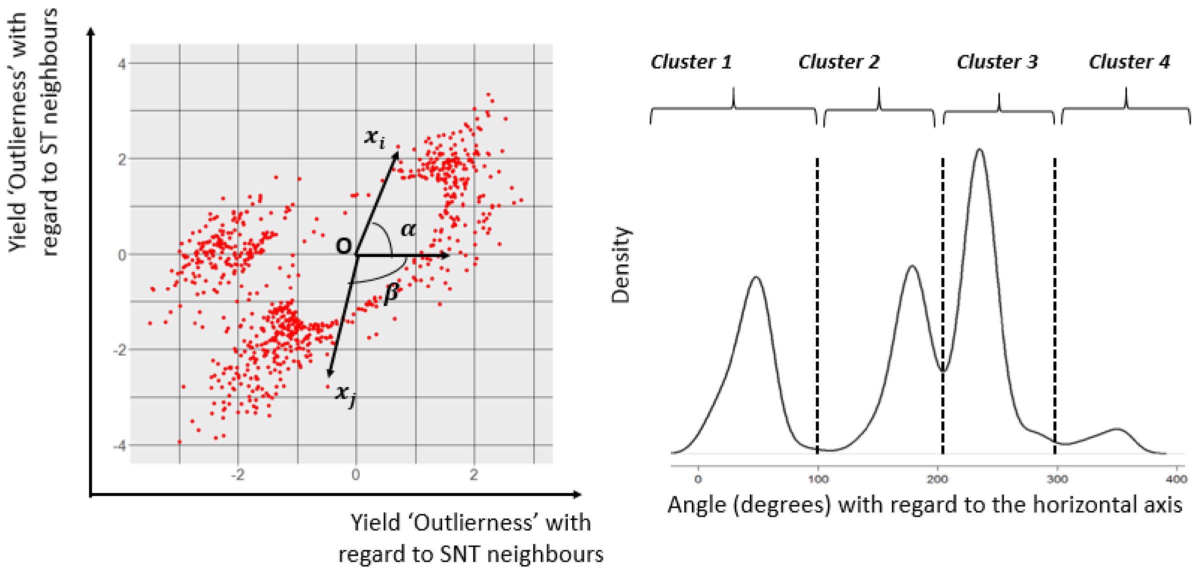

Figure 5. Left – Location of outliers using an angle-based methodology. Right – Clustering of outliers. *Outliers x_i and x_j have respectively an angle α and β with respect to the horizontal axis.*

A kernel density estimation (KDE) was then used to model the distribution of angles within the plot (Fig. 5, right). The number of clusters was chosen as the number of local minima in this distribution (Fig. 5, right). Each cluster was then set to contain all the observations lying between two consecutive local minima (Fig. 5, right). Within this methodology, an attention was paid to avoid the discrepancy between 360° and 0° (observations with these angles would be put in different cluster).

Author-produced version of the article published in Computers and Electronics in Agriculture, 2019, N°156, p. 645-659.
The original publication is available at https://www.sciencedirect.com
Doi: 10.1016/j.compag.2018.12.024

2.2.2 Categorization of outliers

Labelling the clusters of outliers: the first-order label

As the bivariate plot of 'outlierness' solely relies on the yield attribute, each cluster of outliers contains observations that have similar yield outlying characteristics with respect to their ST and SNT neighbours. As a primary description, these clusters can therefore be associated with a first-order label related to the yield component which expresses how this behaviour diverges from that of the cluster of normal observations (Fig. 6). The first-order label regarding the ST and SNT neighbours will be referred to as *Yield ST* and *Yield SNT*. For this first-order label, three classes are provided:

(i) "Low" if the 'outlierness' of the centroid of a cluster of outliers is less than the first 20th percentile value of the distribution of the 'outlierness' values of the cluster of normal observations, e.g. *Low Yield SNT*,

(ii) "Average" if the 'outlierness' of the centroid of a cluster of outliers lies between the first 20th and last 80th percentile values of the distribution of the 'outlierness' values of the cluster of normal observations, e.g. *Average Yield ST*,

(iii) "High" if the 'outlierness' of the centroid of a cluster of outliers is more than the last 80th percentile value of the distribution of the 'outlierness' values of the cluster of normal observations, e.g. *High Yield ST*

For instance, in Figure 6, cluster n°3 is given the following first-order label: "*Low Yield SNT and Average Yield ST*".

Labelling the outliers inside each cluster: the second-order label

However, this first-order label might not be sufficient to discriminate each type of error, especially if some of the errors induce similar yield changes with respect to the ST and SNT neighbours of outliers. For instance, in Figure 3, even if the location of errors is theoretical, multiple sources of errors might risk to be mixed up. As such, within each cluster, the objective was also to propose a second-order label so that each defective observation could be identified more clearly (Fig. 6). To do so, a set of attributes, different from the yield component, was chosen to improve the labelling of outliers. The selection of these attributes was driven by the available knowledge on yield defective observations and by the typicity of spatial observations collected from on-the-go vehicle-based datasets, i.e. yield datasets in that case. Before introducing these attributes, one may question why these variables were not taken into account directly within the process of detecting of outliers. Those reasons are multiple. First, incorporating several new variables makes the detection of outliers more difficult because those defective observations are likely to have outlying characteristics with respect to one variable but not with respect to others. This problem is also referred to as the curse of dimensionality (Beyer et al., 1999). Secondly, multiple attributes are used to compute the yield, e.g. speed, grain flow, width of the cutting bar, which means that if the values of these attributes were to be abnormal, this should be reflected on the yield records. It can also be added that, given the expertise and knowledge available on yield technical errors, it might be better to first detect outlying observations and then to try to explain their origin. Finally, it could be argued that yield datasets are often in different formats and do not necessarily contain the same attributes which may be problematical for creating a general methodology to detect outliers. Something certain is that they contain at least the basic information required to compute the yield.

For each observation x_i, three features were selected: (i) the change in speed between x_i and the previously collected observation x_{i-1} (*Var_Speed*), (ii) the spatial distance between x_i and the nearest harvest pass (*SpDist*) and, (iii) the number of ST neighbours of x_i (N_{ST}). The numbers of ST neighbours were evaluated within a distance of twice the length of the cutting bar. The attribute *Var_Speed* was selected because it should help discriminate the outliers that arise from an abrupt speed change. *SpDist* should bring insight into the operator-based outliers, e.g. narrow finishes, unknown cutting width when entering the crop, harvest turns because those types of errors are very often located close to adjacent passes. Finally, N_{ST} could be helpful to label delay-based outliers as these latter are expected to have lower ST neighbours than the remaining dataset, i.e. these errors are located at the beginning and end of harvest rows.

Author-produced version of the article published in Computers and Electronics in Agriculture, 2019, N°156, p. 645-659.
The original publication is available at https://www.sciencedirect.com
Doi: 10.1016/j.compag.2018.12.024

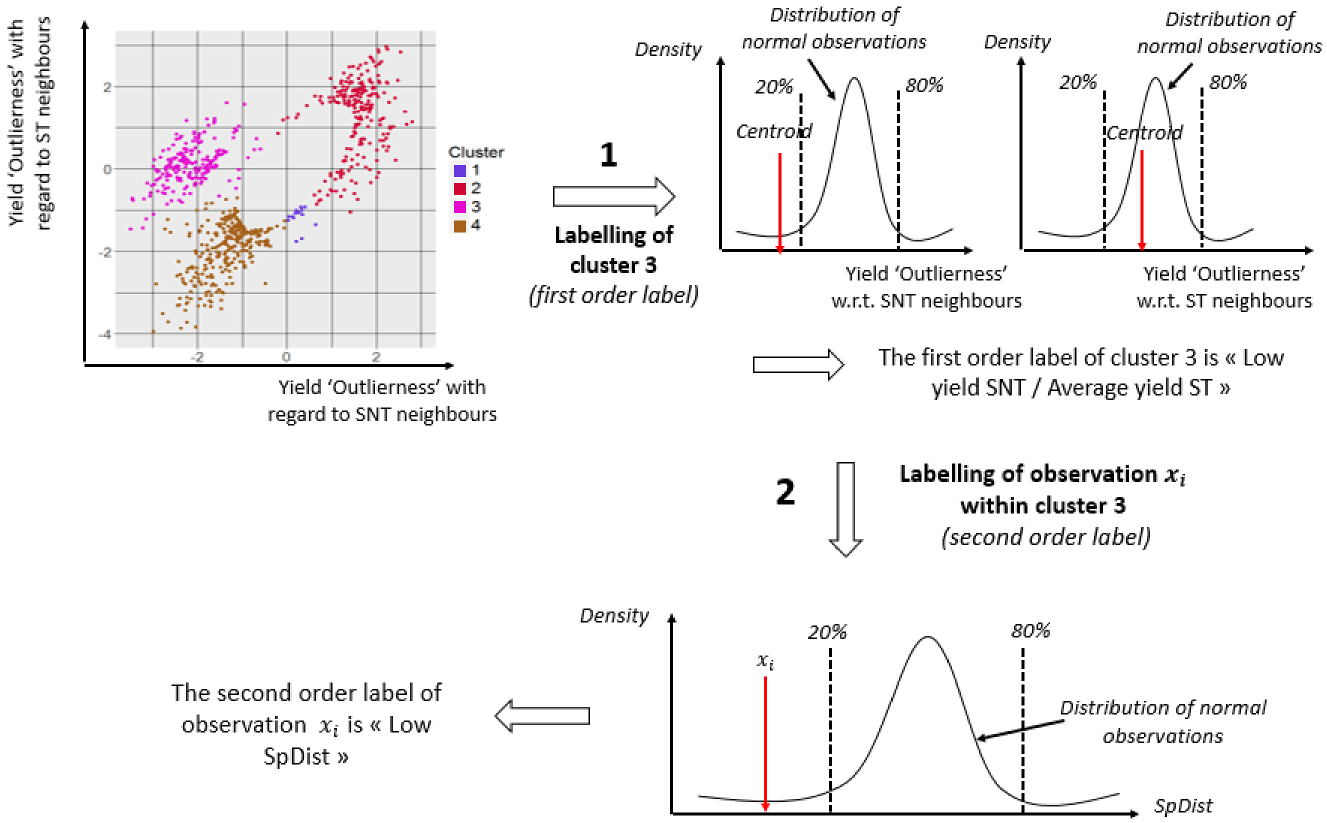

Figure 6. An example of the proposed methodology to label outliers in cluster n°3. *First, each previously defined cluster is given a first-order label. Then, within this cluster, each outlier is given a second-order label.*

To improve the labelling of outliers inside each cluster, the second-order (*Var_Speed*, *SpDist*, N_{ST}) labels were compared to those of the cluster of normal observations (Fig. 5). More specifically, for each attribute, three classes were provided:

(iv) "Low" if the attribute value of the outlier within the considered cluster is less than the first 20th percentile attribute value of the distribution of normal observations, e.g. *Low SpDist*,

(v) "Average" if the attribute value of the outlier within the considered cluster lies between the first 20th and last 80th percentile attribute values of the distribution of normal observations and, e.g. *Average SpDist*,

(vi) "High" if the attribute value of the outlier within the considered cluster is more than the last 80th percentile attribute value of the distribution of normal observations, e.g. *High* N_{ST}

For instance, in Figure 6, the observation x_i within cluster n°3 is given the following second-order label: "*Low SpDist*". Given the first- and second order labels that were put into place, Figure 3 can be improved to provide a classification of the main sources of errors as proposed in Figure 7. An accuracy index was put into place to evaluate whether the proposed classification was able to provide accurate labels to the defective observations. Considering a first and a second-order label, the accuracy index is the ratio of the number of true labels to the number of total observations labelled. In other words, the accuracy reports on the ability of the decision rules to identify a given type of observations. Once again, Figure 3 and Figure 7 are theoretical but will be tested and validated on simulated and real yield datasets. Be aware that all the outlying observations that would not be labelled with our proposed methodology, i.e. that do not belong to our theoretical clusters (Fig. 7), will be solely considered as local outliers.

Author-produced version of the article published in Computers and Electronics in Agriculture, 2019, N°156, p. 645-659.
The original publication is available at https://www.sciencedirect.com
Doi: 10.1016/j.compag.2018.12.024

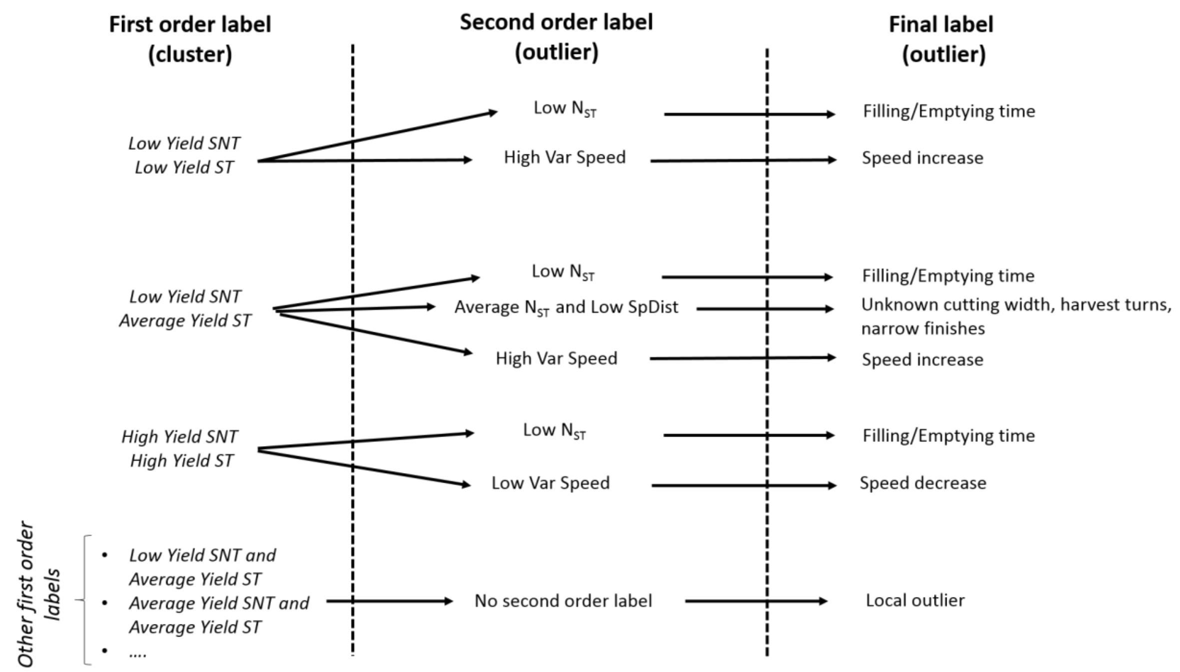

Figure 7. Decision rules to label outlying observations in within field yield datasets.

2.3 Evaluation of the proposed approach

2.3.1. Simulated datasets with labelled outliers

The approach was first validated on simulated datasets with known outliers' labels. The first objective was to locate the main types of outliers within the bivariate plot of 'outlierness' to see whether they could be differentiated. The second goal was to identify the most relevant features of these defective observations. Simulated yield datasets were generated according to the methodology of Leroux et al. (2017). The simulation process starts with the creation of a spatially structured yield dataset to which are added specific yield defective observations reported in the literature (Fig. 8). Yield datasets were created with a mean of approximately 7 T/ha. The magnitude of variation, represented by the coefficient of variation, CV, was set to 30%. Spatial structures (S) were modelled with exponential semi-variogram models. These datasets were set to contain 20% of outliers distributed between the different types of defective observations according to general findings in the literature and in personal datasets (Tab.1). Two yield datasets were simulated (Simu1 and Simu2), differing by the level of variance associated to the outliers to generate a diversity of case studies (Tab. 1) This variance can be understood as the influence of the outliers within the dataset. A low variance associated to the outliers would mean that outliers are relatively similar to their normal neighbours, and as such, are more difficult to identify (Simu1). On the contrary, a higher variance would mean that outliers have more diverging values from those of their normal neighbours (Simu2). In this case, outliers should be more easily identifiable.

Author-produced version of the article published in Computers and Electronics in Agriculture, 2019, N°156, p. 645-659.
The original publication is available at https://www.sciencedirect.com
Doi: 10.1016/j.compag.2018.12.024

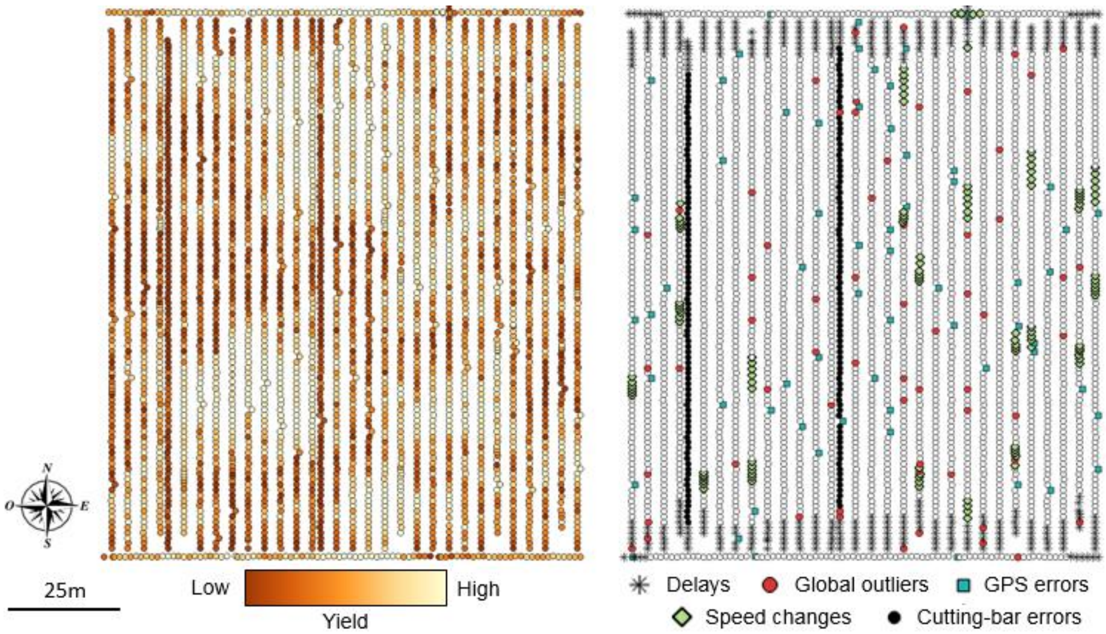

Figure 8. Example of yield simulated dataset (left) along with corresponding simulated errors (right). *These datasets were generated according to the methodology described in Leroux et al. (2017).*

2.3.2 Real datasets with non-labelled outliers

The proposed approach was then tested on four real yield datasets from fields located near Evreux in the North-western part of France. Fields were cropped in wheat and harvested with combines of different brands, especially New Holland (Turin, Italy) and Claas (Harsewinkel, Germany) combines. These datasets were selected for containing (i) different sorts of suspicious observations and (ii) outliers in different proportions (Tab. 2). Indeed, the filtering approach of Leroux et al. (2018) identified between 15 and 48% of outliers in the datasets. Defective observations were found responsible for lowering the mean yield and substantially increase the variability (CV) and skewness of the yield distribution (Tab. 2). Dataset 1 was considered as a typical yield dataset with a strong yield spatial structure, well harvested with mainly delay-based errors. Dataset 2 contains a couple of rows harvested with a not fully-used cutting width in the centre of the field. Dataset 3 was chosen because the wheel passages of a former fertilizer are very visible over the whole field and induced a decrease in yield. Dataset 4 contains two specific features. First, there are multiple narrow finishes within the field. Secondly, when entering the field, the width of the cutting bar was not set appropriately, i.e. lower than it actually was. This width was corrected after a few minutes inside the field. The objective was to see whether these specificities could be observed within the bivariate plot of 'outlierness' and labelled correctly.

Author-produced version of the article published in Computers and Electronics in Agriculture, 2019, N°156, p. 645-659.
The original publication is available at https://www.sciencedirect.com
Doi: 10.1016/j.compag.2018.12.024

Table 1. Description of the two simulated datasets Simu1 and Simu2 with their associated outlying yield observations. *Readers are referred to Leroux et al. (2017) for further details regarding the simulation process.*

		Yield technical errors				
		Filling and emptying times	**Sensor errors**	**GPS errors**	**Speed changes**	**Not fully-used cutting bar**
Amount of errors (Percentage of the total number of outliers)		50%	10%	10% (it can be single or groups of observations)	10%	20 % (all the observations inside a same harvest row are affected)
Simulated dataset	Simu1	Yield underestimation of 40% at the beginning and 20% at the end of the rows [B_k parameter in Leroux et al. (2017)]	20% noise	Lag of 10% of the inter-row distance	20% speed variation	80% of the cutting bar is used
	Simu2	Yield underestimation of 60% at the beginning and 40% at the end of the rows [B_k parameter in Leroux et al. (2017)]	50% noise	Lag of 20% of the inter-row distance	50% speed variation	50% of the cutting bar is used

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

Table 2. Descriptive statistics of the four raw and filtered real yield datasets.

Dataset	Surface (ha)	Raw dataset (with outliers)			Filtered dataset (without outliers)			
		Mean ($t.ha^{-1}$)	CV (%)	Skewness	Mean ($t.ha^{-1}$)	CV (%)	Skewness	Outliers detected (%)
1	14.5	7.1	28.1	-0.6	7.5	12.1	0.1	15.3
2	20.5	7.74	34.1	6.5	8.3	10.2	-0.3	32.5
3	30.9	9.6	28.7	8.4	9.9	6.0	-0.3	34.5
4	2.2	8.7	47.3	0.05	9.5	9.1	-0.5	48.7

3. Results and discussion

3.1. A first insight into the simulated datasets

3.1.1 Location of labelled outliers within the bivariate plot of 'outlierness' for simulated datasets

In simulated datasets, the label of each observation, and more especially that of the outliers along with the type of defective observation, is known. This means that it is possible to locate each outlier generated in the bivariate plot of 'outlierness' to explore how they behave. This will also be a first way to evaluate the veracity of the theoretical location of yield technical errors that was provided in Figure 3. Figure 9 displays the location of all the observations in the simulated datasets within the bivariate plot of 'outlierness' (Fig. 9, left) and also that of all simulated outliers (Fig. 9, right). The two simulated datasets Simu1 and Simu2 are presented. From a general standpoint, it appears that several defective observations have effectively a specific location within the bivariate plot of 'outlierness'. This position within the plot appears relatively consistent with what was proposed in Figure 3 but is also much fuzzier. The bivariate plot of 'outlierness' appears to be clearly impacted by the level of divergence between normal and defective observations (Fig. 9, right). Indeed, for Simu2, one can distinguish much more easily several groups of observations through a visual inspection of the plot. It can be noted that when outliers are more deeply rooted in the dataset, i.e. outliers are more similar to normal observations (Simu1), the bivariate plot of 'outlierness' seems more homogeneous without strong deviations from the centre of the plot.

As expected, observations collected with a not-fully used cutting bar ("Cutting Width") are mostly located on the left-hand part of the plot, i.e. they are very consistent with their ST neighbours but exhibit relatively different values to those of their SNT neighbours (Fig. 9, right). Local outliers can be spotted on the extremities of the plot, i.e. on the bottom-left hand corner because, in this simulated dataset, global outliers were generated with a low yield value. Suspicious observations collected within the filling and emptying time periods, or during a speed change appear on the main diagonal of the plot. However, several observations of these last types of error appear also near the centre of the plot of "outlierness" (Fig. 9, right). The thing is that all the outlying observations do not have the same influence on the dataset quality. For instance, within the filling time period, the underestimation associated to the first few observations will be much stronger than that associated to the last observations collected during this filling time. The primary observations within the filling time will therefore strongly deviate from the normal population while the last ones will be much closer to the distribution of the normal population. To put it simple, observations with the major impact on the yield local distribution will be located far from the centre of the bivariate plot of 'outlierness'. Finally, by observing more carefully the shape drawn by the outliers, it seems that several populations can be depicted within the plot. Indeed, it seems possible to fit straight lines with similar slopes but different intercept, especially for the observations collected with a not-fully used cutting bar with respect to the rest of the data. The change of cutting width, which originated a strong decrease in the yield values, have produced a substantial change in the yield distribution of these specific outliers that is highlighted by a shift in the bivariate plot of 'outlierness'.

Author-produced version of the article published in Computers and Electronics in Agriculture, 2019, N°156, p. 645-659.
The original publication is available at https://www.sciencedirect.com
Doi: 10.1016/j.compag.2018.12.024

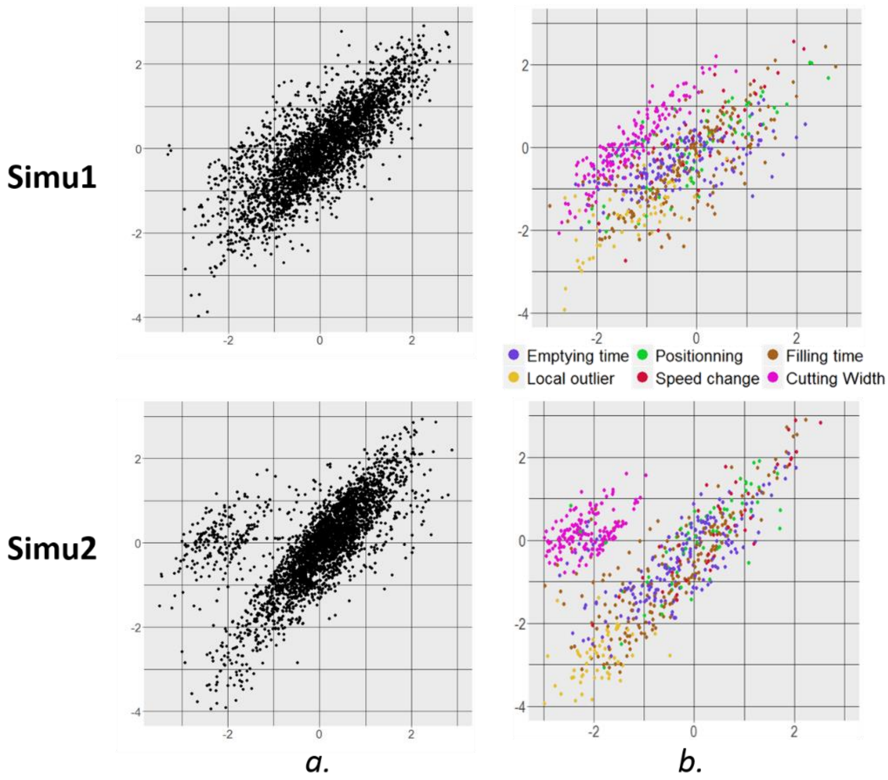

Figure 9. Location of simulated-based outliers in the bivariate plot of 'outlierness'. In Simu1, outliers are relatively similar to normal observations. In Simu2, outliers are more diverging from normal observations. *a. Unlabeled observations. b. Labelled observations*

3.1.2 Automatic detection and clustering of outliers in simulated datasets

Figure 10 reports how outliers are handled by the proposed approach, i.e. detection of outliers (Fig. 10, left) and clustering of outliers (Fig. 10, right). First, it can be seen that multiple outliers are not detected by the approach of Leroux et al. (2018). As discussed in the previous section, these suspicious observations appear near the centre of the plot where observations are considered normal in the aforementioned methodology. These outliers, i.e. much more similar to the normal observations, are more difficult to detect and can be referred to as false-negative outliers. From a practical standpoint, by considering the example of the delay-based outliers, it is much more important to remove the observations at the beginning of the filling period, i.e. outliers that lay far from the centre of the plot, than to remove those when the filling time is almost finished, i.e. outliers that are located in the centre of the plot. In other words, it is more interesting to focus on removing the variance associated to the outliers than a specific number of defective observations.

By using the proposed angle-based methodology, several groups of outliers were identified automatically (Fig. 10, right). The proposed approach has generated two and three major clusters for the simulated datasets Simu1 and Simu2. It appears that this delineation comes out more robust when outliers are relatively different to the normal population, i.e. Simu2 (Fig. 10, bottom right). In this case, clusters effectively correspond to major sources of yield errors. For Simu1, cluster n°5 seems relatively wide as it gathers several types of outliers. However, for Simu1, the relative consistency that exists between outliers and normal observations makes it difficult to properly split the cluster by solely relying on the yield attribute. In both simulated datasets, relatively small clusters are being identified, e.g. clusters n°1,2 and 4 for Simu1. Those clusters will not carry much information as they contain very few data.

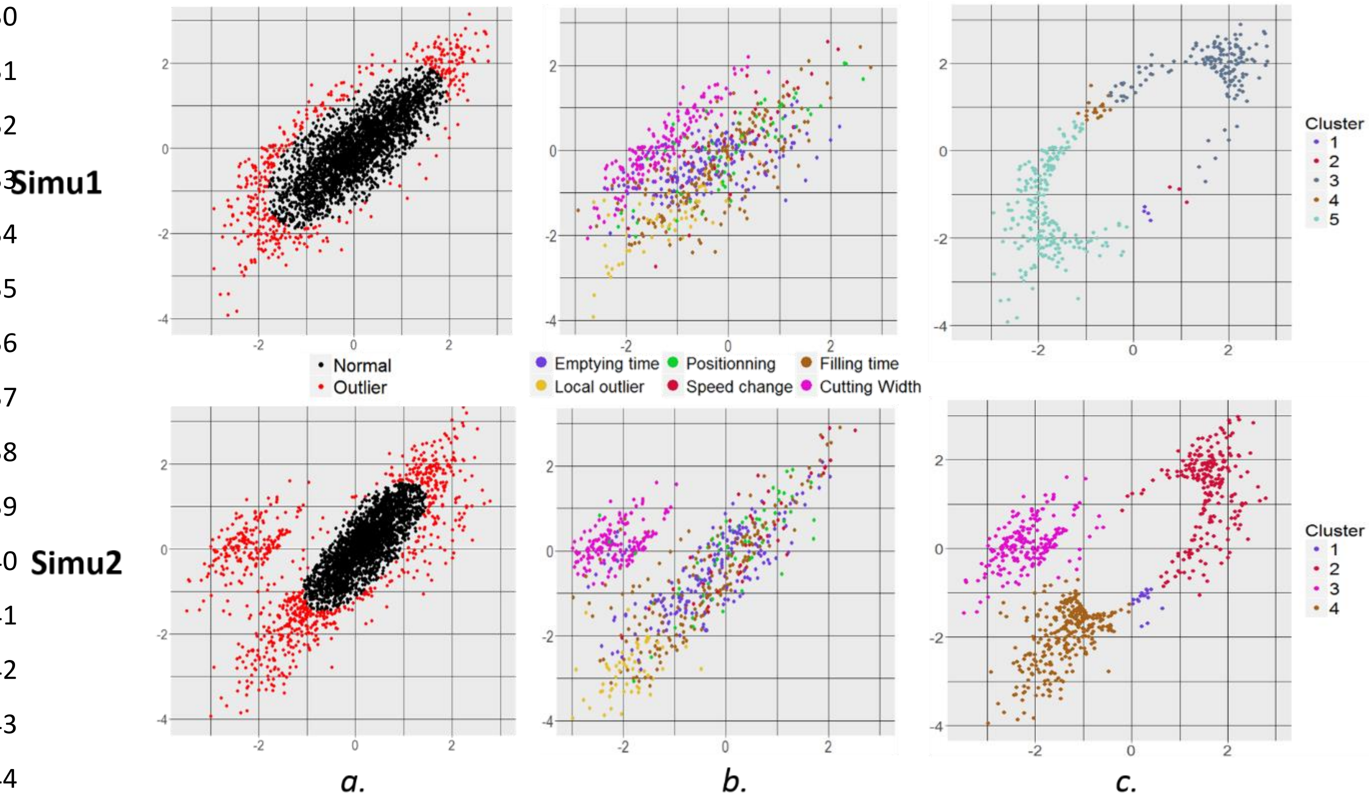

Figure 10. Analysis of outliers in the simulated datasets with low outlier variance (top) and moderate outlier variance (bottom). *a. Identification of outliers by the approach of Leroux et al., (2018). b. Known labelling of the outliers within the dataset. c. Clustering of the outliers within the dataset*

3.1.3 Labelling of outliers in simulated datasets

For both simulated datasets, Table 3 reports: (i) the first-order label associated to each cluster, (ii) the second-order label associated to each outlier within the previously-defined clusters, (iii) the final label associated to each outlying observation and (iv) the accuracy of the labelling (ratio of the number of true labels to the number of total observations labelled). Unsurprisingly, it appears that the labels' accuracy is better for Simu2 because in that case, outliers are more different to the normal population. From a general perspective, the labelling of outliers is relatively accurate. Note that when the first-order label did not match any of the theoretical outlying clusters that were proposed in Figure 7, the respective outlying observations were simply labelled as local outliers. As all types of errors can be considered as specific forms of local outliers, the accuracy index for the label *local outliers* does not make much sense.

Be aware that the labelling is generally bad when it comes to detecting observations acquired during the filling and/or emptying times with a first-order label "High yield ST/SNT" (top right-hand corner of the bivariate plot of 'outlierness'). For the remaining clusters, the accuracy is relatively high enough meaning that the outlying observations can be automatically labelled given the first- and higher-order labels that are provided. This classification, i.e. that of Figure 7, will therefore be used to analyze the real datasets (see next section).

Table 3. Labelling of outliers in simulated datasets.

Simulated Dataset	Cluster	First-order label	Second-order label	Final label	Accuracy (%)
Simu1	1	Average yield SNT Low yield ST	-	Local outliers	-
	2	Average yield SNT Average yield ST	-	Local outliers	-
	3	High yield ST/SNT	Low N_{ST}	Filling/Emptying times	5
			Low *Var_Speed*	Speed decrease	35
	4	Average yield SNT Average yield ST	-	Local outliers	-
	5	Low yield ST/SNT	Low N_{ST}	Filling/Emptying times	45
			High *Var_Speed*	Speed increase	100
Simu2	1	Average yield SNT Low yield ST	-	Local outliers	-
	2	High yield ST/SNT	Low N_{ST}	Filling/Emptying times	13
			Low *Var_Speed*	Speed decrease	78
	3	Low yield SNT Average yield ST	Low N_{ST}	Filling/Emptying times	61
			Average N_{ST} and Low *SpDist*	Partially-used cutting bar	97
			High *Var_Speed*	Speed increase	100
	4	Low yield ST/SNT	Low N_{ST}	Filling/Emptying times	91
			High *Var_Speed*	Speed increase	80

It must be understood that, here, the accuracy shows whether an outlying observation is given a good final label considering the first and second-order labels that are defined. However, it does not specify if, within the whole dataset, all the observations that should have been given a specific label actually received it. For instance, one can be pretty sure that the outlying observations in Simu2 that were given the label "Partially-used cutting bar" are observations that were collected when the width of the cutting bar was not used entirely. Nevertheless, one cannot be entirely sure that all the observations collected with a partially-used cutting bar were found in the whole dataset.

To provide users with a more comprehensive overview of the reliability of each label, the ratio between accurate labelled outliers and the total number of outliers of each type in the whole dataset is presented in Table 4. As should be expected, ratios are lower for Simu1 than for Simu2 given the construction of both datasets. From a general perspective, by looking at Table 4, ratios seem to be relatively low, especially for Simu1. Note also that no observations collected with a partially-used cutting bar could be found in Simu1 given the clusters that were identified in Figure 10 and the associated labelling rules. Obtaining relatively low ratios should not be very surprising given that several outliers were not identified by the filtering approach of Leroux et al. (2018), i.e., those are located near the centre of the bivariate plot of outlierness. As the labelling procedure solely labels observations that were identified as outliers, not all the outliers could be labelled. Be aware that the ratios would have been higher if solely the detected outliers had been considered (and not all the outliers in the dataset). On top of that, it must be clear that those ratios represent solely a percentage of outlying observations and do not convey any information regarding the variance associated with these outliers. For instance, only 43.5% of the observations acquired during a filling or emptying time were correctly labelled for Simu2 but those observations accounted for most of the variance associated to the filling/emptying time label (data not shown). The outlying observations near the centre of the bivariate plot of outliers (that were not labelled) are not that different from their neighbours (the

Author-produced version of the article published in Computers and Electronics in Agriculture, 2019, N°156, p. 645-659.
The original publication is available at https://www.sciencedirect.com
Doi: 10.1016/j.compag.2018.12.024

influence of these outliers is expected to be relatively low) while those far away from the centre of the plot are much more influencing (Fig. 10b). This last statement echoes some of the points that were addressed in section 3.1.1 where it was discussed that not all the outlying observations had the same influence on the quality of the dataset. The same reasoning can be applied to the other outlying observations, e.g. those collected during a speed change. Indeed, some very slight speed changes can also be found near the centre of the bivariate plot of outlierness (Fig. 10b). From a general perspective, the labelling outputs on the simulated datasets necessarily depend on the way yield datasets were simulated (Leroux et al., 2017).

Table 4. Reliability of the labelling in simulated datasets. *The table presents the ratio between accurate labelled outliers and the total number of outliers of each type.*

	Filling/emptying times	Speed changes	Partially used cutting bar
Simu1	16.8	8.7	0
Simu2	43.5	18.6	91.1

3.2. Clustering and labelling of suspicious observations in real yield datasets

The defective observations that were identified in the real datasets by the filtering approach of Leroux et al., (2018) are depicted and clustered in Figure 11. First of all, it can be seen that the structure of the bivariate plots of 'outlierness' shares many similarities with that of the simulated datasets (Fig. 11, left and middle). More specifically, multiple observations expand towards either the top-right, bottom-left or left-hand part of the plot. This aspect is satisfying because it proves the interest of the theoretical data to study and help label outliers. Note that all the datasets seem to have similar types of clusters (the angles that are formed between the cluster and the horizontal axis are very similar). Each dataset also has its own specificities as the number of outliers' clusters varies across the yield datasets, from two to five main clusters between datasets 1 and 5. These groups of outliers are relatively well identified especially for datasets 1, 2 and 4. The delineation of the clusters appears more abrupt for dataset 3, e.g. for instance between clusters n°3 and 4, but there effectively seems to be two different populations in the data. It is acknowledged that the clustering using the proposed angle-based approach can be considered quite brutal at the edges of the outliers' clusters. Some confusion might effectively remain, but it must be noted that the main groups of outliers are being spotted. Interestingly, the aspect of different statistical yield distributions that was previously discussed with respect to simulated datasets, i.e. the impression of parallel straight lines that could be fitted to the data, is particularly visible on dataset 4.

Author-produced version of the article published in Computers and Electronics in Agriculture, 2019, N°156, p. 645-659.
The original publication is available at https://www.sciencedirect.com
Doi: 10.1016/j.compag.2018.12.024

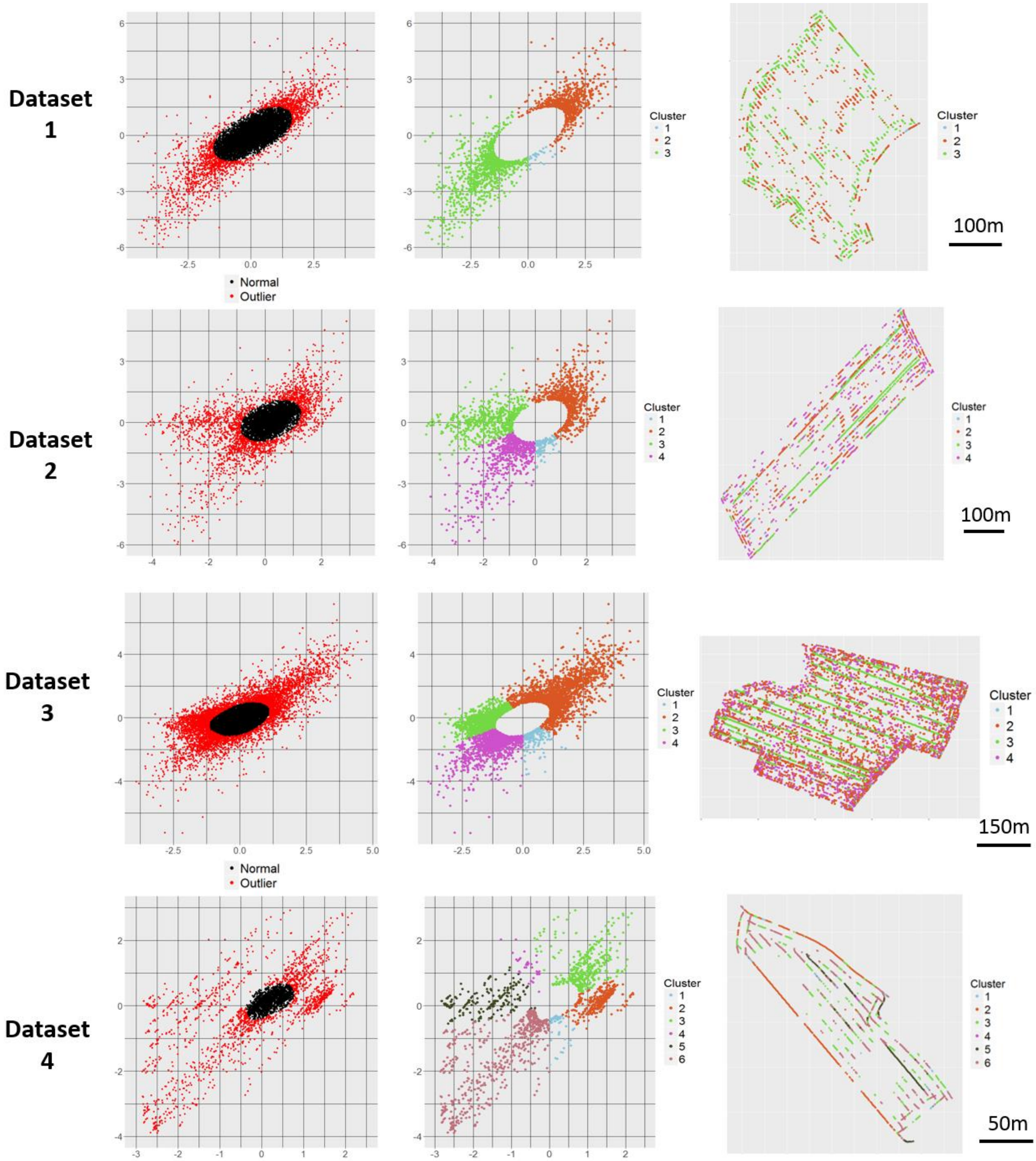

Figure 11. Labelling of outliers in the real yield datasets. *Left. Detection of outliers. Middle. Clustering of outliers. Right. Location of the clusters within the field.*

Given the findings in the simulated datasets and the location of each outlier's clusters within the field (Fig. 11, right), yield outliers could start being labelled. For the four datasets under study, the clusters located on the diagonal of the bivariate plot of 'outlierness' (Low *yield ST/SNT* and High *yield ST/SNT*) are relatively well identified. Observations inside these clusters were labelled as filling/emptying times, speed changes and local outliers following the decision rules that were used for the simulated datasets (Fig. 7). Regarding dataset 1, some observations lying within the clusters n°2 and n°3 appear to be located in the centre of the field. These observations,

Author-produced version of the article published in Computers and Electronics in Agriculture, 2019, N°156, p. 645-659.
The original publication is available at https://www.sciencedirect.com
Doi: 10.1016/j.compag.2018.12.024

that were labelled as local outliers according to the proposed methodology (data not shown), are in fact due to the presence of a change in soil conditions which originated a short-range variation in yield. These observations are therefore not outlying observations but rather expected yield records. Note that without a soil map, this distinction is relatively difficult to make.

In the case of simulated datasets, the cluster on the left-hand side of the plot (Low *yield SNT* and *Average yield ST*), i.e. cluster n°3, was mostly standing for observations collected with a low cutting width. This is why the second-order label "Low *SpDist*" was put into place for this specific cluster. However, when looking at the observations in cluster n°3 within dataset 3, many of these observations appear to be regularly spaced within the field, which is not particularly a feature of passes harvested with a low cutting width (Fig. 11). These observations could be spotted by the second-order label "Average *SpDist*". These observations were found to represent the wheel passages of a former fertilizer or other agricultural machinery. It must be clear that this labelling was not proposed in the initial labelling framework (Fig. 7). Without using the second-order label "Average *SpDist*", these regularly spaced observations would be given the final label 'local outliers'. To provide a better labelling of these observations, it was therefore decided to add a new rule to the labelling framework (Tab. 5). This rule was specifically applied to this dataset, but could certainly be used in a more general perspective in the proposed approach.

Figure 12. Analysis of cluster n°3 in datasets 2 and 3. *The attribute SpDist helps improve the labelling of observations inside this cluster.*

A last interesting aspect to consider was the relatively large cluster n°2 of dataset 4 that expands towards the right-side (High yield SNT and Average yield ST) of the bivariate plot of 'outlierness' (Fig. 11, dataset 4). In the case of dataset 4, these observations have effectively a somewhat questioning behaviour because they can be found mostly on the edges of the field. It was found that this cluster n°2 corresponded to the operator's error in setting the appropriate width of the cutting bar when he started harvesting the field. The cutting bar was effectively set lower than it actually was, which led to an overestimate of the yield (see material and methods section 2.3.2). This dataset enabled to propose an additional rule to the initially proposed labelling framework (Tab. 5). Here again, this rule was specifically applied to this dataset, but could certainly be used in a more general perspective in the proposed approach.

Table 6 sums up the results of the labelling process, i.e. an estimate of the proportion of each type of outlying observations, on the four real datasets using the initial labelling framework (Fig. 7) to which additional rules were joined (Tab. 5). These summary statistics are obviously not perfect and depend on the methodology that was used in this work. Be aware that global outliers (header up, zero yield values, very abnormal yield value…) are not accounted for in Table 6, because they were removed before the spatial outlier detection process in Leroux et al. (2018). Note also that some of these global outliers might have been labelled with one of the main sources of technical errors but these outliers were found so diverging from the normal population that they were removed prior to applying the spatial outlier detection algorithm. Table 6 highlights that all datasets are unique in the sense that they all have different outliers and those latter are present in different proportions. It must be reminded that the label "Local outliers" contains the outlying observations that could not be labelled in any of the other classes of technical yield errors. This is why the percentage of observations having this label is quite high. The labelling of filling and emptying time errors seems slightly low, especially for datasets 2 and 3, when comparing with the literature. This may be due to the removal of such errors with the global filter introduced in Leroux et al. (2018) or because some of these errors were mixed up with others and were labelled as local outliers.

Author-produced version of the article published in Computers and Electronics in Agriculture, 2019, N°156, p. 645-659.
The original publication is available at https://www.sciencedirect.com
Doi: 10.1016/j.compag.2018.12.024

Table 5. Additional decision rules arising from the analysis of the real yield datasets.

Dataset	Cluster	First-order label	Second-order label	Final label
3	3	Low yield SNT and Average yield ST	Low SpDist	Unknown cutting width
			Average SpDist	Wheel passage of a former fertilizer
4	2	High yield SNT and Average yield ST		Error in setting the width of the cutting bar
	5	Low yield SNT and Average yield ST	Low SpDist	Unknown cutting width / Narrow finishes

Table 6. Summary of the technical errors within each real dataset. *The total number of outliers is the sum of the number of each type of outliers.*

Dataset	Filling/Emptying time	Speed change	Unknown cutting width / Narrow finishes	Local outliers	Others (wheel passages, error in settings)	Total number of Outliers
1	4.8%	1.9%	-	8.6%	-	15.3%
2	1.5%	3.7%	5.5%	21.8%	-	32.5%
3	0.7%	5.4%	1.5%	19.6%	7.3%	34.5%
4	12.4%	2.8%	5.3%	18.6%	9.6%	48.7%

From a general perspective, by looking at the labelling rules that are proposed in this study (Fig. 7), one could suggest that the second label alone would be successful to separate each error. It is effectively acknowledged that the second-order labelling could be efficient in itself but it is also stressed that the clustering and first-order labelling of outlying observations have also their interest. First, it is clear that defective yield observations are clustered in specific portions of the bivariate plot of outlierness (Fig. 2, right; Fig. 10, Fig. 11). When looking at these figures, one might be very tempted to intend to group these outliers in terms of their yield behaviour with respect to neighbouring observations to see whether specific patterns can be identified. The approach to automatically split outlying observations in different clusters was done in that sense. Secondly, when focusing on the real yield datasets, it should become clearer that this first order labelling was relevant. In fact, for dataset 3, if cluster n°3 with the first-order label '*Low yield SNT* and *Average yield ST'* is not separated from the rest of the outlying observations, it would not have been possible to identify the wheel passage of a former fertilizer or agricultural machinery. Indeed, these observations have a second order label "Average SpDist". If this labelling rule was used on all the outlying observations, many specific observations would have been mixed. Same goes for cluster n°2 in dataset 4, the settings error in the cutting bar width would not have been clearly separated from the other types of outlying observations.

The proposed approach enables to provide users with a clearer interpretation and analysis of their yield datasets. Some of these results might be used to improve the quality of the datasets by correcting some of these errors instead of removing them (see next section). Another possibility would be to analyze the way operators drive within the fields (speed changes, operator-based outliers) or to characterize the functioning of the harvester. Economic considerations might also come up such as whether investing in systems that measure in real-time the width of the cutting bar is relevant if the outlying-related observations can be spotted and corrected. Once again, these results come along with a given uncertainty, but they might be used to depict general trends in the data. Be aware that the proposed method is a first attempt to provide a label to yield outliers. This approach can be sensitive to the thresholds that have been set, more especially the 20^{th} and 80^{th} percentile values that were used to help label the clusters of outliers and the outliers within each cluster. The choice of these thresholds would require further investigation. One possibility could be for instance to test the sensitivity of the method to the values of these thresholds through a Monte Carlo approach, but this is beyond the scope of this work. Note however that these thresholds are relatively easy to parametrized.

Author-produced version of the article published in Computers and Electronics in Agriculture, 2019, N°156, p. 645-659.
The original publication is available at https://www.sciencedirect.com
Doi: 10.1016/j.compag.2018.12.024

3.3 What can be done with the labelled outliers ?

When outliers are labelled and described with a proper subset of attributes, they become meaningful and understandable. As such, it becomes possible for users to make a decision with regard to these suspicious observations. Two major case studies can be observed. In the first case, outliers can be considered as noise meaning that these observations are not reflecting a real phenomenon and as a consequence should not have been generated. This noise can have multiple causes such as the process of data acquisition in itself, e.g. the pass of the combine harvester within the field, or a technical failure, e.g. loss of positioning signal. To tackle this noise, defective observations can be either corrected or removed. Performing a correction on a defective observation is conceivable when the phenomenon which originated the outlier is fully known and controlled. Here, it is suggested that, when possible, the correction should be preferred to the removal of outliers because the final dataset would contain more information and should therefore be more accurate. If the origin of an outlier is known but the accuracy of the correction could be questioned, the outlier should be removed to make sure the quality of the dataset is not affected. This was especially considered for technical errors such as speed changes or filling and emptying times which have a complex influence on the yield output. In the second case, the outlier might really shed light on a phenomenon of interest which could be either expected or unexpected. In such situations, users should be warned so that they can intend to get a deeper understanding of this specific phenomenon.

Here, the output of the processing that was applied to dataset 4 is displayed in Figure 13. In this case study, more specifically, a correction was applied to the outliers in clusters n°2 and 5 while other defective observations were removed. Indeed, most suspicious observations of cluster n°2 are due to bad settings in the cutting width of the harvester, which can be corrected by weighing the yield values with an appropriate factor depending on what was set by the operator (this information was available in the yield dataset). The outliers belonging to cluster n°5 especially reflect passes harvested with a low cutting width. For these specific observations, a weighing factor, related to the spatial distance to the previously harvested pass, can be applied to calculate the yield that should have been found with the portion of the cutting width that was used.

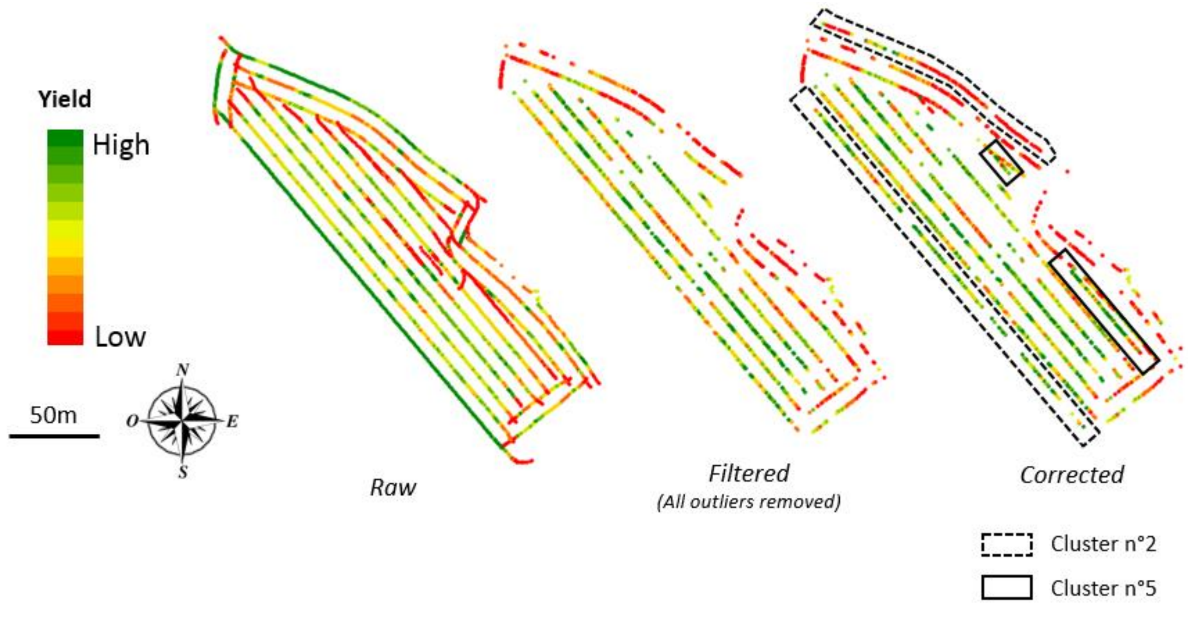

Figure 13. Making value of the labelling of outliers to propose a correction for dataset 4. *Dashed polygons contain the observations that were restored.*

This correction helped retrieve lots of yield observations within the dataset (almost 15%) to improve its quality and reliability (Fig. 13, right). Note for instance that most of the yield information on the edges of the fields were restored. However, it was decided not to propose any correction for the remaining clusters. One effectively knows the general impact that a speed change or the delay-time might cause on the yield attribute, i.e. an increasing or decreasing trend, but it is much more difficult to evaluate it precisely. Some convolution filters might be proposed to cope with that issue, but they were considered relatively complex to put into place as the parameters of the model convolution are not easy to define properly (Arslan and Colvin 2002). Nevertheless, it must be said that yield datasets contain quite a large amount of information which means that removing outliers is not too critical if a proper and accurate correction cannot be proposed. Be aware that this case study is an application example of the proposed methodology and that applying this methodology would require having a discussion with the operator to validate the origin of the errors.

4. Conclusion

This study proposes a methodology to cluster and label outlying observations in yield datasets after that these latter have been detected by a holistic and unsupervised filtering approach. Defective observations are first labelled in terms of yield characteristics with respect to their spatial neighbours. They are then further labelled with appropriate spatial and non/spatial attributes so that they can be classified more accurately into the main types of yield technical errors, e.g. filling/emptying times, speed changes, unknown cutting width when entering the crop, narrow finishes. While some observations are more accurately classified (speed changes or unknown cutting width), others are slightly more complex to be given an appropriate label (filling/emptying times). The proposed labelling approach also enabled to identify specific observations in real yield datasets, i.e. the wheel passages of a former fertilizer or agricultural machinery and settings errors in the cutting bar width. The proposed methodology provides users with a set of interpreted outlying observations which can then be used for multiple purposes: (i) understanding of the main sources of errors in each user's yield dataset, (ii) correction of the outliers instead of removing them if possible, (iii) characterization of the way the operator drives within the field or how the combine works during harvest, and (iv) provision of guidelines for future improvements of equipment and operations processes. Future work will focus on improving the ability of the proposed methodology to properly label outliers and testing the approach on more datasets, i.e. not only related to yield.

5. References

Angiulli, F., Fassetti, F., Palopoli, L. (2009). Detecting outlying properties of exceptional objects. *ACM Transactions on Database Systems*, 34(1):7.

Angiulli, F., Fassetti, F., Palopoli, L. (2012). Discoverying characterizations of the behavior of outlier sub-populations. *IEEE Transactions on Knowledge and Data Engineering*, 25, 1280-1292

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

Baluja, J., Diago, M., Goovaerts, P., & Tardaguila, J. (2012). Assessment of the spatial variability of anthocyanins in grapes using a fluorescence sensor: relationships with vine vigour and yield. *Precision Agriculture, 13*, 457–472.

Beyer, K., Goldstein, J., Ramakrishnan, R., Shaft, U. (1999). When is nearest neighbor meaningful*? In Proceedings of the 7 th ICDT*, Jerusalem, Israel.

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

Colaço, A.F., Rosa, H.J., Molin, J.P. (2014). A model to analyse as-applied reports from variable rate applications, *Precision Agriculture*, 15, 304–320, DOI 10.1007/s11119-014-9358-5

Debuisson, S., Germain, C., Garcia, O., Panigai, L., Moncomble, D., Le Moigne, M., Fadaili, E.M., Evain, S., Cerovic, Z.G. (2010). Using Multiplex® and Greenseeker™ to manage spatial variation of vine vigor in Champagne. *10th International Conference on Precision Agriculture*. Denver, Colorado, July 18–21, (www.icpaonline.org/finalpdf/abstracts.197.pdf

Duan L., Tang, G., Pei, J., Bailey, J., Campbell, A., Tang, C. (2015). Mining outlying aspects on numeric data. *Data Mining Knowledge Discovery*, 29, 1116–1151

Ertoz, L., Eilertson, E., Lazarevic, A., Tan, P., Srivastava, J., Kumar, V., Dokas, P. (2004). The MINDS - Minnesota Intrusion Detection System, in Data Mining, A. Joshi H. Kargupta, K. Sivakumar, and Y. Yesha (Eds.) Next Generation Challenges and Future Directions.

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

Knorr E. M., Ng R. T. (1999). Finding Intensional Knowledge of Distance-based Outliers. *In Proceedings of the 25th International Conference on Very Large Data Bases*, Edinburgh, Scotland, pp. 211-222

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.V. Stafford (Ed.), Advances in Animal Biosciences: Precision Agriculture (ECPA)*, 1-6.

Leroux, C., Jones, H., Clenet, A., Tisseyre, B. (2018). A general method to filter out defective spatial observations

Author-produced version of the article published in Computers and Electronics in Agriculture, 2019, N°156, p. 645-659.
The original publication is available at https://www.sciencedirect.com
Doi: 10.1016/j.compag.2018.12.024

from yield mapping datasets. *Precision Agriculture*. https://doi.org/10.1007/s11119-017-9555-0

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.

Marques, H.O., Campello, R.J., Zimek, A., Sander, J. (2015). On the internal evaluation of unsupervised outlier detection. *In Proceedings of the 27th International Conference on Scientific and Statistical Database Management* (SSDBM '15), Amarnath Gupta and Susan Rathbun (Eds.), ACM, New York, NY, USA, 12 pp

Micenková, B., Ng, R.T., Dang, X.H., Assent, I. (2013). Explaining outliers by subspace separability. *In Proceedings of the 13th IEEE International Conference on Data Mining* (ICDM), Dallas, TX, pages 518–527, 2013.

Oliver, M. A. (2010). *Geostatistical Applications for Precision Agriculture*, Springer, London, UK, 295 pp.

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

Santesteban, L. G., Guillaume, S., Royo, J. B., & Tisseyre, B. (2013). Are precision agriculture tools and methods relevant at the whole-vineyard scale? *Precision Agriculture*,14(1), 2-17.

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 Precision agriculture'13. Wageningen Academic Publishers*, 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.

Vinh, N.X., Chan, J., Romano, S., Bailey, J., Leckie, C., Ramamohanarao, K., Pei, J. (2016). Discovering outlying aspects in large datasets. *Data Mining and Knowledge Discovery*, 1–36.

Zhao, J., Lu, C., Kou. Y. (2003). Detecting Region Outliers in Meteorological Data. *In Proceedings of the 11th ACM International Symposium on Advances in Geographic Information Systems*, 49–55, New Orleans, Louisiana, USA.