The importance of data quality assessment for machinery data in the field of agriculture

Chapter · January 2022

DOI: 10.51202/9783181023952-495

2 authors:

Morteza Abdipourchenarestansofla
John Deere Kaiserslautern
5 PUBLICATIONS **0** CITATIONS
SEE PROFILE

Christof Schroth
Fraunhofer Institute for Experimental Software Engineering IESE
3 PUBLICATIONS **0** CITATIONS
SEE PROFILE

The importance of data quality assessment for machinery data in the field of agriculture

M.Eng **Morteza Abdipourchenarestansofla**,
John Deere, Kaiserslautern;
M.Sc. **Christof Schroth**,
Fraunhofer IESE, Kaiserslautern

Abstract

In the field of agriculture, particularly in the horizon of farming 4.0, more and more smart sensors are being used and telematics brought lots of machinery data production in agricultural. The data obtained can help the farmer with optimization or decision support by means of Artificial Intelligence. In the horizon 2020 Demeter project, we develop a job cost system which aims to enable calculating site-specific costs for fertilization and plant protection applications. Telematics machinery data are leveraged for developing the system. As such field operation data are complex and contaminated by various sources of measurement and documentation errors, it is crucial to have an appropriate Data Quality approach which can reveal issues in such data. In this work we present concrete quality challenges in geospatial machinery sensor data and their metadata documentation, and how we check them based on ISO 25012 and ISO25024 standards via an automated data quality assessment service.

Introduction

Field operations data acquired by agricultural machineries in precision agriculture (PA) provides ever-increasing opportunities for sustainable crop production system, pushing frontier in agricultural economics, and support in designing better agricultural policy. Such data can also be referred to telematics data. Machine telematics data in agriculture includes both fleet data and field operations data [1]. The application for which agricultural machinery is used in precision crop farming is ranging from fertilizer and chemical applications, seeding/planting which are pulled by tractor, and harvest operations. As argued by Carletto et al. [2], despite tremendous progress in agricultural data production, large gaps remain in terms of availability and quality of agricultural data. Although PA supports sustainable crop production system, its usability and feasibility demand high-quality data that ensure the traceability of performance and support in making informative decisions [3]. By studying relevant literature, we found only one recent review paper in the domain of PA that mentions data quality explicitly as a problem in arable farming [4]. There is also a study by Zhang et al. [5] on improving the accuracy of the quality of fertilizer applicators in terms of location accuracy and the time lag switching between different rates for nitrogen to be applied. Generally, the topic of data quality in agriculture is mostly addressed with an economic and geospatial point of view, more particularly focusing on GIS and Remote Sensing data quality [2, 6, 7]. Such studies emphasize on positional and the thematic accuracy of geospatial observation. Corrupted data, missing data, or data points outside a valid range or exceeding credible boundaries must be detected, and appropriate countermeasures like data imputation need to be utilized to avoid misleading results and wrong decisions. Data quality dimensions are including but not limited to accuracy, completeness, and consistency [3, 8]. In this study, we emphasize the need to assess data quality for field operation data and briefly describe our approach to address some of the challenge.

Machinery PA data and quality challenges

Increasing data quality should be motivated based upon applications that properly quantify and communicate the value of data with simple and accessible metrics of impact. This section aims to investigate data quality issues for machinery data which are being used in H2020 Demeter project[1] in a use case (pilot 2.2). This use case develops a job cost calculation system which aims to support the farmer by automating cost calculation associated to a given field operation. This technology, its reliability and accuracy, requires high-quality data that avoids misleading results. The system leverages telematics machinery data with the scope of fertilizer and chemical applicators in small grain. The chemical and fertilizer applications take place several times during the season. In such applications, the sensors measure the applied rate (e.g., kg/ha) which is applied by spreader or sprayer pressure. The observation is associated with a geometric representation of the applied product and machine related performance such as “swath width”, “section id” etc. To assess the quality of telematics machine data, we define two categories:

1) Global positioning and sensor readings quality

Machinery data are contaminated with various sources of errors, such as measurement location [9], operation dynamic, field geometry, speed changes, and sensor deficiencies [10]. These influencing factors can potentially lead to measurement errors in the observation. The positioning accuracy can be disturbed by multiple factors such as atmospheric effects,

[1] https://h2020-demeter.eu/

multipath effects, etc. [11]. Such errors result into unrealistic applied rate and observation that are appeared to be outside the field boundary.

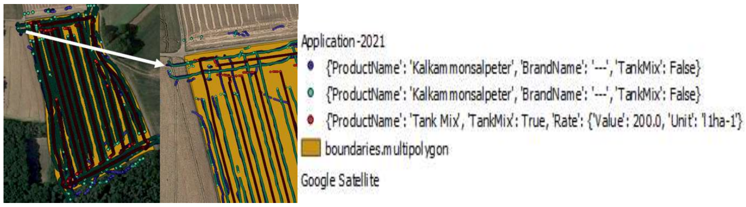
Application-2021
● {'ProductName': 'Kalkammonsalpeter', 'BrandName': '---', 'TankMix': False}
● {'ProductName': 'Kalkammonsalpeter', 'BrandName': '---', 'TankMix': False}
● {'ProductName': 'Tank Mix', 'TankMix': True, 'Rate': {'Value': 200.0, 'Unit': 'l1ha-1'}
boundaries.multipolygon
Google Satellite

Fig. 1: Three different application operations applied in three different dates on a single field with Winter Barley Crop, season 2021.

In Fig. 1 the measurement location of some of the observations are outside the field boundary which is not desirable because this means that for example Nitrogen is applied outside the field boundaries. This might be due to the operator error or unstable GPS connectivity. To demonstrate measurement quality in the applied product rate for the three field operations (see Fig. 1), we provided an overview in Table 1. Sometimes the sensor readings for “applied rate” for each location might not match the desire/target rate which was defined by the farmer. In addition to that, there are observations with zero “applied rate” and in case this differs from the “target rate” it might be due to sensor deficiency. Due to the dynamic nature of the operation and the environmental factors mentioned above, the sensor is sometimes not able to read the exact applied rate for each location.

Table 1: Statistical summary of three-applications in one season for one field,

Product-name	min-max Applied rate	min-max Target-rate	Number of data points measured by sensor	Zeros- applied rate
Kalkammonsalpeter	0-660 kg/he	220-220 kg/he	8935	2%
Kalkammonsalpeter	0-660 kg/he	220-220 kg/he	9301	2.20%
Tank Mix	166-364 l/he	168-237 l/he	8205	0%

2) Metadata documentation and quality related to operator input and adoption to the technology.

The documentation/metadata of the sensor readings to each field, each crop season, and each field operation requires a consistent logic which can also ensure validity, accuracy, completeness, and uniqueness of the data. The farmer skills have a large impact on data quality. He/she needs to know how to operate with advanced PA equipment, so the mistakes won't affect the data quality. For instance, in planning a spraying application an operator can make mistake in inserting a wrong name for a chemical product and may forget to provide information for the product brand. In this research with our corporate farms, we observed several field operations containing such typical mistakes which result into bad data quality in terms of completeness. During metadata quality profiling we found that "crop-season" is not consistent in metadata in almost 60% of the data. For instance, there were two different machines operated on a field with a span of few days for a fertilization application, where the first machine documented "crop-season" as e.g., 2020 but the second machine as 2021.

Approach for data quality assessment of structured data in PA

To address the challenges described above, we studied the ISO 25012 and 25024 standards. Three interviews with subject experts are conducted. In the first round, we asked which dimensions are relevant to their applications and the interviewees explicitly needed to rank their most important dimensions. After the first round of interviews, the most important dimensions turned out to be consistency, accuracy, and completeness. Taking a closer look at the ISO 25012 standard the dimensions can be split in two areas, inherent and system dependent point of view. Our three most important dimensions are all from the inherent point of view. Other important results of the first round are:

- Each measurement result should be scored, i.e., a number between 0% (bad data quality) and 100% (appropriate data quality),
- The results must be provided in a machine-readable format (e.g., JSON) to process the data in an automated fashion,
- The data quality assessment should be held flexible to meet users' need,
- For numerical values, statistical values (mean, median, quantiles and so on) must be delivered right at hand,
- Visualisations (e.g., histograms) are obligatory.

There is a need to decide automatically whether the data should be taken for further analysis or not. For example, corrupted data or measurements that contain outliers should be removed before the job cost calculation starts. The data quality service should have some flexibility to meet specific user needs, e.g., the completeness of the data, the user should be able to specify parts that are relevant for further analysis and can neglect unnecessary ones. For numerical values, it is often necessary to check if they are accurate, e.g., they need to be in a pre-defined interval. For the job cost calculation, this interval can be provided by the operator through the “control rate” plus an offset based on expert opinion. Visualizing the results of the data quality assessment helps to understand the cause of poor data quality. Based on the interview results we implemented a data quality assessment service to identify the issues described in the previous sections, focussing on the inherent point of view of the data. In the follow-up interviews, we asked which metrics are the most relevant ones for their work and to which dimension a metric belongs. According to the results of these interviews, we implemented 21 metrics to measure the different data quality dimensions (4 related to accuracy, 8 completeness, 2 consistency, 5 credibility, 1 precision, and one addressing understandability). An illustration of the process can be found in Fig. 2.

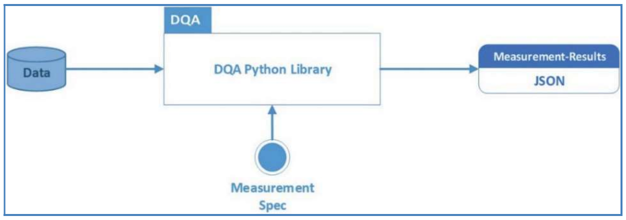

Fig. 2: Procedure of the data quality analysis. The measurement specification is provided as a JSON file.

This service can be used to assess the quality of machinery data including structured data such as JSON files and ESRI shapefile, The test phases are started, and improvements of the service are expected until the end of the Demeter project.

Next steps

By the end of the Demeter project 2023, we will have one solution for the job cost calculation considering the results of the data quality assessment service. The integration of it in the ecosystem of the Demeter. Also, the data quality assessment service will be applied and tested in two other use cases. For example, another use case focusses on engine monitoring data for which it is important that the records do not have any gaps in time. We expect that based on the feedback, adjustments of metrics or implementations of new ones of the data quality assessment service are needed. Furthermore, by considering new environmental settings and systems, other data quality dimensions might be more important, e.g., compliance, confidentiality, and availability. In a long-term perspective, there is a need for appropriate countermeasures in case of missing data. For example, if values in the “target rate” column contains only zero values due to an error, the values must be replaced by average values.

Acknowledgement

Parts of this work was funded through EU H2020 under Grant number XYZ. We would like to thank Anna Maria Vollmer for her valuable comments.

Reference

[1] *Tyler Mark* and *Terry Griffin*, (2016), Defining the Barriers to Telematics for Precision Agriculture: Connectivity Supply and Demand, No 230090, 2016 Annual Meeting, February 6-9, 2016, San Antonio, Texas, Southern Agricultural Economics Association

[2] Carletto, C., Dillon, A., & Zezza, A. (2021). Agricultural Data Collection to Minimize Measurement Error and Maximize Coverage. Global Poverty Research Lab Working Paper, (21-108).

[3] Malaverri, J. E. G., & Medeiros, C. B. (2012). Data Quality in Agriculture Applications. In *GeoInfo* (pp. 128-139).

[4] Villa-Henriksen, A., Edwards, G. T., Pesonen, L. A., Green, O., & Sørensen, C. A. G. (2020). Internet of Things in arable farming: Implementation, applications, challenges and potential. *Biosystems Engineering, 191*, 60-84.

[5] Zhang, J., Liu, G., Huang, J., & Zhang, Y. (2021). A Study on the Time Lag and Compensation of a Variable-Rate Fertilizer Applicator. *Applied Engineering in Agriculture*, *37*(1), 43-52.

[6] Congalton, R. G., & Green, K. (2019). Assessing the accuracy of remotely sensed data: principles and practices. CRC press.

[7] Goodchild, M. F., & Li, L. (2012). Assuring the quality of volunteered geographic information. *Spatial statistics*, *1*, 110-120.

[8] Sidi, F., Panahy, P. H. S., Affendey, L. S., Jabar, M. A., Ibrahim, H., & Mustapha, A. (2012, March). Data quality: A survey of data quality dimensions. In *2012 International Conference on Information Retrieval & Knowledge Management* (pp. 300-304). IEEE.

[9] Borgelt, S. C., Harrison, J. D., Sudduth, K. A., & Birrell, S. J. (1996). Evaluation of GPS for applications in precision agriculture. *Applied engineering in agriculture*, *12*(6), 633-638.

[10] Abay, K. A. (2020). Measurement errors in agricultural data and their implications on marginal returns to modern agricultural inputs. *Agricultural Economics*, *51*(3), 323-341.

[11] Grisso, R. D., Alley, M. M., & Heatwole, C. D. (2005). Precision Farming Tools. Global Positioning System (GPS).