https://doi.org/10.17221/361/2024-PSE

Information sources in agriculture

Jan Jarolímek, Jakub Samek, Pavel Šimek, Michal Stočes, Jiří Vaněk, Jan Pavlík*

Department of Information Technologies, Faculty of Economics and Management, Czech University of Life Sciences Prague, Prague, Czech Republic

**Corresponding author: jarolimek@pef.czu.cz*

Citation: Jarolímek J., Samek J., Šimek P., Stočes M., Vaněk J., Pavlík J. (2024): Information sources in agriculture. Plant Soil Environ., 70: 712–718.

Abstract: The aim of this study is to define data sources and propose methods for effective and secure data management in an agricultural enterprise in the context of using data for decision support. Current developments in information and communication technology (ICT) have contributed towards the increase in the amount of generated data in various fields. The main data sources for agricultural enterprises are the farm itself, suppliers, government, market, and research. The use of smart solutions, artificial intelligence, and other innovative practices in agriculture is discussed at many conferences, in various journals, strategies and project plans. Data is the essential raw material for all these solutions. Large amounts of data cannot be analysed efficiently with spreadsheet programs. Currently, there are trends in the use of data, for example, in business intelligence (decision-making systems), e.g. tools using online transaction processing (OLAP) or process automation or the possibility of e.g. tracing the origin of food. The availability and possibility of creating large data sets bring many challenges related to managing that data. To effectively manage farm data, it is essential to have a well-developed data management plan (DMP) used to formalise the processes related to handling. A DMP mainly addresses archiving, backup, licensing and other important aspects of data management. The challenges and developments in farm data management include incorporating artificial intelligence into data analysis and security. Food is classified as an "Entity of Critical Importance" in the NIS2 EU Directive, which also deals with cybersecurity issues.

Keywords: data map; data quality; usability of data; data analytics; farmers; data management plan

The era of big data in the 21st century has brought unprecedented opportunities for data collection across various industries. Nowhere is this more evident than in agriculture, where data acquisition has become a routine aspect of modern farming practices, often due to legislation. Food safety, health protection, and "from farm to fork" approaches are the objectives of all EU laws and standards in agriculture, as published in Food Safety – EU Action, European Union (2018). However, despite the abundance of data, a critical problem has emerged – much of the data collected over the past two decades has proven less than optimal for meaningful analytical processing, not allowing further rapid development in digitisation and automation.

The agricultural sector has been diligent in gathering data, yet the quality of this data often falls short when it comes to consistency, coherence, and completeness. These limitations stem from a range of factors, including the absence of context, missing metadata, and the pervasive influence of human error. Unlike data collected in more controlled environments, operators frequently enter agricultural data manually rather than through automated processes. This introduces

The knowledge and data presented in the present paper were obtained as a result of the FEM CULS Prague Internal Grant Agency, Grant No. 2022B0006, and supported by the Ministry of Agriculture of the Czech Republic under the ZEME program, Project No. QK 23020058.

https://doi.org/10.17221/361/2024-PSE

a significant margin of error and further complicates the task of leveraging the data for sophisticated analytics. This all leads to misunderstandings among various interest groups, particularly IT analysts and economists, *versus* farmers and lawmakers. For example, farmers can believe they have been collecting data for over two decades; they have been working with them on a daily basis and using them for real-time evaluations. Such real-time input correction is typically based on current conditions and does not necessitate deeper analysis of older data. With data collections tailored for real-time usage, IT analysts are simply unable to perform reliable prediction, extrapolation, interpolation or correlation of the production and economic indicators. In theory, everything is according to the law, mathematical models are suitable, missing values can be imputed, outliers detected (Kar et al. 2020), and all efforts are adequately supported by the European Union. In reality, no adequate progress is being achieved.

Data collection and evaluation, therefore, have to be viewed as issues that combine technical aspects of used ICT hardware and software and social interactions between the stakeholders. Lawmakers, clerks, farmers, vendors, employees as well as consumers have their needs based on their limited perception of reality and knowledge of their immediate surroundings. One of the key factors to understanding the data quality is to enhance the strictly technical approach by understanding the interactions between people involved in the data collection process (Chen et al. 2017).

The sociological point of view plays the same important role as the IT definition itself. What is a farmer's perception of digitalisation and data gathering? It seems that sociological aspects are not geo-related. Research is focused on specific regions or countries, such as Nigeria, South America, or Asia. Despite the different climate in the mentioned regions, scientific papers are accenting or focusing on education, level of digitalisation and internet availability and demographic indicators as published by Oladele and Fawole (2007), Braun et al. (2018), Sylvere and Jean D'amour (2020), da Silveira et al. (2023), and emphasise it as a fundamental obstacle to further digitalisation which would be necessary for additional advanced data processing (Estes-Zumpf et al. 2022). Although the authors are writing about data quality and its evaluation, an adequate definition of quality is missing. This phenomenon of poorly established definitions and standards is not uncommon. Scientists and IT communities have learned over the years to treat the concept of data quality as a dogma, requiring no further or detailed explanation. A different group of scientists noticed that data for agricultural research is often collected by questioning respondents about past seasons and agricultural cycles, which, due to memory biases, leads to a decrease in the overall quality of the research (Beegle et al. 2012). This is, of course, unacceptable for reliable data evaluation and data-driven decision-making (Tantalaki et al. 2019).

Despite the previous paragraphs, which indicate that data quality and availability belong to soft system problematics, data storage and evaluation are still of a technical nature. Long-term agricultural projects and research, such as "Soil quality indicators as influenced by 5-year diversified and monoculture cropping systems", as published by Feng et al. (2020), are designed for data collection from the very beginning. Therefore, there is no need for an explicit definition of the quality or availability itself. Automation and AI seem to be promising elements for advancing data collection. Data quality is a core concern in terms of decision-making processes based on IDSS because farmers will only rely on tools that have ensured reliability based on data quality. And not only farmers or human-controlled processes. Data and their metadata are the key aspects of automated control and response tools, such as precise irrigation systems supported by using a computer-based algorithm of heuristics (Pálková et al. 2012). However, it is worth emphasising that data quality is a shared responsibility, and the farmer is an essential part of ensuring this. Improving and maintaining data quality is an integral part of the data collection process for the IDSS to work properly (Baldin et al. 2021).

Evaluation by established deterministic algorithms using mathematical equations defined specifically for agricultural purposes is not the only approach. As machine learning (ML) is becoming more and more popular, its use in the field of sustainable agriculture is a highly debated topic. Despite the promising results and advantages, there are some limitations in applying ML algorithms. One of the most critical issues is the poor quality and consistency. Indeed, data quality affects ML analysis significantly. Data must have these characteristics: validity, consistency, uniformity, accuracy, and completeness. Operations associated with farming itself, and all the activity carried out in the farms are not fully automated, and hence data may not be properly annotated and may be lost and/or inaccurate (Trapanese et al. 2024).

https://doi.org/10.17221/361/2024-PSE

If improved in the future, ML and AI can appear to be powerful tools for processing complete yet non-structured data, usually stored in a human-readable form such as PDF or TXT files. According to research focused on the possibilities of reusing open data in the sugar beet sector, most of the data are stored in formats unsuitable for subsequent machine processing (Stočes et al. 2018).

The sequences of collected data may not be continuous for many reasons, whether technical or semantic. It is necessary to interpolate the data and remove deviations that do not statistically fit into the dataset. This can be addressed through ex-post processing using various mathematical models and procedures, as suggested by research "A data analysis pipeline for efficient processing and utilization of temporal high-throughput phenotyping data" as published by Kar et al. (2020) and "Detection and correction method of erroneous data using quantile pattern and LSTM" as published by Hwang et al. (2018). Automated collection of agricultural data is currently primarily carried out using inexpensive and accessible IoT devices, as noted by Omar et al. (2020). These devices inherently lack mechanisms to verify their own functionality; however, the expected accuracy surpasses the long-term accuracy of manual data collection.

Conducted research indicates a gap in the definition of data quality and availability. As we navigate the complexity of agricultural data, it is clear that simply accumulating information is not enough. Data must be reliable, standardised, and accompanied by contextual information to be able to use its full potential.

MATERIAL AND METHODS

The term data refers to data used to describe a certain phenomenon or property of an observed object. Data represents information suitably formalised for human and machine communication, interpretation and processing. A set of data relevant to a certain problem, i.e., data in a certain context that is usable and understandable and that can be further analysed in context and try to understand the implications, can be considered information. In the most general sense of the word, information is understood as information about the real environment, its state and the processes taking place in it. Understanding information is actually a piece of knowledge. It is the result of the process of knowing reality. Information becomes knowledge only in the hands of a person or system that knows how to use it and has experience with the decision-making process, as published by Jonák (2003a,b,c,d). Figure 1 shows the process from data to decision to the process that creates more data and information.

The identification of data sources in the agrarian sector is based on long-term cooperation with stakeholders from agricultural, forestry and food process enterprises, technology suppliers and government.

Data management requires creating, maintaining and updating data management plans (DMPs). This involves a systematic approach to ensuring that data is managed efficiently, securely, and in compliance with relevant regulations.

RESULTS AND DISCUSSION

Data map in agriculture

To understand the availability and form of data in agriculture, it is necessary to perceive agricultural data as a complex issue with all its factors, specifics and links. It is not possible to focus only on one particular area. For orientation in the sources and use of agricultural data, the authors constructed a data map in agriculture (Figure 2). It defines the main groups of agricultural data producers and users. The map shows the levels into which we can define the data source: areas of agricultural production, commodities, measured variables, technologies used for data collection up to a specific device/sensor. The colour scale presents the level of use of the given data source; dark corresponds to the highest level of use; orange is more of a goal or just a vision. The given scale is based on the authors' previous knowledge; the exact descriptions are the subject of further research.

Characteristics of the main sources and users of agricultural data

Farm data. The key producers but also the users of data are, or at least should be, farmers/agricultural enterprises. In essence, all technologies, processes and research should serve for efficient agricultural production.

Large amounts of data are generated in agricultural enterprises; specifications and a comprehensive list are a matter of a separate study; here, we will deal with the way they are used. Data acquired and used

https://doi.org/10.17221/361/2024-PSE

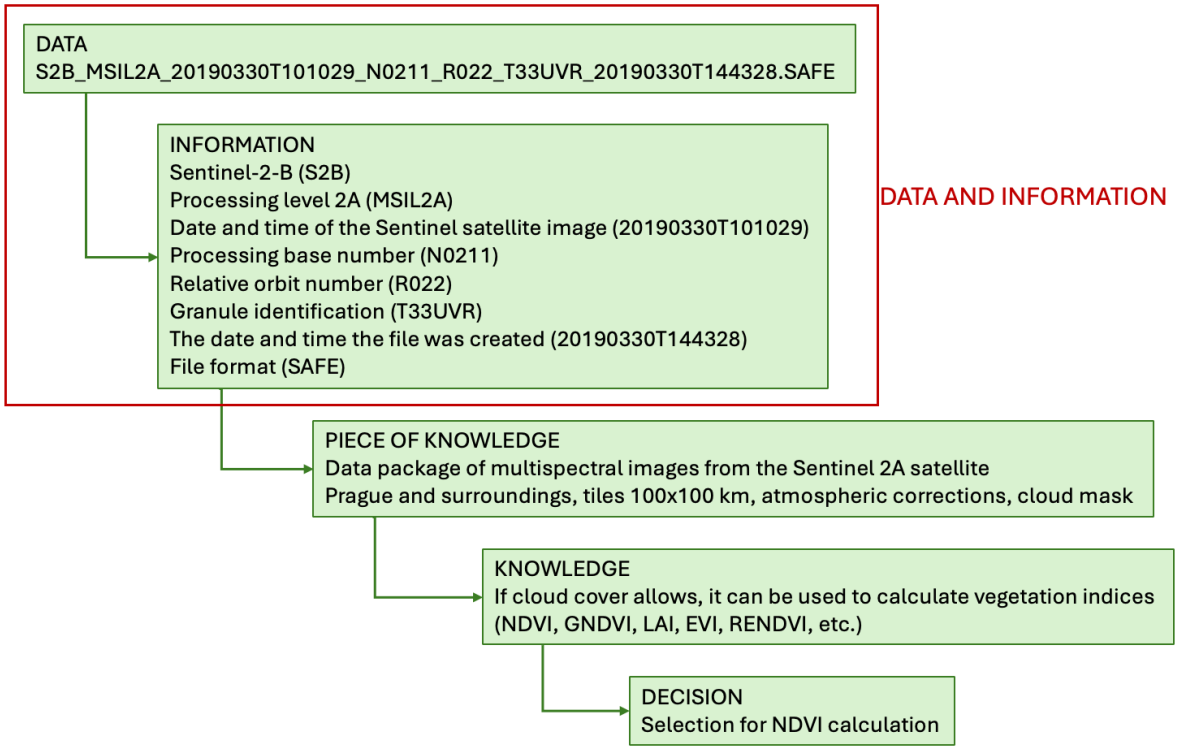

Figure 1. The process from data to decision

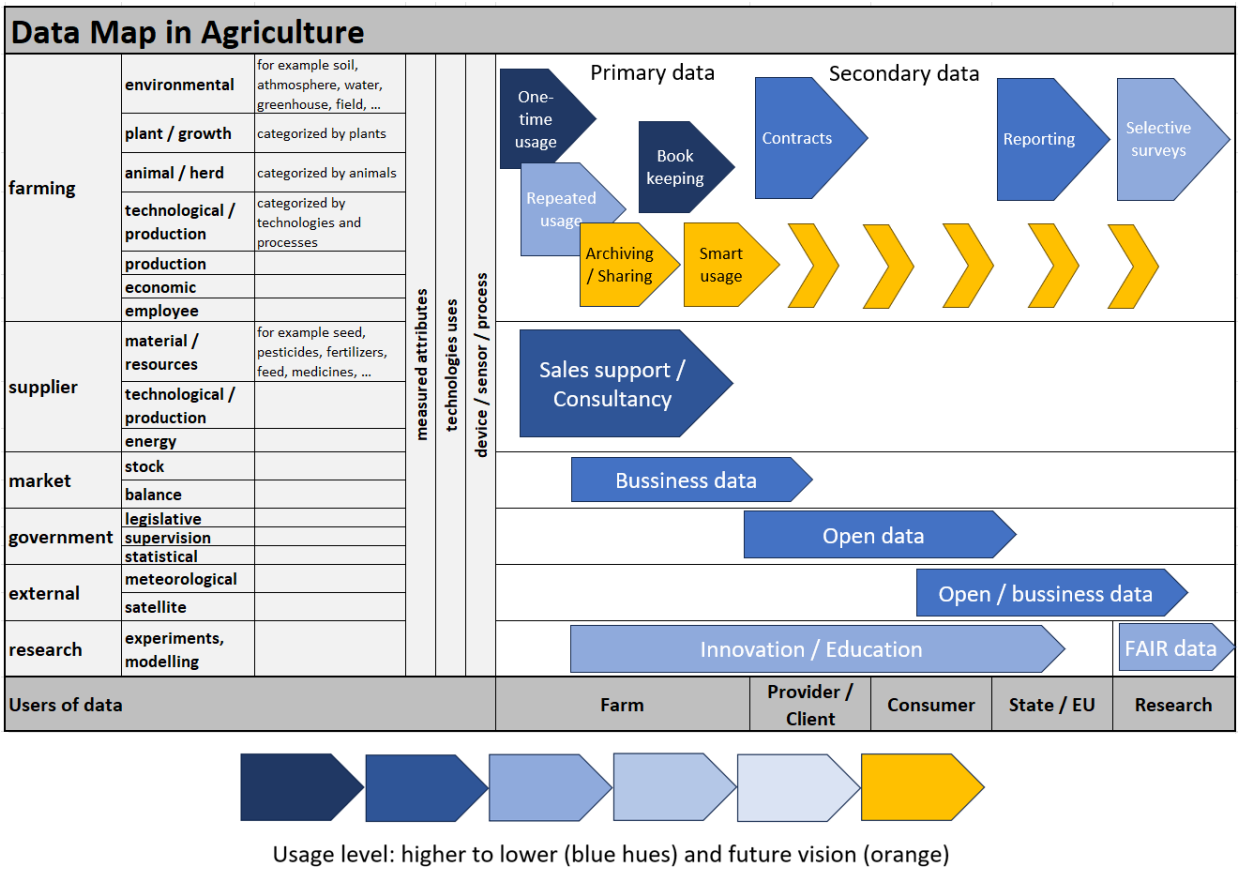

Figure 2. Data map in agriculture (source: author)

https://doi.org/10.17221/361/2024-PSE

in one company can be referred to as primary. The predominant method is one-time use at the point of origin when the data obtained is used to solve the current situation and is often not even stored; it can also be called "use and throw away". Part of this data is also used for various internal records. Only a small part of the obtained data is then used repeatedly. However, the goal should be complete archiving and the possibility of sharing – for further smart use of this data in your own company, as well as by other interested entities – most importantly in research, but also other by stakeholders, of course with the principles of personal data protection and anonymisation in mind.

Currently, archiving farm data in the necessary quality for further use is more of a vision. The data obtained from agricultural enterprises is already, in most cases, secondary data processed and modified for the purpose of use. The state obtains data from mandatory reporting for its needs. Research organisations then obtain the necessary data from state reporting, records, and various sample surveys when solving specific projects. It is often easier to obtain the necessary data for research experimentally than to utilise existing farm data. This may also be one of the causes of the problem of practical application of the results of science and research.

Supplier data. Part of the supply of technologies, pesticides, fertilisers, feed mixtures and all other raw materials and services is also the data and information defining the parameters and method of use of such materials. The provision of some information is dictated by legislation, e.g. pesticide labels. A large part of the information then serves to decide on the use of the given technology/raw material and the use itself. A large part of this information serves and is presented to suppliers for sales promotion and related advice for use in agricultural enterprises. Other entities use this data primarily from freely available sources or based on specific cooperation, e.g., in research.

Market data. Characterise the market for agricultural resources and commodities. Data from commodity exchanges, electronic markets, etc., can be classified as primary. A large part of the data comes from statistical surveys of states, the European Union, and various trade associations at the national and international levels. In most cases, data availability is equivalent for all necessary subjects.

State data. All states present a significant amount of data and information for managing the sector in the field of agriculture. This is legislation, various standards and rules. Other areas are the outputs of control mechanisms, state-organised statistical surveys, etc. For data presented by the state, we are talking about the principle of so-called OPEN data, i.e. freely distributable data with the possibility of machine readability.

Science and research data. These are outputs of research organisations, and availability for farmers and other subjects is mainly in the form of innovations; they are often part of universities' educational programs where research and education are connected. However, transferring research and development results into practice is a separate, not always very successful research chapter.

When sharing this data in the research sphere between different research workplaces, we talk about so-called FAIR data or FAIR principles, respectively.

External data. In the data map in agriculture, this part is used to include data that cannot be classified into the previous groups, even though their importance is considerable. These are, for example, data from such fields as meteorology, remote sensing of the earth, and others. Much of this data is available based on OPEN data principles, but sometimes it is also procured from commercial relationships. However, in most cases, data availability is equivalent for all necessary subjects.

Data management plan. The starting point for solving the availability of data from agricultural enterprises is an awareness of the current data use situation and the wasted potential caused by poor data preservation practices. At conferences and other scientific symposia, scientists often talk about using AI, autonomous machines, production models, etc.; all of this is impossible without enough data. The quality of this data is a separate issue.

After acknowledging the problematic situation, the solution is the deployment of data management plans across agricultural enterprises. The questions that help us realise that there is a problem are:

- Do we have devices that produce data?
- Do we save the acquired data?
- Are we able to reuse the stored data?
- Is it possible to share stored data between different company applications, technologies, and information systems?
- It is possible to share stored data with other entities (suppliers, consumers, the state, research teams, etc.).
- Do we have data covering all the necessary production areas?

https://doi.org/10.17221/361/2024-PSE

Created data management plans must also be maintained in a continuous process that aims to account for changes in technologies, legislation, and data usage. The creation of the data management plan itself needs to consider several key aspects:

- Basic information about data (project of field of data, acronym, keywords, short description, data steward, contact, etc.)
- Source and characteristics of data (data source, data type, data format, file naming, etc.)
- Metadata (basic metadata, special metadata format, standard, ontologies, FAIR principles, licences, etc.)
- Storage data (storage of data, security, backup, archiving, etc.)
- Ethics (processing of sensitive data, collection, protection, rights, etc.)
- Cost (cost of data acquisition, data management, storage, etc.)

The main parts are metadata and rights. Metadata must describe data perfectly for storage, searching, and processing, particularly by machines. Managing data access and sharing according to defined policies and depositing data in designated repositories ensures that data is accessible to authorised users. Monitoring data quality through validation and verification processes helps address any issues promptly.

Maintaining the data management plan requires regular reviews and updates to ensure it remains current. As the project progresses or data management practices evolve, the plan should be revised accordingly. Training team members on best practices in data management is crucial to ensure that everyone knows their roles and responsibilities. Monitoring compliance with the DMP and conducting audits helps ensure adherence to data management protocols. Adapting the DMP to meet new requirements and staying informed about new standards and technologies ensures the plan remains relevant.

Can a question be considered to be the conclusion? Is it even a conclusion, then? In this case, yes. The awareness of the current situation, the lost potential due to inadequate data, and the overall complexity of the problem are necessary starting points for solving it.

REFERENCES

Baldin M., Breunig T., Cue R., De Vries A., Doornink M., Drevenak J., Fourdraine R., George R., Goodling R., Greenfield R., Jorgensen M.W., Lenkaitis A., Reinemann D., Saha A., Sankaraiah C., Shahinfar S., Siberski C., Wade K.M., Zhang F., Cabrera V.E. (2021): Integrated decision support systems (IDSS) for dairy farming: a discussion on how to improve their sustained adoption. Animals, 11: 2017–2025.

Beegle K., Carletto C., Himelein K. (2012): Reliability of recall in agricultural data. Journal of Development Economics, 98: 34–41.

Braun A.T., Colangelo E., Steckel T. (2018): Farming in the Era of Industrie 4.0. Procedia CIRP, 72: 979–984.

Chen D., Wu B., Chen T., Dong J. (2017): Development of distributed data sharing platform for multi-source IOT sensor data of agriculture and forestry. Transactions of the Chinese Society of Agricultural Engineering, 33: 300–307.

Da Silveira F., da Silva S.L.C., Machado F.M., Barbedo J.G.A., Amaral F.G. (2023): Farmers' perception of the barriers that hinder the implementation of agriculture 4.0. Agricultural Systems, 208: 103656.

Estes-Zumpf W., Addis B., Marsicek B., Lee M., Nelson Z., Murphy M. (2022): Improving the sustainability of long-term amphibian monitoring: the value of collaboration and community science for indicator species management. Ecological Indicators, 134: 108451.

Feng H., Abagandura G.O., Senturklu S., Landblom D.G., Lai L., Ringwall K., Kumar S. (2020): Soil quality indicators as influenced by 5-year diversified and monoculture cropping systems. The Journal of Agricultural Science, 158: 594–605.

European Union (2018): Food safety – EU action. Available at: https://european-union.europa.eu/priorities-and-actions/actions-topic/food-safety_en

Hwang C., Kim H., Jung H. (2018): Detection and correction method of erroneous data using quantile pattern and LSTM. Journal of Information and Communication Convergence Engineering, 16: 242–247.

Jonák Z. (2003a): Data. Czech terminological database of librarianship and information science (TDLIS). National Library of the Czech Republic. Available at: https://aleph.nkp.cz/F/?func=direct&doc_number=000000442&local_base=KTD

Jonák Z. (2003b): Information. Czech terminological database of librarianship and information science (TDLIS). National Library of the Czech Republic. Available at: https://aleph.nkp.cz/F/?func=direct&doc_number=000000456&local_base=KTD

Jonák Z. (2003c): Piece of knowledge. Czech terminological database of librarianship and information science (TDLIS). National Library of the Czech Republic. Available at: https://aleph.nkp.cz/F/?func=direct&doc_number=000000484&local_base=KTD

Jonák Z. (2003d): Knowledge. Czech terminological database of librarianship and information science (TDLIS). National Library of the Czech Republic. Available at: https://aleph.nkp.cz/F/?func=direct&doc_number=000000498&local_base=KTD

Kar S., Garin V., Kholová J., Vadez V., Durbha S.S., Tanaka R., Iwata H., Urban M.O., Adinarayana J. (2020): SpaTemHTP: a data analysis pipeline for efficient processing and utilization of temporal high-throughput phenotyping data. Frontiers in Plant Science, 11: 552509.

https://doi.org/10.17221/361/2024-PSE

Oladele O.I., Fawole O.P. (2007): Farmers perception of the relevance of agriculture technologies in South-Western Nigeria. Journal of Human Ecology, 21: 191–194.

Omar N., Zen H., Nicole N., Waluyo W. (2020): View of accuracy and reliability of data in IoT system for smart agriculture. International Journal of Integrated Engineering, 12: 105–116.., Rodný T., Okenka I. (2012): Precise irrigation process support by using a computer based algorithm of heuristics. Agris On-Line Papers in Economics and Informatics, 4: 49–54.

Stočes M., Šilerová E., Vaněk J., Jarolímek J., Šimek P. (2018): Posibilities of using open data in sugar and sugar beet sector. Sugar and Sugar Beet Sector, 134: 117–121.

Sylvere N., Jean D'amour R. (2020): Updates on modern agricultural technologies adoption and its impacts on the improvement of agricultural activities in Rwanda: a review. International Journal of Innovative Science and Research Technology, 5: 222–229.

Tantalaki N., Souravlas S., Roumeliotis M. (2019): Data-driven decision making in precision agriculture: the rise of big data in agricultural systems. Journal of Agricultural and Food Information, 20: 344–380.

Trapanese L., Hostens M., Salzano A., Pasquino N. (2024): Short review of current limits and challenges of application of machine learning algorithms in the dairy sector. Acta IMEKO, 13: 1–7.

Received: July 7, 2024
Accepted: September 9, 2024
Published online: October 4, 2024