

A Novel Cleaning Method for Yield Data Collected by Sensors: A Case Study on Winter Cereals

Antonio Natale[1], Sara Antognelli[1(✉)], Emanuele Ranieri[1], Andrea Cruciani[1], and Antonio Boggia[2]

[1] Agricolus SRL, via Settevalli, 320, 06132 Perugia, Italy
discover@agricolus.com
[2] Department of Agricultural, Food and Environmental Sciences, University of Perugia, Borgo XX Giugno, 74, 06123 Perugia, Italy

Abstract. Winter cereals yield tracking is a common practice since decision support systems can greatly benefit from the integration of these data. However, scientific literature highlights that many systematic errors occur during yield data collection. An efficient and easy to automatize protocol to clean collected field data is still missing despite its development is essential to integrate this useful tool in a smart-farming platform.

This paper focuses on the development of a yield data cleaning procedure, easy to industrialize and performant in different contexts. This method is based on both empirical cleaning steps and statistical analysis on the "moving windows". The developed cleaning procedure enabled the mixing of data coming from different combine harvesters and considered yield data measurements from the farmers to strengthen the results. In order to create readable and complete maps, an interpolation method concludes the procedure. The developed method is applied on a case study on real farm data.

Keywords: Yield maps · Winter cereals · Smart farming · Harvest · Outlier

1 Introduction

Winter cereals are widespread cultivated and farmers often adopt smart farming solutions mainly to i) define field management strategies, ii) optimize the nitrogen fertilizations. In this context, yield maps of winter cereals are very powerful tools to reach both targets. Based on these maps, in fact, it is possible to diagnose field problems and plan new strategies. Most important, these maps could be an input of a machine-learning algorithm for the early-prediction of winter cereals yield [1].

The smart farming platform Agricolus® is integrating yield maps in its Decision Support System (DSS). Agricolus DSS collects and analyses data from different sources (field sensors, remote sensing, crop scouting) to allow farmers to take the best data-driven decision [2]. The addition of yield maps to the other information provided by the

© Springer Nature Switzerland AG 2020
O. Gervasi et al. (Eds.): ICCSA 2020, LNCS 12253, pp. 684–691, 2020.
https://doi.org/10.1007/978-3-030-58814-4_55

system will not just complement the data but it will be a key object in the field-specific activity planning.

However, it is known that different problems can occur in the creation of the yield map. The main one concerns the data-collection in field with yield tracking systems, since bias are very common, and they can easily lead to a dangerous misinterpretation of results. Scientific literature, in the last decades, was focused on analyzing the main causes of errors on the yield data collection [3–5]: i) *header cut-width*: errors affect the sensed area that may be generated from a) errors of the harvester in identifying "active" or "non active" cutting bar sections; b) too wide cutting bar sections for the definition of the local harvested area; ii) *header position*: errors due to the combine harvester sensor or by the driver can cause a misregistration of the header position. This determines recorded points with yield = 0; iii) *lag time (or flow delay)*: this parameter is fixed and it represent the time from the moment the cereal is cut to the moment the yield flow is measured. It is needed to associate the yield flow to the exact location. However, even if it is set accurately, it may vary depending on the field conditions, and this may cause errors in data georeferentiation. iv) *travel distance measurements*: when the combine speed is locally very irregular (stop and start) it may generate area with speed = 0 and high yield, followed by areas with high speed and low yield. This affects both the localization of yield and the sensed area; v) *GNSS Positioning errors*: yield data can have wrong coordinates due to technical errors.

Given that, an efficient protocol to clean collected field data is still missing, despite being recognized very important [4, 6–8]. The aim of this paper is to define a robust and easy to industrialize cleaning methodology for yield data collected by tracking systems. To do that, a literature review on the existing data cleaning methodologies was carried on. Then SWOT analysis was used to compare 5 papers. Then, different steps were extracted from the reviewed methods following a SWOT criterion (strength, weaknesses, opportunities and threats), and merged in a chain process. Finally, a case study was tested and evaluated.

2 Methodology

SWOT analysis was performed on 5 papers - [5, 9–12] - published from 2002 to 2019 on scientific journals (Table 1). The methodologies were analysed with the aim of understanding the robustness and the suitability of the suggested method to be automatized. In fact, the protocol should be feasible in different contexts, without recurring to local empirical knowledge. Plus, the process shouldn't require ancillary data (such as the machinery speed at harvest or the grain humidity), as they might not be always available. As the output would be provided to the final users, it should be also clearly explicable. The SWOT analysis for yield data cleaning methodologies is presented in Table 1. The definition of the cleaning protocol was based on the results of the SWOT analysis and it includes several steps.

The missing data resulted from the cleaning procedure were filled by interpolation through Inverse Distance Weighting method (IDW) [13]. The IDW parameters were: 1) Power = 2, as suggested by [13]; 2) research radius: 12 m to consider at least 3 parallel swaths, and a minimum number of points of 12, to clean properly the field borders. The

spatial resolution of the final raster was set to 10 m. The resolution of the original data was reduced to create usable results, easy to manage and comparable with commonly used satellite data, like Sentinel 2, that would be used in further studies.

Table 1. SWOT analysis of selected methodologies for yield data cleaning

Reference	Strengths (S) and weaknesses (W)	Opportunities (O) and threats (T)
[5]	• Use of empirical criteria, easy to understand (S) • Calculations are easy (S) • Needs a lot of input data (W) • The user has to define many threshold values (W) • Not considering spatial relation between elements (W)	• It can be adapted to different contexts, as many criteria are user-defined (O) • Errors are easy to identify (O) • Need of local knowledge: difficult to industrialize (T) • Some criteria can be correlated, increasing the complexity of the model but not the efficacy (T)
[9]	• Use of empirical criteria, easy to understand (S) • Calculations are easy (S) • Needs a lot of input data (W) • The user has to define many threshold values (W) • Not considering spatial relation between elements (W).	• It can be adapted to different contexts, as many criteria are user-defined (O) • Errors are easy to identify (O) • Need of local knowledge: difficult to industrialize (T) • Some criteria can be correlated, increasing the complexity of the model but not the efficacy (T)
[10]	• Calculations are relatively simple, and can be made in a single step (S) • Considers spatial relations between elements (S) • Input-wise: the method only uses yield data as input (S) • Use of statistical criteria, not immediate to understand (W) • Need the definition of two thresholds by the data analyst (W)	• Easy to industrialize: no need of local knowledge (O) • The model is relatively efficient to calculate, as it only needs a single variable (O) • Statistical based: the process should be tested as the errors are not easily identifiable during the process (T)
[11]	• Considers spatial relations between elements (S) • Input-wise: the method only uses yield data as input (S) • Use of statistical criteria, not immediate to understand (W) • Calculations are relatively complex and made in many steps (W) • Need the definition of many thresholds by the data analyst (W)	• Easy to industrialize: no need of local knowledge (O) • The model is relatively efficient to calculate, as it only needs a single variable (O) • Statistical based: the process should be tested as the errors are not easily identifiable during the process (T)

(*continued*)

Table 1. (*continued*)

Reference	Strengths (S) and weaknesses (W)	Opportunities (O) and threats (T)
[12]	• Considers spatial relations between elements (S) • Input-wise: the method only uses yield data as input (S) • Use of statistical criteria, not immediate to understand (W) • Calculations are relatively complex and made in many steps (W) • Need the definition of many thresholds by the data analyst (W)	• Easy to industrialize: no need of local knowledge (O) • The model is relatively efficient to calculate, as it only needs a single variable (O) • Statistical based: the process should be tested as the errors are not easily identifiable during the process (T)

2.1 Case Study

The methodology was then tested on a case study. The experiment was conducted during 2017-2018 season, using 4 selected fields: B33 (3,45 ha), B34 (8,68 ha), B35 (6,58 ha), and B37 (7,73 ha). The fields are located nearby the Badiola farm centre, on a hilly area nearby Perugia, in the central Italy (geographic coordinates: 12,3034; 42,9986). They were all cultivated with durum wheat (Variety: Odisseo). The fields belong to the Fondazione per l'Istruzione Agraria (Perugia, Italy). They were harvested using two different models of combine harvester: CLAAS Lexion 750, and CLAAS Lexion 630 (Fig. 1).

Preliminary field exploration showed that in the north-western areas of fields B33 and B35 cereals were partially lodged. This usually represents a limit for the harvesters, as they usually encounter problems in localizing the yield correctly.

Yield data were collected through a Volume-flow sensor mounted on the harvest machines. Yield tracking was performed by YieldTrakk (YM-1) system. The resulting dataset had an irregular spatial resolution, depending on the harvester working width (varying from 1,5 and 7,7 m) and its local working speed, as the point were recorded at regular time intervals. Centimetre-level positional accuracy was ensured by the Real Time Kinematic (RTK) positioning system. Data cleaning procedure was implemented in ESRI ArcGIS Pro software. Finally yield maps were produced.

Assuming that a good cleaning methodology should increase the normality of the distribution of the yield data [10], skewness and kurtosis indices were calculated before and after the data cleaning procedure as a performance test of the procedure itself. The skewness and kurtosis indexes assume respectively value of 0 and 3 in case of normal distribution.



Fig. 1. Map of the case study fields. The field areas harvested with different machines are highlighted *(Blue: CLAAS Lexion 630; Green: CLAAS Lexion 750).* (Color figure online)

3 Results and Discussion

Based on the results of SWOT analysis, a data cleaning method based on 6 steps has been defined:

1. Removal of null points: points where yield = 0 were eliminated from the dataset [5, 9, 12];
2. Removal of overlapping points, with the same yield value
3. Application of the cleaning data method suggested by [10]. This method is based on the concept of "moving window": each point is classified as "outlier" or "non outlier" based on values related to the points included in a circular neighborhood, defined by radius = R. The method includes the following steps:

    a. Definition of coefficient of maximum variation (CVmax - threshold of acceptability of the coefficient of variation): 20% was used [10]
    b. Definition of R of neighborhood: [10] suggested to use 1,5 times the working width as a radius, to ensure that at least 3 working width are used. A 12 m radius have been used in the experiment.
    c. Calculation of the number of points included in the neighborhood (N)
    d. Calculation of coefficient of variation (CV) of the points included in the neighborhood.

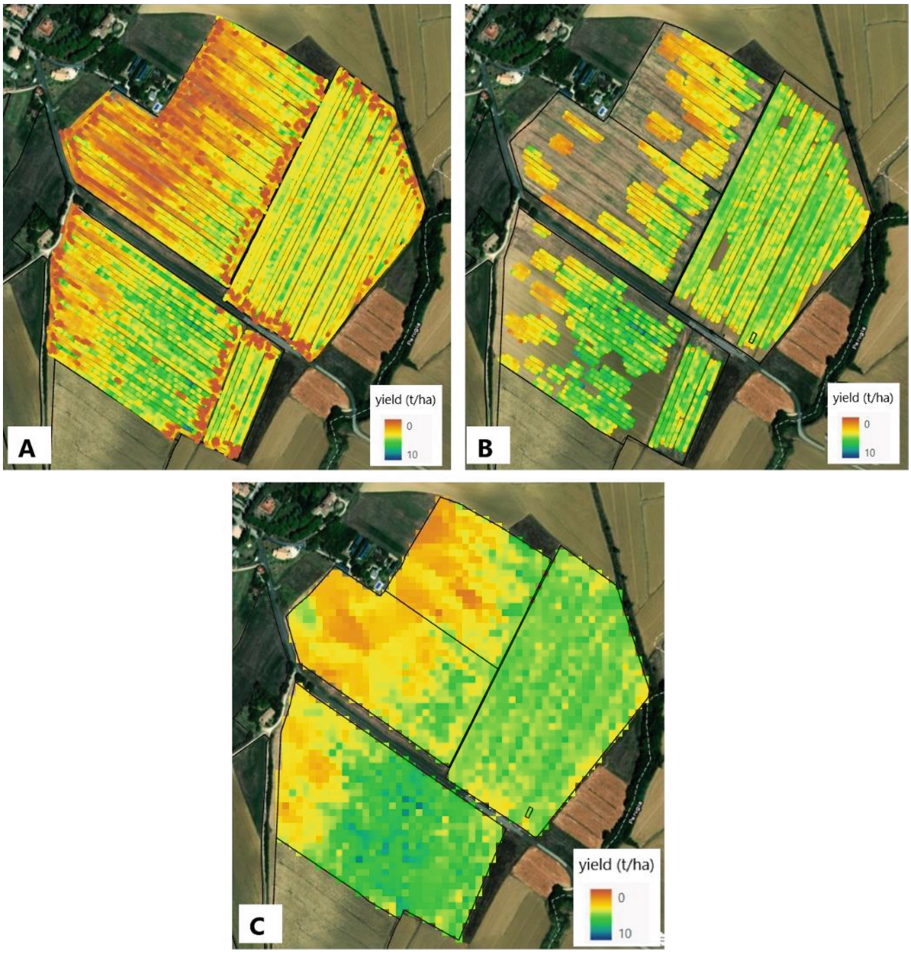

Fig. 2. Yield data (t/ha) of the fields included in the case study. A: raw yield; B: clean yield data; C: interpolated yield.

e. Calculation of the number of points with CV > CVmax within the neighborhood (NCVmax).
f. Definition of the outlier: points where N = NCVmax are identified as outlier.

4. Removal of points not included in the numeric interval between $(\mu - 3\sigma)$ and $(\mu + 3\sigma)$ [12] where μ is the population mean and σ the standard deviation;
5. Co-registration between different combine harvesters. Yield values of one combine harvester were varied proportionally to the difference between the mean of the yield recorded by the two different combines working on the same field.
6. Post-calibration of the co-registered yield with the measured yield (at the mill), where values were varied proportionally to the difference between the two means

Cleaned yield maps show a clear visual improvement of the yield data collected (Fig. 2). The method eliminated most of the values located in the areas of the field where lodging occurred, where the data recorded by the harvester combine resulted very locally variable. However, the interpolation filled the missing data coherently with the spatial trends observed in the original data. The post-calibration step avoided the risk of overestimating or underestimating the average yield per field.

Data distribution started with a skewness index ranging from 0,79 and 20,61 and a kurtosis ranging from 1,60 and 860,86 in the different fields. After the data cleaning, skewness ranged between 0,20 and 0,53, while kurtosis ranged between 2,81 and 3,88 after the implementation of the 6 steps (Table 2).

Table 2. Skewness and kurtosis indices of the yield data distribution in each test field before and after the application of the data cleaning procedure (harvester combine model is reported in parentheses).

Field name	Skewness before data cleaning (L630)	Kurtosis before data cleaning (L630)	Skewness before data cleaning (L750)	Kurtosis before data cleaning (L750)	Skewness after data cleaning	Kurtosis after data cleaning
B33	9,03	142,85	1,16	3,06	0,53	3,88
B34	20,61	860,86	5,74	112,71	0,23	3,77
B35	5,94	52,69	0,79	1,67	0,24	2,81
B37	2,57	20,42	2,06	19,46	0,20	3,12

4 Conclusion

The critical review of literature allowed the definition of a complete and flexible novel methodology for yield data cleaning. The methodology proved to be effective in eliminating most of the yield data errors. Plus, it is easily automatable as it is mainly based on statistical calculation rather than local knowledge. The points identified as outliers were located nearby the field margin, in the manoeuvring area of the tractor, as well as in the centre of the field, especially in the north-western area, where the yield appeared lower and irregular. In this area, in fact, the crop was damaged during the season by wind and hail. This specific problem clearly influenced the results obtained in fields B33, B35 and B37. Field B34 was less influenced by atmospheric events. Consequently, the number of outliers was lower.

The methodology is being tested on a wider dataset, including yield data of around 400 ha of winter cereals, harvested in 2018, 2019 and 2020, giving promising results. Then, further studies will be needed to test and improve it. After this stage, the procedure would be implemented in the Agricolus platform, to give to the farmer a powerful tool to produce correct and easy to read yield maps from the different combine harvester they may use in their fields.

Acknowledgements. We acknowledge Mr. Mauro Brunetti of the Foundation for Agricultural Education of Perugia for his very helpful and valuable support and collaboration in the data collection activities, and the University of Perugia to have supported the research activities as part of the PhD programme.

Funding. This research was developed within the framework of the project "RTK 2.0—Prototipizzazione di una rete RTK e di applicazioni tecnologiche innovative per l'automazione dei processi colturali e la gestione delle informazioni per l'agricoltura di precisione"—RDP 2014–2020 of Umbria—Meas. 16.1—App. 84250020256.

References

1. Luque Reyes, J.R., et al.: Early prediction of winter cereals yield. In: Global IoT Summit (GIoTS). IEEE (2020)
2. Guidotti, D., Marchi, S., Antognelli, S., Cruciani, A.: Water management: agricolus tools integration. In: Global IoT Summit (GIoTS). IEEE (2019)
3. Luck, J.D., Fulton, J.D.: Best management practices for collecting accurate yield data and avoing errors during harvest. University of Nebraska Cooperative extension, Institute of Agriculture and Natural Resources (2004)
4. Adamchuck, V.I., Dobermann, A., Ping, J.: Listening to the story told by yield maps. University of Nebraska Cooperative extension, EC 04-704 (2004)
5. Sudduth, K.A., Drummond, S.T.: Yield editor: software for removing errors from crop yield maps. Agron. J. **99**, 1471–1482 (2007)
6. Arslan, S., Colvin, T.S.: Grain yield mapping: yield sensing, yield reconstruction, and errors. Precis. Agric. **3**, 135–154 (2002)
7. Kharel, T., et al.: Yield monitor data cleaning is essential for accurate corn grain/silage yield determination. Agron. J. **111**, 509–516 (2018)
8. Blackmore, S.: The interpretation of trends from multiple yield maps. Comput. Electron. Agric. **26**(1), 37–51 (2000)
9. Kleinjan, J., Clay, D.E., Carlson, C.G., Clay, S.A.: Developing productivity zones from multiple years of yield monitor data. Site-Specific Management Guidelines (SSMG) Series, International Plant Nutrition Institute (IPNI) (2002)
10. Spekken, M., Anselmi, A.A., Molin, J.P.: A simple method for filtering spatial data. In: 9th European Conference on Precision Agriculture, ECPA (2013)
11. Maldaner, L., Corrêdo, L., Rodrigues Tavares, T., Mendez, L.G., Duarte, C., Molin, J.: Identifying and filtering out outliers in spatial datasets. In: 14th International Conference on Precision Agriculture, Montreal, Quebec, Canada (2018)
12. Vega, A., Córdoba, M., Castro-Franco, M., Balzarini, M.: Protocol for automating error removal from yield maps. Precis. Agric. **20**(5), 1030–1044 (2019). https://doi.org/10.1007/s11119-018-09632-8
13. Souza, E., Bazzi, C., Khosla, R., Uribe, O.M., Reich, R.M.: Interpolation type and data computation of crop yield maps is important for precision crop production. J. Plant Nutr. **39**, 531–538 (2016)