Czech University of Life Sciences Prague
Faculty of Economics and Management

agris on-line

Papers in Economics and Informatics

http://online.agris.cz

ISSN 1804-1930
XVIII, 2026, 3

International scientific journal
Prague

Agris on-line Papers in Economics and Informatics

The international reviewed scientific journal issued by the Faculty of Economics and Management of the Czech University of Life Sciences Prague.

The journal publishes original scientific contributions from the area of economics and informatics with focus on agriculture and rural development.

Editorial office

AGRIS on-line Papers in Economics and Informatics
Department of Economics of FEM CZU Prague
Kamýcká 129, 165 00 Praha-Suchdol
Czech Republic
Phone: +420 224 382 056
E-mail: agrisonline(at)pef.czu.cz

Publisher

Faculty of Economics and Management
Czech University of Life Sciences Prague
Kamýcká 129, 165 00 Praha-Suchdol
Czech Republic
Reg. number: 60460709

ISSN 1804-1930

XVIII, 2026,
30th of September 2026
Prague

Agris on-line
Papers in Economics and Informatics

ISSN 1804-1930
XVIII, 2026, 3

Agris on-line Papers in Economics and Informatics

Volume XVIII Number 3, 2026

Content:

Volume XVIII Number 3, 2026

Smart Agriculture with Machine Learning and Internet of Things Techniques: Literature Review and Bibliometric Analysis

Karima El Majdouby[1], Mohamed El Mohadab[2, 3], Karim Abouelmehdi[1]

[1] ELITES laboratory, Faculty of Sciences, University of Chouaib Doukkali, El Jadida, Morocco

[2] Laboratory of Information Processing and Decision Support, Higher School of Technology Beni Mellal, Sultan Moulay Slimane University, Beni Mellal, Morocco

[3] Laboratory of Information Processing and Decision Support, Faculty of Sciences and Technology, Sultan Moulay Slimane University, Morocco

Abstract

A booming field known as "smart agriculture" has emerged because of the rapid revolution and transformation of agriculture. It makes use of cutting-edge agricultural technologies, like the internet of things and machine learning (ML) algorithms, to enhance farm management tasks, boost agricultural productivity, and lessen environmental impact by empowering farmers to react swiftly and efficiently to climate change. A key component of the shift to smart farming is the Internet of Things (IoT), which makes it easier for sensors and devices to connect and share data. It also gives farmers access to actionable information and knowledge, enabling them to make well-informed decisions using machine learning algorithms to enhance crop management and boost yields. Numerous tasks, including weed identification, fertilization requirement analysis, irrigation adjustment, and pest and soil management, are made possible by the automation made possible by these technologies. They allow farmers to make targeted interventions by continuously monitoring crop health, which lowers labor costs and the environmental impact. The effects of technological advancements, such as machine learning and the Internet of Things, on the agriculture industry are thoroughly examined in this paper. It draws attention to the useful applications of these instruments, which help farmers better, more automatically, and more precisely manage their resources and output..

Keywords

Smart agriculture, machine learning, internet of things, bibliometric.

El Majdouby, K., El Mohadab, M. and Abouelmehdi, K. (2026) "Smart Agriculture with Machine Learning and Internet of Things Techniques: Literature Review and Bibliometric Analysis", *AGRIS on-line Papers in Economics and Informatics*, Vol. 18, No. 3, pp. 3-16. ISSN 1804-1930. DOI 10.7160/aol.2026.180301.

Introduction

A major revolution has recently taken place in agriculture. Farmers have begun to implement some advanced agricultural techniques and technologies to optimize the efficiency of their daily tasks. They suggest innovative ways of integrating artificial intelligence (AI) and Internet of Things (IoT) into farming activities to stimulate growth and increase production yields (Kasera and Acharjee, 2024). The advent of digital technologies such as the Internet of Things (IoT) and Machine Learning (ML) is radically transforming human food production techniques, giving rise to what is known as smart farming.

The Internet of Things (IoT) is an important technology that offers efficient and reliable solutions in a variety of agricultural practices, (IoT) plays a central role in smart farming by giving farmers the ability to effectively monitor and regulate their products in real-time and with remote management such as irrigation and crop and livestock health diagnostics using several sensors that can detect parameters such as soil NPK, humidity, nitrate, pH, electrical conductivity, temperature, CO_2, light, humidity, water level, weather station (Morchid et al., 2024). Intelligent agriculture uses technological advances that combine the Internet of Things (IoT) and machine learning (ML) to increase production, reduce waste and costs, and reduce the environmental footprint (Ribeiro Junior et al., 2022). Machine learning enables farmers to solve problems such as pests, optimize irrigation, and facilitate informed decisions on crop selection and predictive modeling of future challenges.

Intelligent agriculture using the Internet of Things (IoT) and artificial intelligence techniques in the agricultural sector responds to many challenges, providing a better understanding of soil, crops, and climatic variations. System sensors help monitor and manage resources, while machine learning approaches based on this collected data facilitate crop prediction. This enables sustainable agriculture by guiding farmers towards the right crop for each season, guaranteeing a significant increase in production, while offering low-cost remote monitoring.

This paper proposes a systematic review and bibliometric analysis focusing on the use of artificial intelligence and the Internet of Things in the smart agriculture sector to deepen current knowledge. The rest of the document is structured into sections as follows: Section 2 discusses the definition of smart agriculture, Section 3 describes modern technologies integrated into the agricultural domain, such as the Internet of Things and machine learning, and then Section 4 presents preferred Internet of Things and machine learning applications in agricultural operations. Section 5 presents the methodology followed to select and analyze articles collected in various databases, then the sixth part deals with bibliometrics. Related work on agriculture is presented in Section 6, and a conclusion is offered in the final section. The structure of our paper, including all of the contributions, is shown in Figure 1.

Materials and methods

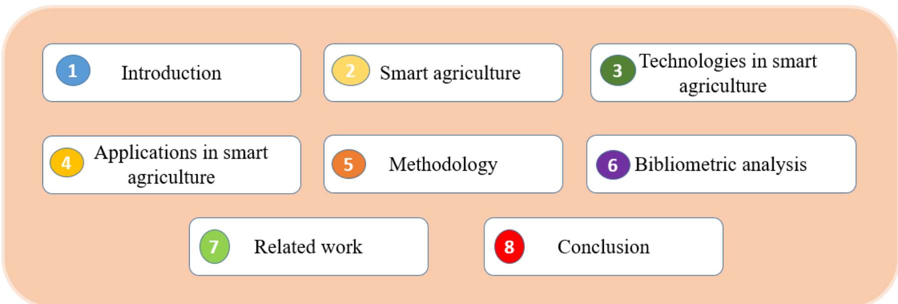
Machine learning

Artificial intelligence can be found in virtually every sector, from agriculture to healthcare. Machine learning is a branch of artificial intelligence that involves training machines to a degree of autonomy that enables them to predict outcomes independently. In agriculture, a range of applications based on artificial intelligence are feasible, such as agricultural robots that surpass human performance in managing crop and soil monitoring, as well as climate change forecasting (Liakos et al., 2018). The fundamental principle of artificial intelligence in agriculture is flexibility, rapid efficiency, accuracy and cost-effectiveness. Artificial intelligence in the agricultural sector not only supports farmers in improving their farming skills, it also involves increasing their yield and ensuring quality while reducing costs. The application of artificial intelligence to wireless sensors boosts efficiency in a variety of sectors, while offering methods for the challenges faced by many branches of agriculture, such as harvesting, irrigation and soil sensitivity.

Artificial intelligence can identify plant diseases, pest invasions, and nutrient shortages on farms. In addition, AI- enabled sensors can monitor and manage agricultural indicators (Murugamani et al., 2022). ML applications are used in many fields, such as land management, crop health monitoring, disease detection, and weed infestation control (Abbasi et al., 2022).

ML is a data-driven method that develops a model capable of generalizing by learning rules and knowledge from large datasets. Thanks to their ability to analyze large volumes of data collected by IoT sensors, machine learning algorithms facilitate the optimization of agricultural processes, resulting in increased productivity while minimizing resource consumption. Machine learning techniques outperform all traditional methods, offering a cost-effective and efficient solution for agricultural activities (Aldossary et al., 2024).

Machine learning is used in a variety of fields, from image detection and linguistic analysis to forecasting and autonomous systems



Source: Authors

Figure 1: Organization of the paper.

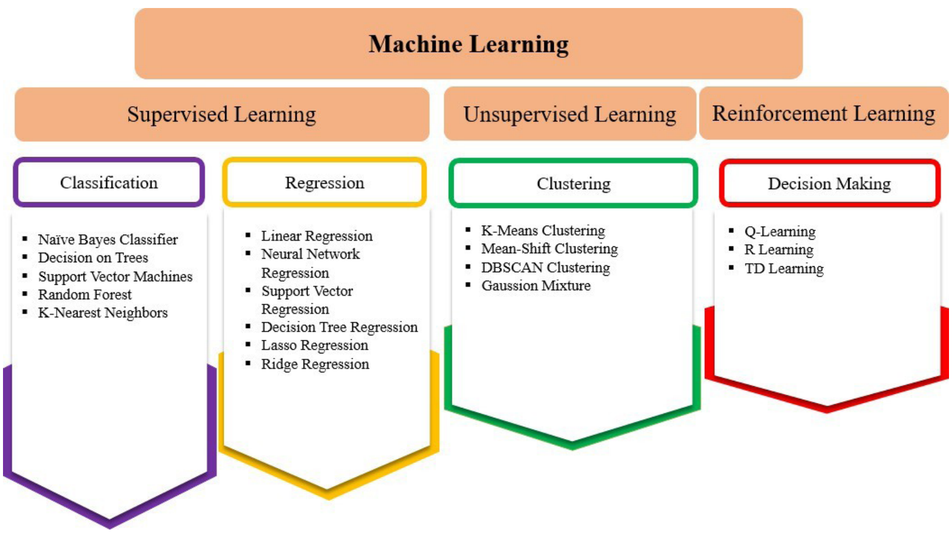
(Wang et al., 2025). Machine learning aims to create algorithms and models that enable machines to learn, predict or decide without the need for precise programming. Machine learning models analyze the information provided, detect consequences and use statistical analysis to make predictions or classifications (Choudhary et al., 2025). This field includes a variety of algorithms, classified into three main categories: 1) supervised learning (SL), where models are trained on labeled data; 2) unsupervised learning (USL), which operates on unlabeled data; and finally, 3) reinforcement learning (RL), which focuses on sequential decision making (El Majdouby et al., 2024).

- Supervised learning is a technique that utilizes models trained on a set of labeled data to predict an outcome based on an input. This method is used to design machine learning models for regression (when the results are numerical) and classification (when the results are categorical) (Staff, 2025).
- Semi-supervised learning is a machine learning approach that combines labeled and unlabeled data to train algorithms, aiming to enhance the efficiency of predictive models. Indeed, algorithms start training with a limited amount of labeled data to initiate their learning, and are then progressively fed with a much larger volume of unlabeled data to improve the model (Fu et al., 2025).
- Reinforcement learning consists of training artificial intelligence models to perform a sequence of choices in which an autonomous agent learns to decide by interacting with its environment through trial and error based on its experiences and actions. This genre is often used in fields such as robotics, interactive video games, or autonomous cars, where the agent must adapt to constantly changing situations (Saad Alotaibi et al., 2024).

Machine learning (ML) is an innovative method that enables machines to simulate the learning activities of humans, continuously improving performance and achieving set goals. Machine learning (ML) techniques are applied on agricultural farms for disease detection, weather forecasting, irrigation scheduling, weed detection, crop quality, and soil conditions to reduce crop costs and maximize yields (Maduranga and Abeysekera, 2020). Major ML algorithms applied in IoT-based agriculture major ML algorithms can be applied such as support vector machine (SVM), naive Bayes, fuzzy clustering, discriminant analysis, K-means clustering, K-nearest neighbor (K-NN), Gaussian mixture models, artificial neural networks (ANN), decision making, and deep learning. Figure 2 illustrates the structure of ML classification, specifying the different divisions of machine learning.

Machine learning (ML) has several uses in agriculture, such as:

- Disease detection: Conventional approaches to identifying the type and severity of leaf diseases frequently rely on human observation and expert



Source: Authors

Figure 2: Categories of machine learning.

judgment. However, thanks to the evolution of deep learning methods, it is possible to improve the accuracy and efficiency of detecting complex pathologies by processing leaf images (Wang et al., 2025).

- Crop management: The inclusion of machine learning algorithms in the agricultural sector to anticipate future harvests has produced positive results in helping to better understand the influence of climate and agricultural conduct on future yield, while facilitating the management of problems that lead to poor agricultural performance and food shortages linked to low yields (Zhang et al., 2025).
- Irrigation planning: Machine learning-based irrigation programs can autonomously calculate the ideal volume of water, using current weather data to guide the irrigation of agricultural fields (Gao et al., 2023).
- Weed detection: In agriculture, machine learning is crucial in identifying weeds, their position in each row, and their spacing, thus promoting optimum productivity. Machine learning models are based on images of cultivated fields, soils, and weeds (Kumar Nagothu et al., 2023).
- Crop quality assessment: Abiotic stresses, which threaten food security worldwide, are among the reasons for crop failure. However, thanks to deep learning (DL) methods, rapid identification and accurate prediction of indicators of these abiotic stresses are now possible (Subeesh and Chauhan, 2025).
- Weather forecasting: Machine learning models are an effective way of predicting future weather conditions, such as wind speed, temperature, humidity, and pressure, to optimize the productivity of agricultural crops (Kumari and Muthulakshmi, 2024).
- Water management: Machine learning (ML) techniques facilitate real-time monitoring of water quality and prediction of parameters such as microbial indicators, heavy metals, and chlorine (Li et al., 2024).

Internet of things

The concept of the Internet of Things refers to the interconnection of physical devices through sensors, networks, and data analytics that facilitate the collection and analysis of data from sensors placed on the farm site, monitoring various criteria such as soil moisture, nutrient concentrations, temperature, humidity levels, and even plant well-being. This information guides farmers in improving techniques such as fertilization, pest control, and irrigation, leading to increased production while making more efficient use of available resources (Choudhary et al., 2025). IoT integration in agriculture aims to provide farmers with the tools they need to adopt automation strategies and improve their decision-making. These technologies increase crop quality, productivity, and profitability by effectively integrating data, expertise, and a variety of services. Intelligent algorithms and sensors can be utilized to precisely recommend methods for preserving field quality and ensuring high-quality harvests (Venkatesh et al., 2022).

The Internet of Things (IoT) refers to a network of digital devices, sensors, machines, and equipment interconnected over the web, each with its own identity and the ability to observe and control remotely (Javaid et al., 2025). The IoT is based on a five-level structure: the perception layer (hardware devices), the network layer (information transmission), the middleware layer (device management and interoperability capability), and the service layer (cloud computing) and the application layer (information integration and analysis). Finally, we reach the end-user layer, which provides the user interface (Razaque et al., 2025).

- Sensor layer: This physical layer is mostly made up of hardware components, such as actuators and sensors. These devices are used to gather environmental data, including brightness, wind direction, wind speed, soil moisture, pH, temperature, air pressure, and humidity (Lefoane et al., 2025).
- Network layer: Through communication protocols like Bluetooth, WiFi, 3G, GSM, and others, sensors transmit the data they have collected. Sending data to the cloud for remote control or analysis at a later time is made simple by this layer (El Majdouby et al., 2024).
- Data processing layer: a layer where gathered data is processed and stored, typically situated between the application and the network. It provides the ability to analyze data more thoroughly and make decisions automatically (Sarohe et al., 2025).

- Service layer: Data processed and stored in the previous layer can be sent to a cloud environment. This layer makes it possible to centralize data and offer scalable services, which is essential for managing the large quantities of data generated by connected objects. The cloud can also host artificial intelligence services for advanced analysis (Huo et al., 2024).

- Application layer: Applications and user interfaces are developed to enable interaction with the end- user. It shows end-users information about the data collected from the previous layer to help them make decisions. The application layer deals with the creation of efficient and useful applications for services, and the user can control IoT devices via a smartphone application, PC, tablet, etc. (El Majdouby et al., 2024). Figure 3 shows the four layers of Internet of Things technology.

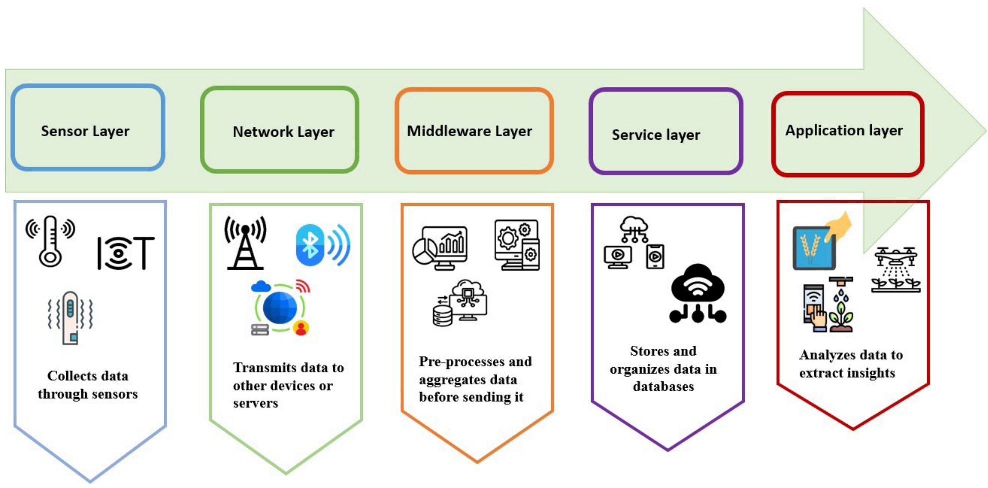
In smart farming, various IoT sensors have been employed to gather information on various agricultural aspects, including soil, crop, environment, and more. For substrate monitoring, electronic sensors were used to collect substrate data such as soil and water temperature, humidity, and nitrogen levels. Also, pH sensors were used to assess the level of acidity or alkalinity of the water in hydroponic cultures. In addition, for environmental monitoring, electronic sensors are integrated into IoT solutions to gather environmental information such as temperature, luminosity, and humidity. For crop monitoring, multispectral sensors and cameras have been used to gather images of crops. These devices can be mounted on a drone to obtain aerial photos of vast crops or integrated into robots to achieve precise capture of a plant's leaves (Navarro et al., 2020).

Applications in smart agriculture

In recent years, new solutions and technologies have been introduced in the agricultural sector to increase crop production. A variety of IoT applications for smart farming help farmers make good decisions about their crops and automate farm management to reduce labor costs (Quy et al., 2022). Intelligent agriculture consists of a range of services that could be provided, for example, irrigation management, disease management, input management (pesticides, fungicides, and herbicides), yield prediction, remote monitoring, farm machinery control, product traceability, vehicle guidance, animal tracking...

- Monitoring of climate conditions: IoT facilitates the real-time collection and analysis of agricultural soil composition and the transmission of information to farmers or landowners via the Internet (soil condition, temperature, and humidity, as well as forecasts of natural factors such as precipitation and weather). The data collected was also used to automate farm management to reduce labor costs.

- Precision farming: In smart precision farming, we find the use of drones in monitoring and farming activities. Some common agricultural tasks using drones include pesticide spraying, seed sowing, fertilization, crop evaluation and mapping, and crop growth monitoring (Zhou



Source: Authors

Figure 3: The architecture of the Internet of Things.

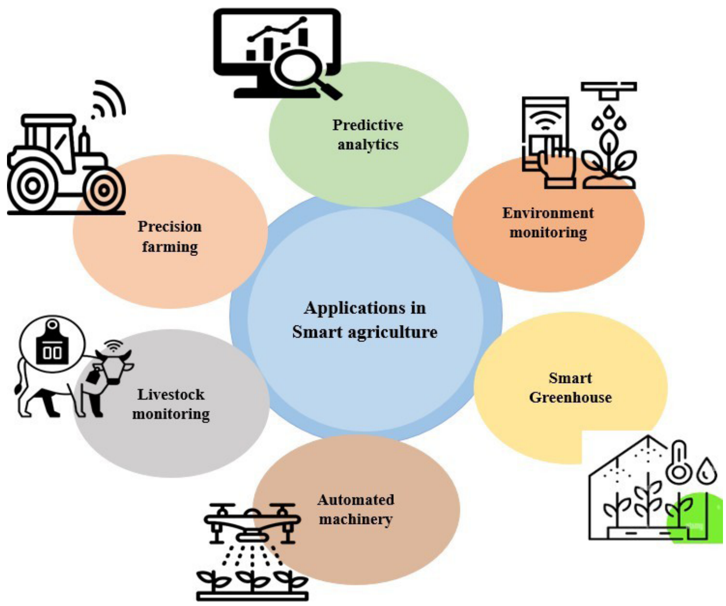
and Zhou, 2020). Precision farming relies on technologies such as irrigation systems, UAVs (unmanned aerial vehicles), and intelligent farm machinery, which can be programmed to operate autonomously according to various criteria or controlled remotely by the farmer via the Internet. To achieve healthier, more productive crops while minimizing costs, farmers can consult various parameters such as light levels, temperature, soil conditions, humidity, and pest infestations to determine the ideal quantities of water, fertilizers, and pesticides needed for their plantations.

- Livestock monitoring: There are IoT sensors specifically designed for agriculture, which can be attached to farm animals to monitor their health and record their performance. These devices can monitor and collect information relating to the health, well-being, and geographical position of animals. The use of intelligent agricultural sensors (collar tags) provides detailed information on the health, temperature, activity, and nutrition of each animal, as well as collective information on the herd (Shalimov, 2023).

- Smart greenhouse management: Greenhouse agriculture is considered the oldest method of smart farming, but it involves precise monitoring of environmental parameters to manage and maintain the local climate. Using an IoT-based prototype, this method monitors greenhouses to measure internal parameters and obtain detailed, real-time information on greenhouse conditions, including illuminance, soil condition, temperature, humidity, and pressure (Ayaz et al., 2019).

- Agricultural drones: The use of drones (unmanned aerial vehicles) is particularly well suited to agricultural data collection, thanks to their high surveillance capacity. Drones can also perform a multitude of activities that previously required human intervention: plant cultivation, pest and infection control, spraying of agricultural products, crop monitoring.

- Predictive analysis: IoT is a technology based on intelligent sensors that provide crucial data in real time. Exploiting the analysis of this data enables farmers to better understand it and make significant forecasts: crop harvesting time, production quantity, risks of disease and infestation, etc. Farmers can anticipate the volume and quality of their crops, as well as their vulnerability to adverse weather conditions, such as flooding and drought. It also gives growers the ability to optimize the supply of water and nutrients for each plant, while allowing them to select yield characteristics to improve quality.



Source: Authors

Figure 4: Applications in smart agriculture.

- Smart irrigation and water management: Farmers can maintain proper humidity levels while maximizing the effectiveness of the resources used by using smart irrigation systems and Internet of Things (IoT) technology to monitor water. Drip irrigation systems can be automatically managed based on region, crop cycle, and growth phases thanks to the use of soil sensors and weather data. One of the best digital agriculture solutions provided by the Internet of Things is intelligent irrigation. Real-time soil property monitoring is made possible by the placement of sensors in fields (Razaque et al., 2025). In order to ensure more effective water management, this provides farmers with the chance to enhance and optimize their irrigation systems while accounting for the crops, local weather patterns, and soil types. While guaranteeing that every plant receives the precise moisture needed for maximum production, this resource-saving technology reduces water waste.

This study's primary goal is to systematically compile recent research on smart agriculture that integrates Internet of Things (IoT) and artificial intelligence (AI) technologies, such as machine learning (ML). This study aims to investigate the ways in which these advanced technologies have been used in the agricultural industry, particularly for managing livestock, administering land, identifying weeds, controlling plant diseases, and optimizing irrigation.

Our research methodology used in this systematic review was inspired by the SLR process suggested by Kitchenham and Charters (Kitchenham, no date). It consists of the following sequential steps (Figure 5).

Research questions

This in-depth literature review aims to outline the major issues and trends in smart agriculture, focusing on the associated technologies, applications, and consequences. Seven research questions are presented in Table 1 to address the objectives of this study.

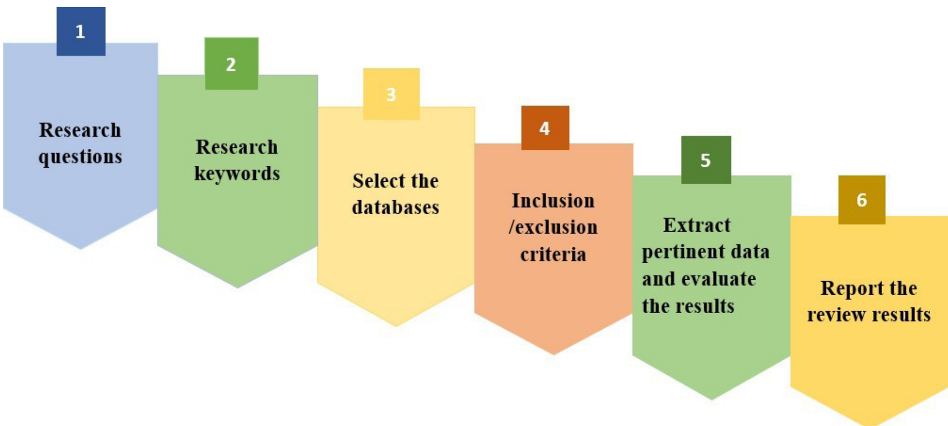
No	Research Question (RQ)
RQ1	Which ML techniques have been used in smart agriculture?
RQ2	Which Internet of Things and machine learning applications are most favored in intelligent agriculture?
RQ3	Which machine Learning|IoT techniques are well-suited to smart agriculture?
RQ4	What is the overall performance of ML techniques for intelligent agriculture?
RQ5	What is the benefit of combining IoT and Machine Learning in environmental applications?
RQ6	Which countries are the most influential in the field of smart agriculture?
RQ7	What trends and advancements may we expect in intelligent agriculture in the future?

Source: Authors

Table 1: Research questions of the study.

Data sources

This systematic review relies on an in-depth search to gather significant articles by exploiting scientific databases and indexing services such as Scopus and Web of Science, to identify conference



Source: Authors

Figure 5: Systematic review process.

and journal articles related to agricultural technologies and innovations.

Keyword selection strategy

This study used several keyword combinations to locate relevant literature related to smart agriculture, including terms such as "smart farming", "precision agriculture", "smart farming", "machine learning", "artificial intelligence", and "internet of things". These combined terms were employed in different configurations, using logical operators (AND, OR, NOT) to refine results and optimize the relevance of selected searches. This method favored the inclusion of various technical, theoretical, and practical studies, while excluding works that lacked relevance or were too far removed from the subject.

Inclusion criteria

- To ensure the relevance and quality of the studies included, several elements must be specified:
- Only studies that integrate ML and IoT in the context of smart agriculture.
- The study is published in a peer-reviewed journal. The study is written in English.
- The studies focus on the use of IoT sensors to collect agricultural data and the use of machine learning algorithms to analyze and make decisions.
- Only research carried out between 2020 and 2024 will be considered to incorporate recent work in the agricultural field.

Exclusion criteria

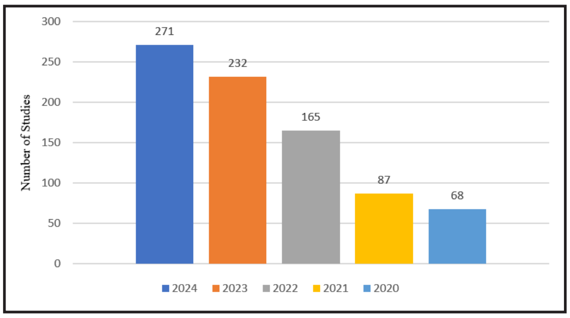
Studies were excluded based on the following criteria:

- The study does not focus on ML and IoT integration.
- Studies not written in English are excluded.
- The study is a conference abstract, editorial, letter, or opinion piece with no original research data.

Results and discussion

Bibliometric analysis

Scopus database

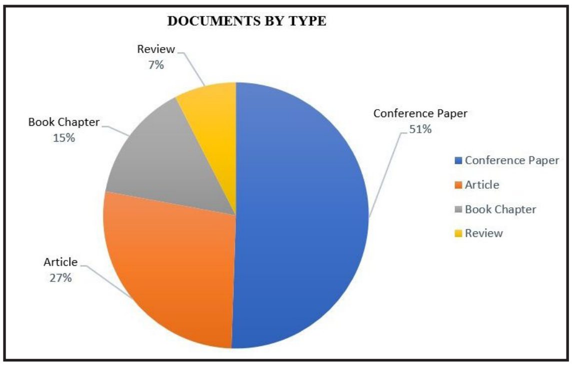


Source: Authors

Figure 6: Documents published in Scopus.

The Scopus database contains a total of 823 research articles focusing on the use of relevant keywords. The following figure 6 illustrates the annual progression of publications from 2020 to 2024, indicating a growth in the number of works related to the integration of IoT and ML in the smart agriculture sector. The surge in recently published articles reflects the growing debate and relevance of the smart agriculture theme.

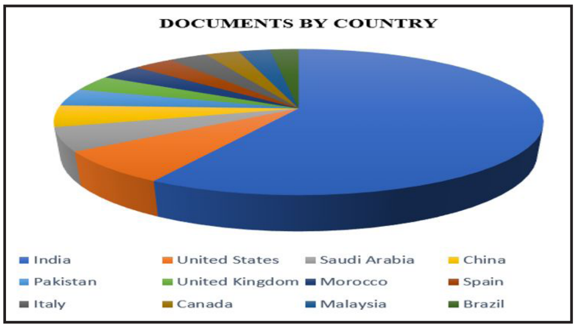
The publications found on smart agriculture were divided into various documents: Conference paper (51%, n = 416), Article (27%, n = 225), Review (7%, n = 62), and Book chapter (15%, n = 120) as shown in Figure 7.



Source: Authors

Figure 7: Type of documents published in Scopus.

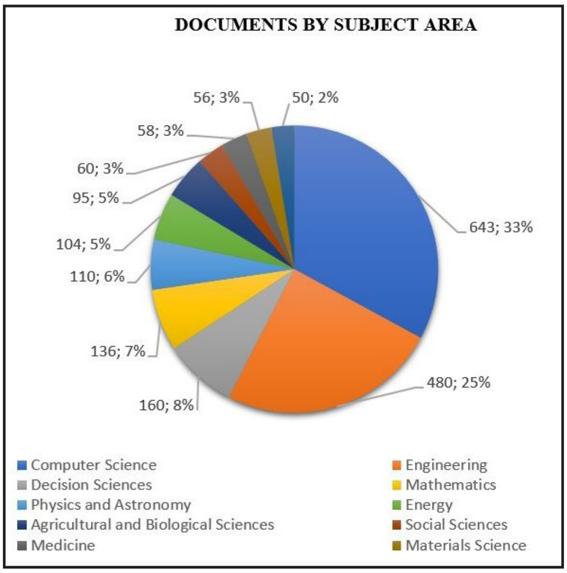
The graph below (Figure 8) shows that India leads the way in the production of documents relating to technological agriculture, followed by other countries such as the United States, Saudi Arabia, China, Pakistan, the United Kingdom, Morocco, and various other countries focused on agricultural research.



Source: Authors

Figure 8: Documents contributed by country.

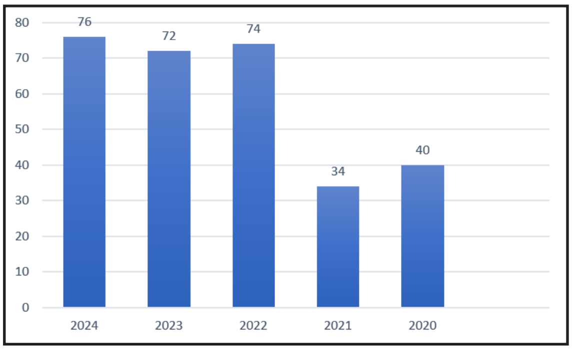
More than 25% of the works listed in the Scopus database focus on computer science and engineering, reflecting a predominant preference for technological dimensions.



Source: Authors

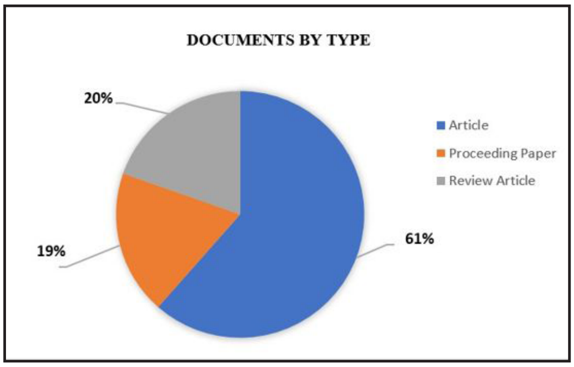
Figure 9: Documents contributed by subject.

The data used in this analysis comes from the Web of Science (WOS) database, and the same filter was applied to the Scopus database. We found 296 academic publications: 61% articles, 19% proceedings papers, and 20% review articles (Figure 11) on the concept of smart agriculture. The integration of ML-based algorithms into IoT systems for various applications has grown exponentially in recent years. This is reflected in the annual scientific output shown in Figure 10. This is a clear sign that researchers are considering the joint use of ML/ AI and IoT to solve farmers' problems.



Source: Authors

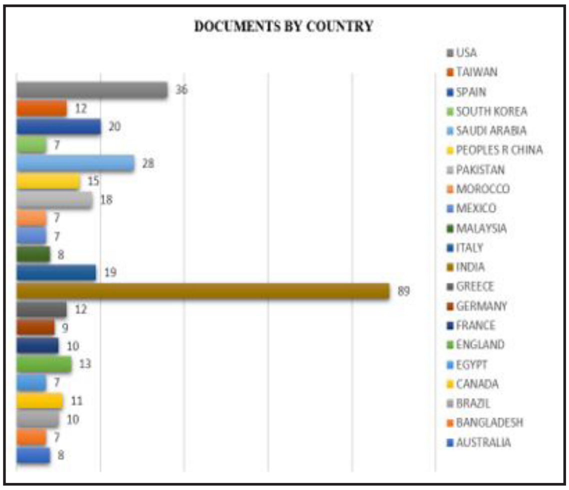
Figure 10: Number of articles by years of publication in WOS.



Source: Authors

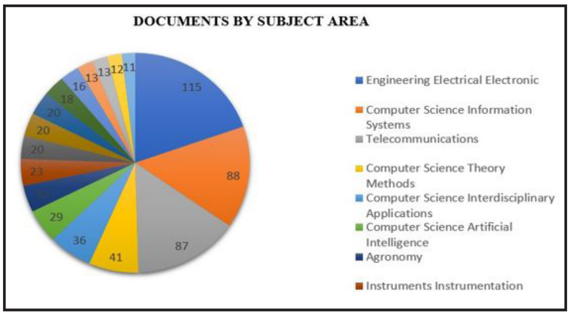
Figure 11: Document type published in WOS.

The graphs below show, the main countries involved in in-depth agricultural research are India, the USA, and Saudi Arabia, followed by a number of other countries. However, researchers are showing a particular interest in the fields of computing, engineering, and telecommunications, indicating that the world's attention is increasingly focused on new technologies and the it sector.



Source: Authors

Figure 12: Document published by the country in WOS.



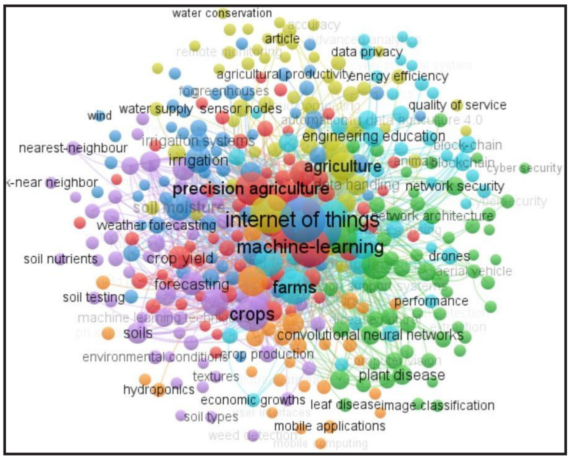
Source: Authors

Figure 13: Documents contributed by the Subject area in WOS.

VOSviewer analysis

Bibliometric analysis is a crucial tool for measuring the impact and evolution of scientific literature. In recent years, this method has been widely applied in various areas of academic research, mobilizing visual knowledge to illustrate the current situation, development, and emerging trends in a specific field (Chen et al., 2025).VOSviewer provides in-depth bibliographic data, facilitating the identification of current and fast-growing research topics to create relational network diagrams for comprehensive, systematic, and unbiased analysis.

In VOSviewer, the minimum occurrence of keywords has been set at 5, enabling a total of 361 elements to be visualized. This network Figure 14 visualizes the relationships between keywords, with larger nodes representing the most frequent keywords. Keyword occurrences are indicated by the size of the circles, which reflects the importance of the keyword, and the connections between elements represent the relationships. The network highlights the importance of terms such as "internet of things" and "precision agriculture", exposing their strong associations with related notions such as "artificial intelligence", and ML algorithms in agriculture, and the proximity of these two indicates a degree of correlation in terms of strength and thematic relevance. The findings in Figure 14 show that these subjects are highly valued and have a significant impact on the advancement of smart agriculture. This enables us to create a map by grouping the keywords into seven clusters, which are dispersed as follows: Cluster 1 and Cluster 2 each contain 60 items, 57 items, 56 items, 55 items, 45 items, and 28 items, respectively.

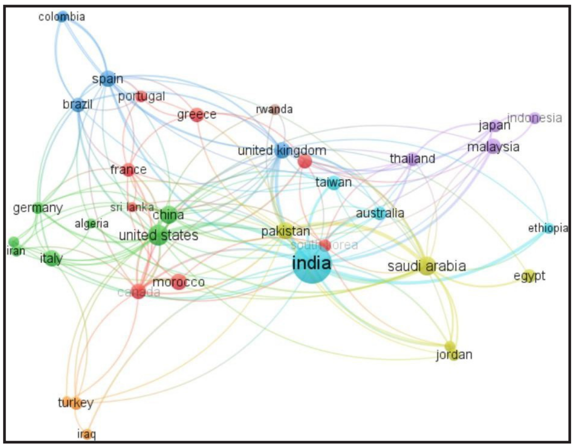


Source: Authors

Figure 14: Keyword occurrence in Scopus.

Figure 15 explicitly demonstrates the geospatial distribution of selected articles from various countries/regions within the research area examined. The nations/regions are segmented into various interconnected nodes, with India capturing the most essential node. In this network of cooperation, the USA, along with a few countries such as Saudi Arabia, China, Pakistan, and the UK, are considered the six main contributors in this sector.

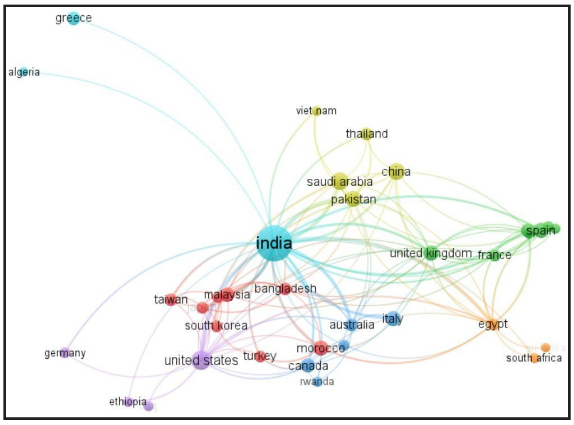
Furthermore, in terms of connecting lines, India is the country with the highest number of links to other nations/regions. Eight groups have been divided up as follows: Cluster 1 consists of eight items, Cluster 2 of seven, Cluster 3, 4, five, 6, four, three, and eight items, and Cluster 8 of one item.



Source: Authors

Figure 15: Keyword visualization by countries in Scopus.

The visualization of citations by country in Scopus makes it possible to identify the leading countries in terms of their strong scientific output and to analyze the geographical distribution of scientific contributions and the impact of publications worldwide. The graphical representation in Figure 16 provides an overview of publication trends, international collaborations, and the geographical areas where research is most influential. Countries that rank among the leaders in terms of scientific citations. Seven clusters are distributed as follows: Eight pieces make up Cluster 1, seven pieces make up Cluster 2, four pieces make up Cluster 3, four pieces make up Cluster 5, and seven pieces make up Cluster 6.



Source: Authors

Figure 16: Citation visualization by countries in Scopus.

Conclusion

In conclusion, this study demonstrates that the integration of innovative agricultural management techniques based on the Internet of Things (IoT) and artificial intelligence represents a modern, effective, predictive, and automated approach that has helped refine agricultural practices to optimize crop production and, consequently, increase yields. These innovations are transforming the agricultural sector, paving the way for smarter and more sustainable farming in various areas such as disease and weed detection, crop and livestock monitoring, crop productivity assessment, improved irrigation, soil and water conservation, crop yield estimation, as well as robotics and automation. This bibliometric study highlights the considerable potential of these innovations: IoT sensors automatically collect agricultural data such as humidity, soil moisture, and temperature in real time, and facilitate connectivity and information exchange between various devices, while ML algorithms analyze this data to predict ideal environmental conditions and optimize farmers' decision-making processes. Although this literature review emphasizes the importance of these technologies in smart agriculture and provides an overview of machine learning and IoT technologies applied to smart agriculture, as well as their application in the agricultural sector to make it more efficient and environmentally friendly. This study examines the dynamic evolution of global publications on smart agriculture over the past few years, from 2020 to 2024. During this period, the topic of smart agriculture has gained increasing prominence, as evidenced by the number of scientific publications in the Scopus and Web of Science databases. A bibliometric analysis provides insight into the evolution of this research field.

Corresponding author:
Karima El majdouby
Computer Science department, ELITES laboratory, Faculty of Sciences
University of Chouaib Doukkali, Avenue Jabran Khalil Jabran, El Jadida, Morocco
Phone: +212 6 46 453022, Email: elmajdouby.karima@ucd.ac.ma

References

[1] Abbasi, R., Martinez, P. and Ahmad, R. (2022) "The Digitization of Agricultural Industry - A Systematic Literature Review on Agriculture 4.0", *Smart Agricultural Technology*, Vol. 2, p. 100042. ISSN 2772-3752. DOI 10.1016/j.atech.2022.100042.

[2] Aldossary, M., Alharbi, H. A. and Anwar Ul Hassan, C. (2024) "Internet of Things (IoT)-Enabled Machine Learning Models for Efficient Monitoring of Smart Agriculture", *IEEE Access*, Vol. 12, pp. 75718-75734. ISSN 2169-3536. DOI 10.1109/ACCESS.2024.3404651.

[3] Ayaz, M., Uddin, A., Sharif, Z., Mansour, A. and Aggoune, E.-H. M. (2019) "Internet-of-Things (IoT)-Based Smart Agriculture: Toward Making the Fields Talk", *IEEE Access*, Vol. 7, pp. 129551-129583. ISSN 2169-3536. DOI 10.1109/ACCESS.2019.2932609.

[4] Chen, X., Tian, W. and Fang, H. (2025) "Bibliometric Analysis of Natural Language Processing Using CiteSpace and VOSviewer", *Natural Language Processing Journal*, Vol. 10, p. 100123. ISSN 2949-7132. DOI 10.1016/j.nlp.2024.100123.

[5] Choudhary, V., Guha, P., Pau, G. and Mishra, S. (2025) "An Overview of Smart Agriculture Using Internet of Things (IoT) and Web Services", *Environmental and Sustainability Indicators*, Vol. 26, p. 100607. ISSN 2665-9727. DOI 10.1016/j.indic.2025.100607.

[6] El Majdouby, K., El Mohadab, M., Ait Daoud, R. and Abouelmehdi, K. (2024) "Exploring the Use of Machine Learning and the Internet of Things in Intelligent Agriculture: Short Survey", *2024 10th International Conference on Optimization and Applications (ICOA)*, pp. 1-6. DOI 10.1109/ICOA62581.2024.10754078.

[7] Fu, R., Chen, Z., Tian, H., Hu, J., Bu, F., Zheng, P., Chi, L., Xue, L., Jiang, Q., Li, L. and Zhu, L. (2025) "A Review on the Applications of Machine Learning in Biomaterials, Biomechanics, and Biomanufacturing for Tissue Engineering", *Smart Materials in Medicine*, Vol. 6, No. 2, pp. 171-204. ISSN 2590-1835. DOI 10.1016/j.smaim.2025.06.003.

[8] Gao, H. et al. (2023) "How Can Agricultural Water Production Be Promoted? A Review on Machine Learning for Irrigation", *Journal of Cleaner Production*, Vol. 414, p. 137687. ISSN 0959-6526. DOI 10.1016/j.jclepro.2023.137687.

[9] Huo, D., Zhangzhong, L., Zheng, W., Chen, G. (2024) "Mapping Smart Farming: Addressing Agricultural Challenges in Data-Driven Era", *Renewable and Sustainable Energy Reviews*, Vol. 189, p. 113858. ISSN 1364-0321. DOI 10.1016/j.rser.2023.113858.

[10] Javaid, M., Haleem, A., Sing, R. P., Suman, R., Khan, S. and Rab, S. (2025) "Adoption of Internet of Things for Smart City-Enabled Smart Hospitals", *Intelligent Hospital*, p. 100011. ISSN 2949-8716. DOI 10.1016/j.inhs.2025.100011.

[11] Kasera, R. K. and Acharjee, T. (2024) "A Comprehensive IoT Edge Based Smart Irrigation System for Tomato Cultivation", *Internet of Things*, Vol. 28, pp. 101356. E-ISSN 2542-6605, ISSN 2543-1536. DOI 10.1016/j.iot.2024.101356.

[12] Kitchenham, B. (2004) "*Procedures for Performing Systematic Reviews*", Corpus ID: 54019416. [Online]. Available: https://www.semanticscholar.org/paper/Procedures-for-Performing-Systematic-Reviews-Kitchenham/29890a936639862f45cb9a987dd599dce9759bf5 [Accessed: Feb. 15, 2026].

[13] Kumar Nagothu, S., Anitha, G., Siranthini, B., Anandi, V. and Prasad, P. S. (2023) "Weed Detection in Agriculture Crop Using Unmanned Aerial Vehicle and Machine Learning", *Materials Today: Proceedings*. ISSN 2214-7853. DOI 10.1016/j.matpr.2023.03.350.

[14] Kumari, S. and Muthulakshmi, P. (2024) "SARIMA Model: An Efficient Machine Learning Technique for Weather Forecasting", *Procedia Computer Science*, Vol. 235, pp. 656-670. ISSN 1877-0509. DOI 10.1016/j.procs.2024.04.064.

[15] Lefoane, M., Ghafir, I., Kabir, S. and Awan, I.-U.(2025) "Internet of Things Botnets: A Survey on Artificial Intelligence Based Detection Techniques", *Journal of Network and Computer Applications*, Vol. 236, p. 104110. ISSN 1084-8045. DOI 10.1016/j.jnca.2025.104110.

[16] Li, Z., Ma, W., Zhong, D., Ma, J., Zhankg, Q., Yuan, Y., Liu, X., Wang, X. and Zou, K. (2024) "Applications of Machine Learning in Drinking Water Quality Management: A Critical Review on Water Distribution System", *Journal of Cleaner Production*, Vol. 481, p. 144171. ISSN 0959-6526. DOI 10.1016/j.jclepro.2024.144171.

[17] Liakos, K. G., Busato, P., Moshou, D., Pearson, S. and Bochtis, D. (2018) "Machine Learning in Agriculture: A Review", *Sensors*, Vol. 18, No. 8, p. 2674. ISSN 1424-8220. DOI 10.3390/s18082674.

[18] Maduranga, M. W. P. and Abeysekera, R. (2020) "Machine Learning Applications in IoT Based Agriculture and Smart Farming: A Review", *International Journal of Engineering Applied Sciences and Technology*, Vol. 4, No. 12, pp. 24-27. ISSN 2455-2143. DOI 10.33564/IJEAST.2020.v04i12.004.v04i12.004.

[19] Morchid, A., El Alami, R., Raezah, A. A. and Sabbar, Y. (2024) "Applications of Internet of Things (IoT) and Sensors Technology to Increase Food Security and Agricultural Sustainability: Benefits and Challenges", *Ain Shams Engineering Journal*, Vol. 15, No. 3, p. 102509. ISSN 2090-4479. DOI 10.1016/j.asej.2023.102509.

[20] Murugamani, C., Shithart, S., Hemalatha, S., Ksirsagar, P. R., Riyazuddin, K., Naveed, Q. N., Islam, S., Mazher Ali, S. P. and Batu, A. (2022) "Machine Learning Technique for Precision Agriculture Applications in 5G-Based Internet of Things", *Wireless Communications and Mobile Computing*, Vol. 2022, No. 1, p. 6534238. ISSN 1530-8677. DOI 10.1155/2022/6534238.

[21] Navarro, E., Costa, N. and Pereira, A. (2020) "A Systematic Review of IoT Solutions for Smart Farming", *Sensors (Basel, Switzerland)*, Vol. 20, No. 15, p. 4231. ISSN 1424-8220. DOI 10.3390/s20154231.

[22] Quy, V. K., Hau, N. V., Anh, D. V., Quy, N. M., Ban, N. T., Lanza, S., Randazzo, G. and Muzirafuti, A. (2022) "IoT-Enabled Smart Agriculture: Architecture, Applications, and Challenges", *Applied Sciences*, Vol. 12, No. 7, p. 3396. ISSN 2076-3417. DOI 10.3390/app12073396.

[23] Razaque, A., Hariri, S., Alajlan, A. M. and Yoo, J. (2025) "A comprehensive review of cybersecurity vulnerabilities, threats, and solutions for the Internet of Things at the network-cum-application layer", *Computer Science Review*, Vol. 58, p. 100789. ISSN 1574-0137. DOI 10.1016/j.cosrev.2025.100789.

[24] Ribeiro Junior, F. M., Bianchi, R. A. C., Prati, R. C., Koleihmainen, K., Soinen, J.-P. and Kamienski, C. A. (2022) "Data reduction based on machine learning algorithms for fog computing in IoT smart agriculture", *Biosystems Engineering*, Vol. 223, pp. 142-158. ISSN 1537-5110. DOI 10.1016/j.biosystemseng.2021.12.021.

[25] Saad Alotaibi, B., Sherma, A. I., Ibrahim, A. U., Abuhussain, M. A., Abdulmalik, H., Dodo, Y. A. and Atakara, C. (2024) "Assimilation of 3D printing, Artificial Intelligence (AI) and Internet of Things (IoT) for the construction of eco-friendly intelligent homes: An explorative review", *Heliyon*, Vol. 10, No. 17, p. e36846. ISSN 2405-8440. DOI 10.1016/j.heliyon.2024.e36846.

[26] Sarohe, S., Harit, S. and Kumar, M. (2025) "A systematic and comprehensive survey of load balancing techniques in Software Defined Network based Internet of Things", *Computer Networks*, Vol. 269, p. 111412. ISSN 1389-1286. DOI 10.1016/j.comnet.2025.111412.

[27] Shalimov, A. (2023) "*IoT in Agriculture: 9 Technology Use Cases for Smart Farming (and Challenges to Consider)*", Eastern Peak - Technology Consulting & Development Company. [Online]. Available: https://easternpeak.com/blog/iot-in-agriculture-technology-use-cases-for-smart-farming-and-challenges-to-consider/ [Accessed: Feb. 6, 2025].

[28] Staff, C. (2025) "*Qu'est-ce que l'apprentissage automatique? Définition, types et exemples*", Coursera. [Online]. Available: https://www.coursera.org/fr-FR/articles/what-is-machine-learning [Accessed: 23 Sept. 23, 2025]. (In French).

[29] Subeesh, A. and Chauhan, N. (2025) "Deep learning based abiotic crop stress assessment for precision agriculture: A comprehensive review", Journal of Environmental Management, Vol. 381, p. 125158. ISSN 0301-4797. DOI 10.1016/j.jenvman.2025.125158.

[30] Venkatesh, V., Raj, P. and Anushia Devi, R. (2022) "Chapter Ten - A smart framework through the Internet of Things and machine learning for precision agriculture", In Raj, P., Saini, K. and Surianarayanan, C. (eds.) "*Advances in Computers*", Elsevier (Edge/Fog Computing Paradigm: The Concept Platforms and Applications), pp. 279-306. ISSN 0065-2458. DOI 10.1016/bs.adcom.2022.02.007.

[31] Wang, J., Liu, Z., Yang, X. Li, M. and Lyu, Z. (2025) "The Internet of Things under Federated Learning: A Review of the Latest Advances and Applications", *Computers, Materials & Continua*, Vol. 82, No. 1, pp. 1-39. ISSN 1546-2218. DOI 10.32604/cmc.2024.058926.

[32] Wang, X., Tang, S. H., Mohd Ariffin, M K. A. B, Ismail, M. I. S. B. and Zhao, R. (2025) "LeafMamba: A novel IoT-integrated network for accurate and efficient plant leaf disease detection", *Alexandria Engineering Journal*, Vol. 123, pp. 415-424. ISSN 1110-0168. DOI 10.1016/j.aej.2025.03.033.

[33] Zhang, E., Liu, Y., Tan, Q., Chen, J. and Pan, T. (2025) "Projected changes in future wheat yield across China under CMIP6 climate simulations and agricultural management scenarios", *Journal of Cleaner Production*, Vol. 508, p. 145570. ISSN 0959-6526. DOI 10.1016/j.jclepro.2025.145570.

[34] Zhou, X. and Zhou, J. (2020) "Data-Driven Driving State Control for Unmanned Agricultural Logistics Vehicle", *IEEE Access*, Vol. 8, pp. 65530-65543. ISSN 2169-3536. DOI 10.1109/ACCESS.2020.2983424.

Volume XVIII Number 3, 2026

Protein Consumption and Socioeconomic Determinants in Indonesia: An Empirical Analysis of the Lowest-Income Households

Mohammad Nafi Izzuddin[1], David Forgenie[2], Lia Rohmatul Maula[1,4], Sukardi Sukardi[3], Nikmatul Khoiriyah[1,4]

[1] Department of Agribusiness, Faculty of Agriculture, Universitas Islam Malang, East Java, Indonesia

[2] Department of Agricultural Economics and Extension, Faculty of Food and Agriculture, The University of the West Indies, St Augustine Campus, Trinidad and Tobago

[3] Public Administration Department, Faculty of Political and Social Science, University of Merdeka Malang, Indonesia

[4] Centre for Food Security Studies, Faculty of Agriculture, Universitas Islam Malang, Indonesia

Abstract

Adequate protein consumption is essential for improving nutritional outcomes in developing countries, yet empirical evidence on the determinants of protein consumption among the poorest households in Indonesia remains scarce. This study investigates the socioeconomic determinants of protein consumption among the lowest-income households using nationally representative data from the 2023 Survei Sosial Ekonomi Nasional (Susenas), focusing on 67,918 households in the lowest income quintile (Q1). Protein-rich foods were classified into ten categories: eight animal-sourced foods (fish, seafood, beef, mutton, poultry, eggs, milk, and other meats) and two plant-based proteins (tofu and tempeh). A binary probit model was applied to examine the influence of household income, household size, and the relative prices of protein-rich foods on the probability of consumption. Marginal effects were calculated to interpret the change in probability associated with a one-unit change in each explanatory variable. The results reveal contrasting patterns across protein sources. Beef consumption is strongly influenced by all three socioeconomic factors — income, household size, and price — reflecting its status as a luxury protein largely inaccessible to the poorest households. In contrast, eggs emerge as a stable and affordable protein source consumed by nearly 79% of Q1 households, remaining largely insensitive to own-price changes and unaffected by household size. Household income significantly increases the probability of consuming nearly all protein categories, while price sensitivity drives poor households to rely more heavily on tofu and tempeh as affordable alternatives. This study provides novel empirical evidence on the probability-based determinants of protein consumption among Q1 households, offering practical insights for food subsidies, local production support, and nutrition education programs in Indonesia.

Keywords

Protein consumption, low-income households, socioeconomic determinants, household nutrition, food security, Indonesia.

Izzudin, M. N., Maula, L. R., Forgenie, D., Sukardi, S. and Khoiriyah, N. (2026) "Protein Consumption and Socioeconomic Determinants in Indonesia: An Empirical Analysis of the Lowest-Income Households", *AGRIS on-line Papers in Economics and Informatics*, Vol. 18, No. 3, pp. 17-29. ISSN 1804-1930. DOI 10.7160/aol.2026.180302.

Introduction

Adequate protein consumption is essential for improving nutritional outcomes and reducing inequality in developing countries (Akter et al., 2025; Bayati et al., 2025; Singha et al., 2025). However, in Indonesia protein intake remains uneven, with low-income households particularly vulnerable to deficiency due to limited access and affordability (Al Hasan et al., 2022; Khusun et al., 2022). This issue aligns directly with Sustainable Development Goal 2, which targets the elimination of hunger and the improvement of nutrition worldwide (Sporchia et al., 2024). Understanding the socio-economic determinants of protein consumption among the poorest households is therefore critical for designing effective nutritional policies.

Indonesia is characterized by one of the lowest levels of protein consumption in the ASEAN region. As illustrated in Figure 1, at 67 grams per capita, average protein intake falls well below Malaysia (159 grams), Thailand (141 grams), and the Philippines (93 grams), despite Indonesia being the fourth most populous country in the world. This low intake reflects nutritional imbalances associated with stunting, malnutrition, and developmental disorders in children (Budzulak et al., 2022). The Government of Indonesia, through Ministry of Health Regulation No. 75 of 2013, officially sets the daily minimum protein requirement at 57 grams per capita, yet many low-income households remain unable to meet this threshold. Consumption of animal-sourced protein such as fish, meat, eggs, and milk is limited by high prices, while plant-based sources such as tofu and tempeh — though affordable — are often consumed without sufficient dietary diversification (Forgenie, Khoiriyah, Mahase-Forgenie, et al., 2023). This imbalance reflects both economic and educational barriers to achieving nutritional adequacy in poor households.

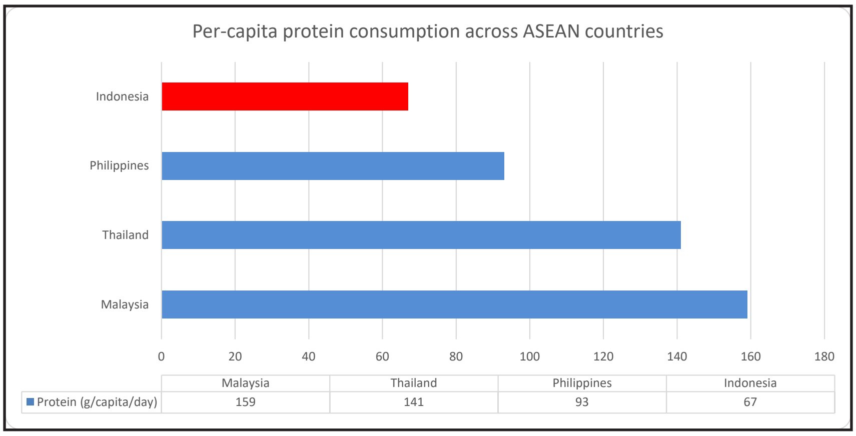
The lowest income quintile (Q1) represents the 20% of the population with the lowest income and the highest vulnerability to poverty and nutritional deficiencies (Khoiriyah et al., 2020). Although previous studies have analyzed food consumption in Indonesia using probit and related approaches (Cele and Mudhara, 2024; Kolog et al., 2023), empirical evidence specifically focused on protein consumption patterns among Q1 households remains scarce. Existing work tends to treat the population in aggregate or compare across income groups without isolating the nutritional behavior of the poorest segment. Furthermore, the joint effects of protein prices, household income, and household size on the probability of consuming specific protein sources have not been systematically examined for this group. Addressing this gap is essential for designing subsidy, production, and nutrition education policies that directly target Indonesia's most economically disadvantaged households.

This study therefore investigates the socio-economic determinants of protein consumption among the lowest-income households in Indonesia. Specifically, it analyzes how the prices of protein-rich foods, household income, and household size influence the probability of consuming ten categories of protein sources — eight animal-sourced foods (fish, seafood, beef, mutton, poultry, eggs, milk, and other meats) and two plant-based proteins (tofu and tempeh). By providing new empirical evidence on Q1 households, the study contributes to the understanding of nutritional inequality in developing countries and offers practical insights for food and nutrition policy in Indonesia.

Material and methods

Specification of the demand function: Probit model

This research employs a binary probit regression model to analyze the probability of households consuming particular varieties of protein-rich foods. The dependent variable is binary, indicating whether a household consumes a specific protein source ($Y_i = 1$) or does not ($Y_i = 0$). The model incorporates several independent variables, including the prices of protein sources, household size, and household income. This model is appropriate for the analysis



Source: FAO Food Balance Sheets, adapted from Al Hasan et al. (2022)

Figure 1: Per-capita protein consumption across ASEAN countries.

of cross-sectional data with a dichotomous dependent variable (Nandiroh et al., 2024; Wen et al., 2025).

The probit model was selected over the logit model based on both theoretical and empirical considerations. Theoretically, the probit model assumes that the latent propensity of households to consume protein-rich foods follows a normal distribution, which is consistent with the distributional properties of socio-economic variables such as income and household size observed in large nationally representative surveys (Greene, 2012). Empirically, the probit specification has been widely adopted in food consumption studies in Indonesia and comparable developing-country contexts, providing a basis for methodological consistency and comparability with prior work. Furthermore, given the large sample size of 67,918 households, the choice between probit and logit has negligible practical impact on the estimated marginal effects, as both models converge asymptotically with sufficient observations.

The probit model is a statistical probability model based on the cumulative normal probability distribution (Allam, 2026). The binary dependent variable Y_i takes on values of 0 and 1, where 1 indicates that a household consumes the protein source and 0 indicates otherwise (Jose et al., 2020). The probability of household consumption can be expressed as follows (Greene, 2012; Jahan and Kamruzzaman, 2022) (Equation 1):

$$P_i = prob\,[Y_i = 1|X] = \int_{-\infty}^{x_i'\beta} (2\pi)^{-1/2} \exp\left(-\frac{t^2}{2}\right) dt = \Phi(x_i'\beta) \quad (1)$$

Where Φ represents the cumulative distribution function of a standard normal random variable, x'_i is a vector of explanatory variables (price of each protein source, household size, and household income), and β is a vector of parameters to be estimated. The unobserved random variable ε_i is assumed to be independently and identically distributed according to a standard normal distribution, that is $\varepsilon_i \sim N\,(0,1)$.

Since probit coefficients reflect effects on the latent index rather than direct changes in the observed probabilities, marginal effects must be calculated for meaningful interpretation (Busenbark et al., 2021). The marginal effect for continuous explanatory variables X_k on the probability $P\ (Y_i = 1 \mid X)$, holding all other variables constant at their sample mean values, is derived as (Equation 2):

$$\frac{\delta y_i}{\delta x_{ik}} = \phi(x_i'\beta)\beta_k \quad (2)$$

Where ϕ represents the probability density function of the standard normal distribution. For dummy or discrete variables, the marginal effect is instead estimated as the discrete change in predicted probability (Equation 3):

$$\Delta = \Phi(x\beta, d = 1) - (\bar{y}\beta, d = 0) \quad (3)$$

Marginal effects provide insights into how explanatory variables shift the probability of protein consumption among low-income households (Jahan and Kamruzzaman, 2022). In this study, marginal effects were calculated for each variable while holding all other variables constant at their sample mean values. Overall model significance was evaluated using the Wald chi-square test, while the significance of individual coefficients was assessed using z-statistics and their corresponding p-values. The pseudo R-square (McFadden) was used as a measure of goodness of fit for each of the ten protein consumption models.

Data and data source

This study uses secondary data sourced from the results of the National Socio-economic Survey (Susenas) in 2023, conducted by the Central Bureau of Statistic (BPS) Indonesia. The data was chosen because of its wide availability, national representation, and comprehensive coverage of information related to household socioeconomic conditions. This survey covers various aspects, such as demographic conditions, education, health, and consumption patterns, and records household income data. The data used in this study are household consumption and expenditure data in Indonesia. The process begins with the division of the region into strata that reflect geographical diversity, such as provinces, cities or districts, sub-districts, and villages, as well as considering urban and rural characteristics. The sample frame is obtained from the last Population Census, and within each stratum, census blocks are randomly selected using the Probability Proportional to Size method, which determines the probability of selecting census blocks based on the number of households. Furthermore, within each selected census block, sample households were randomly selected. In collecting the data, BPS used the multistage random sampling technique, which is sampling conducted in several stages to ensure that the selected samples represent various social and economic characteristics in Indonesia. In addition, BPS also applied stratified sampling to ensure that the sample included socioeconomic variations, such as differences between urban and rural areas or household income levels.

The data in this study are of the cross-section type. In simple terms, the concept of cross-sectional data is data that has many objects in the same year or data collected at one time on many objects. This study examined ten sources of protein, consisting of eight animal-sourced foods (fish, seafood, beef, mutton, poultry, eggs, milk, and other meats) and two plant-based proteins (tofu and tempeh). The data used are quantity consumption and expenditure data in Indonesia. Unit prices were calculated by dividing expenditure by the quantity consumed, as the Susenas dataset does not provide the actual prices paid by households. Moreover, (Nsabimana, 2024; Vu, 2020) said common practice has been to calculate unit values dividing expenditures by corresponding quantities and use them as a direct substitute for market prices. To examine how price and income influence protein consumption, household income data is divided into five quintiles: Q1 (lowest 20%), Q2 (20%-<40%), Q3 (40%<60%), Q4 (60%-<80%), and Q5 (highest 20%). Special attention is directed toward Q1 households, which represent the poorest segment of the population. A total of 67,918 households fall into this lowest income quintile. By analyzing this group, the study investigates the probability that low-income households consume different types of protein foods, offering valuable insights into food accessibility, consumption patterns, and nutritional vulnerability among Indonesia's most economically disadvantaged households. In this study, the classification of protein sources follows the groupings used in Susenas to make the analysis more standardized and in line with official data. This grouping is important for understanding variations in household protein consumption and distinguishing between animal and plant-based protein sources. The protein category in Susenas is presented in Table 1. This classification was used as the basis for analyzing protein consumption in the study.

Groups	Protein source-food group	Foods
1	Fish	Yellow tail, cob, tuna, cakalang dencis, mackerel, selar, mackerel, lema or tatare, banyar or banyara, wet anchovies, milkfish, cork, mujaer, goldfish, tilapia, catfish, snapper, catfish, other wet fresh fish, pomfret, gourami, baronang, preserved buffalo or peda, preserved mackerel, preserved cob, tuna or skipjack, preserved anchovies, preserved selar, preserved sepat, preserved milkfish, preserved cork, canned fish, canned tuna, etc.), other preserved fish, preserved shrimp (ebi, rebon), squid, cuttlefish, preserved, preserved shrimp and other aquatic animals, and cooked fish
2	Seafood	Shrimp, lobster, squid, cuttlefish, octopus, clams, octopus, snails, snails, mussels, preserved shrimp, and other freshwater animals.
3	Beef	Beef
4	Mutton	Mutton, lamb, and sheep
5	Poultry	Purebred chicken meat and free-range chicken meat
6	Eggs	Purebred chicken eggs, free-range chicken eggs, duck or manila duck eggs
7	Milk	Factory liquid milk, sweetened condensed milk, powdered milk, baby milk powder, other milk, and other milk products
8	Tofu	Tofu
9	Tempeh	Tempeh
10	Other meat	Other fresh meat, other cured meat, other (liver, offal, ribs, legs, oxtail, head, etc.), cutlets, and brisket

Source: Own summary based on data from SUSENAS 2023

Table 1: Protein source-food groups in Indonesia.

Results and discussions

Testing probit models

The results from ten probit models analyzing Indonesian household consumption of various protein sources are presented in Table 2. These models cover the consumption of fish, seafood, beef, mutton, poultry, milk, eggs, tofu, tempeh, and other meats. All models demonstrate strong statistical validity, with highly significant Wald chi² test results ($p < 0.01$ for all models), indicating that the explanatory variables—household size, income, and prices—simultaneously contribute to predicting consumption decisions. The models also showed varying levels of explanatory power, as reflected in the pseudo-R^2 values. Beef (0.6344), mutton (0.5998), and seafood (0.4685) consumption models exhibited the highest pseudo-R^2, suggesting these protein choices are more strongly influenced by the selected household and economic factors. In contrast, relatively low pseudo R^2 values for eggs (0.0355), tofu (0.118), and tempeh (0.1384) suggest that variables beyond income, price,

Protein model	Wald Chi^2	p-value (Wald Chi^2)	Log-likelihood	Pseudo R^2
Fish	26560.70	0.000	-29186.568	0.3127
Seafood	5591.26	0.000	-3171.325	0.4685
Beef	894.24	0.000	-257.665	0.6344
Mutton	85.36	0.000	-28.479	0.5998
Poultry	13986.89	0.000	-18604.282	0.2732
Eggs	7182.42	0.000	-22108.077	0.1397
Milk	2899.72	0.000	-39343.898	0.0355
Tofu	10941.57	0.000	-40894.896	0.1181
Tempeh	12838.39	0.000	-39975.361	0.1384
Other meat	3441.89	0.000	-5327.182	0.2442

Source: Own summary based on data from SUSENAS

Table 2: Probit model analysis of protein consumption.

and household size significantly influence their consumption. Indeed, cultural preferences strongly shape consumption as Indonesians perceive tempeh as part of their cultural heritage—a widely valued, nutritional staple consumed across generations (Anwar et al., 2023).

Log-likelihood values across the models also support these trends, with more negative values for frequently consumed items such as tofu (-40,894.90) and tempeh (-39,975.36), reflecting larger sample sizes and variability. This study includes all protein sources commonly consumed by households in Indonesia, excluding pork due to its limited consumption among the population (Hanifasari et al., 2024).

Protein consumption and socioeconomic determinants in Indonesia

The probit models investigate how protein prices, household size, and household income influence the probability of protein consumption (Anindita et al., 2022; Dogbe et al., 2024; Seo and Hwang, 2022). Table 3 presents the estimated probit coefficients, while Table 4 reports the corresponding marginal effects, which represent the change in the probability of consumption associated with a one-unit change in each explanatory variable, holding all others constant at their sample means. It is important to emphasize that probit coefficients and marginal effects reflect changes in the probability of consumption, not changes in the quantity consumed (Budzulak et al., 2022; McCabe et al., 2022; Silva et al., 2024). As shown in Tables 3 and 4, the prices of both animal-sourced and plant-based proteins significantly affect the probability of protein consumption in Indonesia.

Animal-sourced foods are a key group of nutrient-dense foods that provide highly bioavailable micronutrients such as iron, zinc, vitamin B12, vitamin A, vitamin D3, iodine, calcium, folic acid, and essential fatty acids (Scott et al., 2023). The price of food is one of the most important factors influencing what households choose to eat (Downs et al., 2022). Furthermore, Dana et al., (2021) note that prices are among the most important factors determining consumers' food choice decisions. When the price of a certain protein source rises, families may seek cheaper alternatives or adjust their diets by combining different foods to meet nutritional needs. This means that price changes do not only affect the consumption of one product, but can also influence the demand for other protein sources, depending on whether they act as substitutes or complements in the household menu. Animal-sourced foods like fish, meat, and poultry often compete with each other, while plant-based options such as tofu and tempeh can either replace or accompany these animal proteins.

Fish is the most widely consumed protein source in Indonesia, especially for low-income households, due to its relatively affordable price and high availability. The probit results confirm that fish price has a significant negative effect on the probability of fish consumption ($\beta = -2.122$, $p < 0.01$, Table 3), with a corresponding marginal effect of -0.849 (Table 4). This indicates that, holding other variables constant, a one-unit increase in fish price reduces the probability of fish consumption by approximately 0.849 percentage points. Substitution patterns are evident, as poultry and eggs act as alternatives when fish prices rise. In contrast, tofu and tempeh show negative coefficients, suggesting their role as complementary proteins commonly consumed alongside fish.

Independent variable	Dependent variable									
	Fish	Seafood	Beef	Mutton	Poultry	Eggs	Milk	Tofu	Tempeh	Other Meat
α_i (Constant)	11.833	-8.463	9.544	5.841	-1.761	-1.889	-8.780	1.555	0.109	-12.652
	(0.000)	**(0.000)**	**(0.000)**	**(0.443)**	**(0.000)**	**(0.000)**	**(0.000)**	**(0.000)**	**(0.780)**	**(0.000)**
Fish Price	-2.122	0.530	0.435	0.413	-0.089	-0.058	0.183	-0.510	-0.545	-0.312
	(0.000)	**(0.000)**	**(0.003)**	**(0.203)**	**(0.000)**	**(0.002)**	**(0.000)**	**(0.000)**	**(0.000)**	**(0.000)**
Seafood Price	0.018	-0.724	0.011	0.036	0.032	0.033	0.031	0.047	0.025	-0.017
	(0.051)	**(0.000)**	**(0.861)**	**(0.682)**	**(0.002)**	**(0.000)**	**(0.005)**	**(0.000)**	**(0.002)**	**(0.319)**
Beef Price	0.020	-0.009	-0.186	0.087	0.008	0.006	0.016	-0.010	-0.004	0.037
	(0.000)	**(0.396)**	**(0.000)**	**(0.090)**	**(0.244)**	**(0.203)**	**(0.075)**	**(0.063)**	**(0.441)**	**(0.110)**
Mutton Price	-0.144	0.130	0.080	-0.211	0.038	0.015	0.041	0.022	0.036	0.091
	(0.000)	**(0.000)**	**(0.011)**	**(0.000)**	**(0.000)**	**(0.003)**	**(0.000)**	**(0.000)**	**(0.000)**	**(0.000)**
Poultry Price	0.061	0.034	0.037	0.054	-0.261	-0.021	0.019	-0.053	-0.065	0.011
	(0.000)	**(0.000)**	**(0.095)**	**(0.337)**	**(0.000)**	**(0.000)**	**(0.000)**	**(0.000)**	**(0.000)**	**(0.148)**
Eggs Price	-0.421	-0.703	0.403	-0.390	-1.508	-0.015	0.221	-2.439	-3.061	-1.376
	(0.000)	**(0.000)**	**(0.055)**	**(0.800)**	**(0.000)**	**(0.712)**	**(0.000)**	**(0.000)**	**(0.000)**	**(0.000)**
Milk Price	0.031	0.008	-0.003	0.016	0.001	-0.006	-0.134	-0.002	-0.007	-0.023
	(0.000)	**(0.158)**	**(0.888)**	**(0.679)**	**(0.623)**	**(0.001)**	**(0.000)**	**(0.462)**	**(0.001)**	**(0.000)**
Tofu Price	0.002	-0.010	-0.001	-0.071	-0.003	-0.008	-0.001	-0.049	-0.022	-0.022
	(0.464)	**(0.130)**	**(0.939)**	**(0.373)**	**(0.249)**	**(0.000)**	**(0.685)**	**(0.000)**	**(0.000)**	**(0.000)**
Tempeh Price	0.023	0.009	0.034	-0.063	-0.015	-0.020	-0.009	-0.020	-0.044	-0.012
	(0.000)	**(0.150)**	**(0.015)**	**(0.385)**	**(0.000)**	**(0.000)**	**(0.001)**	**(0.000)**	**(0.000)**	**(0.016)**
Other Meat Price	0.016	-0.016	0.040	0.036	0.004	-0.002	0.007	-0.024	-0.033	-0.297
	(0.007)	**(0.353)**	**(0.025)**	**(0.505)**	**(0.603)**	**(0.759)**	**(0.255)**	**(0.000)**	**(0.000)**	**(0.000)**
Households Size	-0.425	-0.066	-0.442	-0.563	-0.166	0.009	0.362	-0.134	-0.095	0.197
	(0.000)	**(0.023)**	**(0.000)**	**(0.073)**	**(0.000)**	**(0.390)**	**(0.000)**	**(0.000)**	**(0.000)**	**(0.000)**
Income	0.007	0.003	0.004	0.005	0.006	0.003	0.002	0.003	0.003	0.003
	(0.000)	**(0.000)**	**(0.000)**	**(0.074)**	**(0.000)**	**(0.000)**	**(0.000)**	**(0.000)**	**(0.000)**	**(0.000)**

Source: Own calculation based on data from SUSENAS

Table 3: Probit model analysis of protein consumption.

These findings reaffirm the central role of fish in the diet of low-income households and highlight the sensitivity of its consumption probability to price fluctuations. Seafood, while nutritionally important, remains less accessible to the lowest income quintile because of its relatively high price compared to fish. Seafood price significantly affects the probability of seafood consumption ($\beta = -0.724$, $p < 0.01$, Table 3), with a marginal effect of -0.290 (Table 4), indicating that an increase in seafood price substantially reduces the probability that a household consumes seafood. The substitution pattern suggests that fish, beef, poultry, and tempeh replace seafood when prices rise, while tofu, eggs, and milk act as complements. These findings suggest that the probability of seafood consumption is highly responsive to price in low-income groups, and households readily shift to cheaper proteins when prices increase. Consequently, seafood is often consumed only occasionally, rather than as a regular source of protein, in these households.

Beef is considered one of the most nutrient-dense animal proteins but is largely inaccessible to households in the lowest quintile due to its high price. Furthermore, Angeles-Agdeppa et al. (2021) note that moderate and severely food-insecure families have significantly lower mean consumption of meat. The regression analysis indicates a significant negative relationship between beef price and the probability of beef consumption ($\beta = -0.186$, $p < 0.01$, Table 3), with a marginal effect of -0.074 (Table 4), confirming its sensitivity to price fluctuations. When beef prices increase, low-income households tend to substitute it with fish, mutton, poultry, eggs, milk, and other meats, as reflected by their positive coefficients. This result is in line with research from Khoiriyah, Forgenie, et al., (2024), who reported that the price of beef leads to a relatively larger decrease

Independent Variable	Fish	Seafood	Beef	Mutton	Poultry	Eggs	Milk	Tofu	Tempeh	Other Meat
A. Price Variable										
Fish Price	-0.8488***	0.2120***	0.1740***	0.1652	-0.0356***	-0.0232***	0.0732***	-0.2040***	-0.2180***	-0.1248***
Seafood Price	0.0072*	-0.2896***	0.0044	0.0144	0.0128***	0.0132***	0.0124***	0.0188***	0.0100***	-0.0068
Beef Price	0.0080***	-0.0036	-0.0744***	0.0348*	0.0032	0.0024	0.0064*	-0.0040*	-0.0016	0.0148
Mutton Price	-0.0576***	0.0520***	0.0320**	-0.0844***	0.0152***	0.0060***	0.0164***	0.0088***	0.0144***	0.0364***
Poultry Price	0.0244***	0.0136***	0.0148*	0.0216	-0.1044***	-0.0084***	0.0076***	-0.0212***	-0.0260***	0.0044
Eggs Price	-0.1684***	-0.2812***	0.1612*	-0.1560	-0.6032***	-0.0060	0.0884***	-0.9756***	-1.2244***	-0.5504***
Milk Price	0.0124***	0.0032	-0.0012	0.0064	0.0004	-0.0024***	-0.0536***	-0.0008	-0.0028***	-0.0092***
Tofu Price	0.0008	-0.0040	-0.0004	-0.0284	-0.0012	-0.0032***	-0.0004	-0.0196***	-0.0088***	-0.0088***
Tempeh Price	0.0092***	0.0036	0.0136**	-0.0252	-0.0060***	-0.0080***	-0.0036***	-0.0080***	-0.0176***	-0.0048**
Other Meat Price	0.0064***	-0.0064	0.0160**	0.0144	0.0016	-0.0008	0.0028	-0.0096***	-0.0132***	-0.1188***
B. Household Characteristics										
Household Size	-0.1700***	-0.0264**	-0.1768***	-0.2252*	-0.0664***	0.0036	0.1448***	-0.0536***	-0.0380***	0.0788***
Income	0.0028***	0.0012***	0.0016***	0.0020*	0.0024***	0.0012***	0.0008***	0.0012***	0.0012***	0.0012***

Note: Statistical significance: *** = $p < 0.01$; ** = $p < 0.05$; * = $p < 0.10$.
Marginal effects were approximated at the mean using the standard normal density approximation ($\varphi \approx 0.4$), following (Greene, 2012).
Marginal effects represent the change in probability of consumption associated with a one-unit change in each explanatory variable, holding all other variables constant at their sample means.

Source: Own calculation based on data from SUSENAS 2023

Table 4: Marginal effects of protein consumption determinants among lowest-income households in Indonesia.

in the quantity demanded, as consumers are more responsive to changes in the price of beef compared to other food items. Conversely, tofu and tempeh show negative coefficients, suggesting their complementary role alongside beef in household diets. These results illustrate that beef remains a luxury protein, consumed more frequently by higher-income groups, while low-income households prefer more affordable protein sources when beef prices rise.

Mutton and goat meat are among the least consumed protein sources in Indonesia, with average consumption below 0.001 kg per week and household expenditure of only IDR 40 per week in the lowest quintile. The regression results show that mutton price has a significant negative effect on the probability of mutton consumption (β = -0.211, $p < 0.01$, Table 3), with a marginal effect of -0.084 (Table 4). Affordability appears to be the main barrier, as households tend to prefer beef despite its slightly higher price, largely due to perceptions of greater nutritional value. Substitution patterns also appear, as households replace mutton with fish, poultry, milk, and other meats when prices rise, while eggs, tofu, and tempeh act as complements. These findings underscore the position of mutton as a luxury food consumed primarily during religious or cultural events, rather than as part of daily diets in low-income households.

Poultry remains an important source of animal protein for Indonesian households, with rural consumption averaging 0.106 kg per week and expenditures of IDR 4,745. The analysis shows that the prices of fish, seafood, mutton, poultry itself, eggs, and tempeh significantly affect the probability of poultry consumption. Notably, the negative coefficient for poultry price (β = -0.261, $p < 0.01$, Table 3), with a marginal effect of -0.104 (Table 4), indicates that price increases reduce the probability of poultry consumption. This prompts households — particularly those in the lowest income quintile — to shift toward substitute proteins such as fish, seafood, and milk. Conversely, tofu and tempeh function as complements, often consumed alongside poultry.

In contrast, eggs represent a more central and stable protein source for low-income households. Field data indicate an average expenditure per household of IDR 3,998 per week, equivalent to approximately 2-3 eggs per household, underscoring their affordability and accessibility. In Ethiopia, eggs were consumed one to two times per week by 20% and one to two times per month by 19% of households (Daba et al., 2021). The regression results further reveal that the probability of egg consumption is not significantly affected by its own price ($p = 0.712$, Table 3), confirming that eggs are price-inelastic and categorized as a normal good (Khoiriyah, Forgenie, et al., 2024). Daba et al., (2021) also noted that availability, positive attitudes toward quality, relative affordability, the limited effort needed to prepare and cook, storage, and favorable beliefs about health benefits were found to be facilitators of egg consumption. Furthermore, Forgenie, Khoiriyah and Elbaar (2023) also reported that eggs were the least responsive to changes in prices, since they are a relatively cheap source of protein compared to other sources of animal protein. Nevertheless, eggs exhibit a substitutive role when the prices of seafood, beef, and mutton rise, and act as complements to fish, poultry, milk, tofu, and tempeh. This finding suggests eggs are a substitute for higher-priced proteins like mutton, while also being consumed alongside staples such as fish and poultry. Taken together, these findings highlight eggs as a reliable and flexible protein choice for the poorest households, less vulnerable to price volatility compared to other animal protein sources.

For low-income households in Indonesia, milk consumption is largely in the form of baby formula, powdered milk, and sweetened condensed milk, rather than fresh milk. The regression results show that the prices of fish, seafood, mutton, eggs, and tempeh significantly affect the probability of milk consumption ($p < 0.05$), while beef and other meats do not ($p > 0.05$). Importantly, milk price itself has a significant negative effect on the probability of milk consumption ($\beta = -0.134$, $p < 0.01$, Table 3), with a marginal effect of -0.054 (Table 4), meaning that higher milk prices reduce the probability of household milk consumption. This confirms earlier findings by Headey, (2023) and Muhammad, (2017), who reported that rising milk prices lead to reduced intake, especially in economically vulnerable groups. Moreover, the positive coefficients for fish, seafood, mutton, poultry, and other meats suggest that milk functions as a substitute protein source when the prices of these items increase. By contrast, tofu and tempeh act as complements, reflecting their role in mixed dietary patterns where plant- and animal-based proteins are consumed together. These results emphasize that milk, despite its relatively high price compared to eggs and tempeh, remains an important alternative protein choice, particularly when the affordability of other animal protein sources is constrained.

Other meats — such as liver, offal, ribs, feet, tail, head, meat slices, and brisket — remain relatively popular protein sources in Indonesia, especially among low-income households. These cuts are cheaper than prime beef or mutton, yet still provide important nutritional value and are often chosen as affordable alternatives. Khoiriyah et al. (2024) also emphasize that other meats play a key role for households with limited purchasing power because they are more accessible compared to premium animal proteins. The analysis shows that the prices of fish, mutton, eggs, milk, tofu, and tempeh significantly affect the probability of other meat consumption ($p < 0.05$). By contrast, beef, poultry, and seafood do not have significant effects on the probability of other meat consumption ($p > 0.05$). The price coefficient for other meats is significantly negative ($\beta = -0.297$, $p < 0.01$, Table 3), with a marginal effect of -0.119 (Table 4), indicating that when their prices rise, the probability of consuming other meats decreases. Furthermore, fish, seafood, eggs, milk, tofu, and tempeh also have negative coefficients, suggesting they function as complements to other meats in household diets. In contrast, beef, mutton, and poultry show positive coefficients, indicating their role as substitutes. Thus, when the prices of these prime proteins increase, low-income households are more likely to shift toward consuming other meats as a more affordable option.

Household size plays an important role in shaping protein consumption patterns. While parents may have control over how and where they feed their children, the fundamental issue for low-income households is poverty, food prices, and access to food outlets (Ravikumar et al., 2022). Larger households naturally require greater expenditure on food, as also emphasized by Rashid et al., (2024), who found that food expenditure increases significantly with household size in developing country contexts. In this study, data from 67,918 households in the lowest income quintile show that the average household size is 2-3 members. The analysis confirms that household size significantly affects the probability of consuming fish, seafood, beef, poultry, milk, tofu, tempeh, and other meats. For example, larger household size significantly reduces the probability of fish consumption ($\beta = -0.425$, $p < 0.01$, Table 3; marginal effect = -0.170, Table 4) and beef consumption ($\beta = -0.442$, $p < 0.01$, Table 3; marginal effect = -0.177, Table 4). By contrast, household size does not significantly influence the probability of mutton or egg consumption. The insignificant result for mutton is likely due to its high price and very low consumption rate, with only 0.1% of the sample reporting consumption. Conversely, the insignificant effect on eggs may stem from their affordability and widespread accessibility, making them the most consumed protein source, with 78.9% of the 67,918 households reporting egg consumption. This suggests that while household size matters for most protein sources, it has little influence on the extremes: very expensive items like mutton and very cheap staples like eggs.

Income is another central determinant of protein consumption. Income directly influences food purchasing power and allocation. Increase in income stimulates household expenditures, which are one of the most important components directly influenced by the income level of individuals. Lévay et al. (2021) says richer household spend higher shares from richer households spend a higher portion than poorer households. Additionally, Olufemi-Phillips et al. (2024) state that income levels directly influence purchasing power and the ability of households to access food. Households in the lowest income quintile were found to have an average monthly income of IDR 1,603,544. The analysis shows that income has a significant positive effect on the probability of consumption for nearly all protein groups, with significance values of $p < 0.05$ for fish, seafood, beef, poultry, eggs, milk, tofu, tempeh, and other meats. Although the marginal effects of income are numerically small (for example, a marginal effect of 0.003 for fish), they are statistically significant, indicating that even modest income gains meaningfully raise the probability that poor households consume protein-rich foods. The only exception is mutton ($p = 0.074$, Table 3), reflecting its low consumption level among poor households. When income increases, households are more inclined to purchase beef rather than mutton, which explains the lack of statistical significance. These findings align with the theoretical view that higher incomes reduce the share of food expenditures while increasing the capacity for dietary diversification (Mehraban & Ickowitz, 2021; Rashid et al., 2024b). In practice, however, protein choices among the poorest households remain highly sensitive to income changes, except in the case of very low-demand items like mutton.

Policy implications

The findings of this study carry important implications for food and nutrition policy in Indonesia's lowest-income households. First, improving household purchasing power through targeted food subsidies, conditional cash transfers, and price stabilization is essential, as income significantly raises the probability of consuming nearly all ten protein categories. Second, larger households experience markedly lower probabilities of consuming beef (marginal effect = -0.177) and fish (-0.170), highlighting the need for family-size-calibrated nutrition assistance. Third, the high price sensitivity for fish (marginal effect = -0.849) and beef calls for stronger supply chains, local production support, and maintained accessibility of plant-based proteins such as tofu and tempeh as critical substitutes. Fourth, eggs — consumed by 78.9% of Q1 households and statistically insensitive to own-price — represent an ideal vehicle for nutritional interventions such as school feeding and maternal nutrition programs. Finally, policy targeting must explicitly prioritize the lowest-income quintile through means-tested subsidies and integration with existing social protection schemes such as Family Hope Program. Together, these strategies can reduce nutritional inequality, prevent malnutrition, and strengthen public health outcomes for Indonesia's most vulnerable populations.

Conclusion

This study provides novel empirical evidence on the socioeconomic determinants of protein consumption among Indonesia's lowest-income households. Using a binary probit model applied to 67,918 Q1 households in the 2023 Susenas, the analysis confirms that household income significantly raises the probability of consuming nearly all protein categories, while household size reduces it for most animal-sourced foods. Price sensitivity varies substantially across proteins, revealing two contrasting patterns: beef consumption is strongly constrained by all three socioeconomic factors, positioning it as a luxury protein inaccessible to the poor, whereas eggs emerge as a stable and affordable staple consumed by nearly 79% of Q1 households and largely insensitive to price and household size. These findings underscore the persistence of nutritional inequality and the central role of tofu and tempeh as affordable protein substitutes. Effective policy responses must therefore combine targeted food subsidies, supply chain and local production support for animal-sourced proteins, nutrition education promoting dietary diversification, and explicit Q1 targeting through existing social protection programs. Together, these strategies can strengthen food security, reduce malnutrition, and improve public health outcomes in Indonesia and comparable developing countries.

Acknowledgements

The authors would like to express their sincere gratitude to the Indonesian Central Bureau of Statistics (Badan Pusat Statistik, BPS) for their excellent service and timely support in providing access to the Susenas data used in this study. Appreciation is also extended to all members of the research team for their valuable contributions, dedication, and collaboration throughout the completion of this work.

Corresponding author:
Nikmatul Khoiriyah
Department of Agribusiness, Faculty of Agriculture, Universitas Islam Malang, Indonesia
Centre for Food Security Studies, Faculty of Agriculture, Universitas Islam Malang, Indonesia
E-mail: nikmatul@unisma.ac.id

References

[1] Akter, S., Kaya, F., Sumon, S. A., Hassan, M. M. and Das, M. K. (2025) "Exploring the impact of income, aquaculture production, protein supply, and food production on life expectancy", *Green Technologies and Sustainability*, Vol. 3, No. 4. ISSN 2949-7361. DOI 10.1016/j.grets.2025.100237.

[2] Al Hasan, S. M., Saulam, J., Mikami, F., Kanda, K., Ngatu, N. R., Yokoi, H. and Hirao, T. (2022) "Trends in per Capita Food and Protein Availability at the National Level of the Southeast Asian Countries: An Analysis of the FAO's Food Balance Sheet Data from 1961 to 2018", *Nutrients*, Vol. 14, No. 3, pp. 1-14. ISSN 2072-6643. DOI 10.3390/nu14030603.

[3] Allam, I. M. A. (2026) "A Python Implementation of Probit Regression Without High-Level Statistical Libraries", *Research Gate*. ISSN 7615-0077. DOI 10.13140/RG.2.2.11870.91207.

[4] Angeles-Agdeppa, I., Toledo, M. B. and Zamora, J. A. T. (2021) "Moderate and severe level of food insecurity is associated with high calorie-dense food consumption of Filipino households", *Journal of Nutrition and Metabolism*, p. 5513409, 15 p., ISSN 2090-0732. DOI 10.1155/2021/5513409.

[5] Anindita, R., Khoiriyah, N. and Sa'Diyah, A. A. (2022) "Food consumption pattern Far Away from Home as a source of household food protein in Indonesia", *IOP Conference Series: Earth and Environmental Science*, *2nd International Conference on Environmental Ecology and Food Security* 19.-20.2.2022 Mataram, Indonesia, Vol. 1107. ISSN 1755-1307. DOI 10.1088/1755-1315/1107/1/012118.

[6] Anwar, K., Maskar, D. H., Prasetyo, I. N. and Kusumawati, I. (2023) "Tempeh as a Cultural Heritage in Indonesia: Intergenerational Perception", *Jurnal Gizi Dan Pangan*, Vol. 18, Supp.1, pp. 75-77. ISSN 1978-1059. DOI 10.25182/jgp.2023.18.supp.1.75-77.

[7] Bayati, M., Arkia, E. and Emadi, M. (2025) "Socio-economic inequality in the nutritional deficiencies among the world countries: Evidence from global burden of disease study 2019", *Journal of Health, Population and Nutrition*, Vol. 44, No. 1, 8 p. ISSN 2072-1315. DOI 10.1186/s41043-025-00739-z.

[8] Budzulak, J., Majewska, K. and Kędzia, A. (2022) "Malnutrition as the cause of growth retardation among children in developed countries", *Annals of Agricultural and Environmental Medicine*, Vol. 29, No. 3, pp. 336-341. ISSN 1898-2263. DOI 10.26444/aaem/148010.

[9] Busenbark, J. R., Graffin, S. D., Campbell, R. J. and Lee, E. Y. (2021) "A Marginal Effects Approach to Interpreting Main Effects and Moderation", *SAGE*, pp. 1-23. ISSN 1552-7425. DOI 10.1177/1094428120976838.

[10] Cele, T. and Mudhara, M. (2024) "Impacts of Crop Production and Value Chains on Household Food Insecurity in Kwazulu-Natal: An Ordered Probit Analysis", *Sustainability,* Vol. 16, No. 2. ISSN 2071-1050. DOI 10.3390/su1602070.

[11] Daba, A. K., Murimi, M., Abegaz, K. and Hailu, D. (2021) "Determinants and constraints to household-level animal source food consumption in rural communities of Ethiopia", *Journal of Nutritional Science*, Vol. 10, pp. 1-10, p. 700. ISSN 2048-6790. DOI 10.1017/jns.2021.52.

[12] Dana, L. M., Chapman, K., Dixon, H., Miller, C., Neal, B., Kelly, B., Ball, K. and Pettigrew, S. (2021) "The relative importance of primary food choice factors among different consumer groups: A latent profile analysis", *Food Quality and Preference*, Vol. 94, Sept. 2019, p. 104199. ISSN 0950-3293. DOI 10.1016/j.foodqual.2021.104199.

[13] Dogbe, W., Wang, Y. and Giha, C. R. (2024) "Nutritional implications of substituting plant - based proteins for meat: Evidence from home scan data", *Agricultural and Food Economics*, Vol. 8. No. 30. ISSN 2193-7532. DOI 10.1186/s40100-024-00324-8.

[14] Downs, S. M., Fox, E. L., Zivkovic, A., Mavros, T., Sabbahi, M., Merchant, E. V., Mutuku, V., Okumu-Camerra, K. and Kimenju, S. (2022) "Drivers of food choice among women living in informal settlements in Nairobi, Kenya", *Appetite*, Vol. 168, Oct. 2021, p. 105748. ISSN 0195-6663. DOI 10.1016/j.appet.2021.105748.

[15] Forgenie, D., Khoiriyah, N., Mahase-Forgenie, M. and Adeleye, B. N. (2023) "An error-corrected linear approximate almost ideal demand system model for imported meats and seafood in Indonesia", *Heliyon*, Vol. 9, No. 11, p. e21390. ISSN 2405-8440. DOI 10.1016/j.heliyon.2023.e21390.

[16] Forgenie, D., Khoiriyah, N. and Elbaar, E. F. (2023) "Estimating household price and income elasticities for animal derived sources of food using the QUAIDS model: The case of Jakarta, Indonesia", *Tropical Agriculture*, Vol. 100, No. 4., pp. 317-328. ISSN 1804-1930.

[17] Greene, W. (2012) "*Econometric analysis*". 7th ed., Prentice Hall. ISBN 978-0-13-139538-1. DOI 10.1007/s00362-010-0315-8.

[18] Hanifasari, D., Masudin, I., Zulfikarijah, F., Rumijati, A. and Restuputri, D. P. (2024) "Millennial generation awareness of halal supply chain knowledge toward purchase intention for halal meat products: Empirical evidence in Indonesia", *Journal of Islamic Marketing*, Vol. 15, No. 7, pp. 1847-1885. ISSN 1847-1885. DOI 10.1108/JIMA-01-2023-0012.

[19] Headey, D. (2023) "Can dairy help solve the malnutrition crisis in developing countries? An economic analysis", *Animal Frontiers*, Vol. 13, No. 1, pp. 7-16. ISSN 2160-6064. DOI 10.1093/af/vfac083.

[20] Jahan, M. and Kamruzzaman, M. (2022) "Economic Assessment of Commercial Turkey and Broiler Farming in Bangladesh: An Application of Binary Probit Model", *Canadian Journal of Business and Information Studies*, pp. 65-71. ISSN 2663-7812. DOI 10.34104/cjbis.022.065071.

[21] Jose, A., Philip, M., Prasanna, L. T. and Manjula, M. (2020) "Comparison of Probit and Logistic Regression Models in the Analysis of Dichotomous Outcomes", *Current Research in Biostatistics*, Vol. 10, No. 1, pp.1-19. ISSN 2524-2210. DOI 10.3844/amjbsp.2020.1.19.

[22] Khoiriyah, N., Anindita, R., Hanani, N. and Muhaimin, A. W. (2020) "Animal food demand in Indonesia: A quadratic almost ideal demand system approach", *Agris on-line Papers in Economics and Informatics*, Vol. 12, No. 2, pp. 85-97. ISSN 1804-1930. DOI 10.7160/aol.2020.120208.

[23] Khoiriyah, N., Apriliawan, H., Maula, L. R., Sa'diyah, A. A., Forgenie, D., Susyanti, J. and Nendissa, D. R. (2024) "Examining Indonesian protein consumption patterns and factors: A probit model", *BIO Web of Conferences*, Vol. 143, p. 01003. ISSN 2117-4458. DOI 10.1051/bioconf/202414301003.

[24] Khoiriyah, N., Forgenie, D., Sookhai, S. and Joko Saputro, A. (2024) "Demand Elasticities of Animal-Sourced Food: Empirical Study in Yogyakarta, Indonesia", IJFAEC (International Journal of Food and Agricultural Economics), Vol. 12, No. 3. ISSN 2149-3766. DOI 10.22004/ag.econ.344845.

[25] Khusun, H., Februhartanty, J., Anggraini, R., Mognard, E., Alem, Y., Noor, M. I., Karim, N., Laporte, C., Poulain, J. P., Monsivais, P. and Drewnowski, A. (2022) "Animal and Plant Protein Food Sources in Indonesia Differ Across Socio-Demographic Groups: Socio-Cultural Research in Protein Transition in Indonesia and Malaysia", *Frontiers in Nutrition*, Vol. 9, No. February, pp. 1-8. ISSN 2296-861X. DOI 10.3389/fnut.2022.762459.

[26] Kolog, J. D., Asem, F. E. and Mensah-bonsu, A. (2023) "The state of food security and its determinants in Ghana: An ordered probit analysis of the household hunger scale and household food insecurity access scale", *Scientific African Journal*, Vol. 19, p. e01579. ISSN 2468-2276. DOI 10.1016/j.sciaf.2023.e01579.

[27] Lévay, P. Z., Vanhille, J., Goedemé, T. and Verbist, G. (2021) "The association between the carbon footprint and the socio-economic characteristics of Belgian households", *Ecological Economics*, Vol. 186, p. 107065. ISSN 0921-8009. DOI 10.1016/j.ecolecon.2021.107065.

[28] McCabe, C. J., Halvorson, M. A., King, K. M., Cao, X. and Kim, D. S. (2022) "Interpreting Interaction Effects in Generalized Linear Models of Nonlinear Probabilities and Counts", *Multivariate Behavioral Research*, Vol. 57, No. 2-3, pp. 243-263. ISSN 1532-7906. DOI 10.1080/00273171.2020.1868966.

[29] Mehraban, N. and Ickowitz, A. (2021) "Dietary diversity of rural Indonesian households declines over time with agricultural production diversity even as incomes rise", *Global Food Security*, Vol. 28, p. 100502. ISSN 2211-9124. DOI 10.1016/j.gfs.2021.100502.

[30] Muhammad, A. (2017) "The Influence of Income and Prices on Global Dietary Patterns by Country, Age, and Gender", United States Department of Agriculture, Report ID: ERR-225. DOI 10.22004/ag.econ.256711.

[31] Nandiroh, U., Khoiriyah, N., Forgenie, D. and Arifatus, A. (2024) "Exploring income-based disparities in plant-based protein consumption: Evidence from indonesia using probit analysis", *BIO Web of Conferences*, Vol. 143, p. 01008, pp. 1-14. ISSN 2117-4458. DOI 10.1051/bioconf/202414301008.

[32] Nsabimana, A. (2024) "Analysis of household demand patterns using household data: Re-thinking the use of unit values or community prices", WIDER Working Paper, Vol. 2024, UNU-WIDER. ISBN 978-92-9267-480-9. DOI 10.35188/UNU-WIDER/2024/480-9.

[33] Olufemi-Phillips, A. Q., Igwe, A. N., Ofodile, O. C., Eyo-Udo, N. L. and Toromade, A. S. (2024) "Analyzing economic inflation’s impact on food security and accessibility through econometric modeling", *GSC Advanced Research and Reviews*, Vol. 21, No. 2, pp. 102-128. ISSN 2582-4597. DOI 10.30574/gscarr.2024.21.2.0411.

[34] Rashid, F. N., Sesabo, J. K., Lihawa, R. M. and Mkuna, E. (2024a) "Determinants of household food expenditure in Tanzania: Implications on food security", *Agriculture & Food Security*, Vol. 13, No. 1, 13 p. ISSN 2048-7010. DOI 10.1186/s40066-023-00462-0.

[35] Rashid, F. N., Sesabo, J. K., Lihawa, R. M. and Mkuna, E. (2024b) "Determinants of household food expenditure in Tanzania: Implications on food security", *Agriculture and Food Security*, Vol. 13, No. 1, pp. 1-16. ISSN 2048-7010. DOI 10.1186/s40066-023-00462-0.

[36] Ravikumar, D., Spyreli, E., Woodside, J., McKinley, M. and Kelly, C. (2022) "Parental perceptions of the food environment and their influence on food decisions among low-income families: A rapid review of qualitative evidence", *BMC Public Health*, Vol. 22, No. 1, pp. 1-16. ISSN 1471-2458. DOI 10.1186/s12889-021-12414-z.

[37] Scott, J. M., Belton, B., Mahrt, K., Thilsted, S. H. and Bogard, J. R. (2023) "Food systems transformation, animal-source foods consumption, inequality, and nutrition in Myanmar", *Food Security*, Vol. 15, No. 5, pp. 1345-1364. ISSN 1876-4525. DOI 10.1007/s12571-023-01380-6.

[38] Seo, H. and Hwang, J. (2022) "Analysis of Decisive Elements in the Purchase of Alternative Foods Using Bivariate Probit Model", *Sustainability (Switzerland)*, Vol. 14, No. 7, pp. 1-12. ISSN 2071-1050. DOI 10.3390/su14073822.

[39] Silva, A., Rivera, M., Durán-Agüero, S. and Sactic, M. I. (2024) "Single-Person Households: Insights from a Household Survey of Fruit and Vegetable Purchases", *Nutrients*, Vol. 16, No. 17, p. 2851. ISSN 2072-6643. DOI 10.3390/nu16172851.

[40] Singha, S., Alabed, N., Singha, R. and Jaheer Mukthar, K. P. (2025) "Pathways for Sustainable Food Security: Status, Challenges, and Opportunities in Developing World, In: "*The Paradigm Shift from a Linear Economy to a Smart Circular Economy: The Role of Artificial Intelligence-Enabled Systems, Solutions and Legislations*", pp. 185-201. ISBN 978-3-031-87550-2. DOI 10.1007/978-3-031-87550-2_14.

[41] Sporchia, F., Antonelli, M., Aguilar-martínez, A., Bach-faig, A., Caro, D., Davis, K. F., Sonnino, R. and Galli, A. (2024) "Zero hunger: Future challenges and the way forward towards the achievement of sustainable development goal 2", *Sustainable Earth Reviews*, Vol. 7, No. 10. ISSN 2520-8748. DOI 10.1186/s42055-024-00078-7.

[42] Vu, L. H. (2020) "Estimation and Analysis of Food Demand Patterns in Vietnam", *Economies*, Vol. 8, No. 1, p. 11. ISSN 2227-7099. DOI 10.3390/economies8010011.

[43] Wen, H., Huang, K., Ma, Z. and Huang, J. (2025) "Determinants of driver injury severity in electric vehicle crashes: A random parameters binary probit model with heterogeneity in means and variances", *Journal of Transportation Safety & Security*, Vol. 17, No. 4, pp. 427-455. ISSN 1943-9970. DOI 10.1080/19439962.2024.2415960.

Agris on-line Papers in Economics and Informatics

Volume XVIII Number 3, 2026

Digital Maturity for Sustainable Livestock Process Management

Oksana Makovoz[1, 2], Sergii Lysenko[2]

[1] IHI Zittau, Dresden University of Technology, Zittau, Germany

[2] National Technical University “Kharkiv Polytechnic Institute”, Kharkiv, Ukraine

Abstract

Digital transformation in livestock enterprises creates managerial value only when digital tools are embedded in business processes and translated into measurable sustainability effects. The paper develops a digital maturity and portfolio prioritization approach for sustainable livestock process management. The research uses a structurally balanced pilot sample of 20 Ukrainian livestock en-terprises, documentary evidence, expert-documentary scoring, KPI baseline comparison and multi-criteria normalization. The results show uneven maturity across strategy, process automation, data analytics, digital architecture and cyber-resilience. The highest priority is assigned to BI and data warehouse initiatives, followed by ERP-based feed logistics integration and operational control solutions.

Keywords

Business process management, digital maturity, digital transformation, key performance indicators, livestock enterprises, sustainability.

Makovoz, O. and Lysenko, S. (2026) "Digital Maturity for Sustainable Livestock Process Management", *AGRIS on-line Papers in Economics and Informatics*, Vol. 18, No. 3, pp. 31-40. ISSN 1804-1930. DOI 10.7160/aol.2026.180303.

Introduction

Digital transformation is changing the way agricultural and rural enterprises plan resources, control production cycles, collect operational data and communicate managerial decisions. In livestock en-terprises this change is especially sensitive because biological cycles, feed logistics, veterinary mea-sures, productivity accounting, biosecurity procedures and product traceability generate a conti-nuous flow of operational information. If this information remains fragmented across paper journals, spreadsheets and isolated applications, digital technologies do not become a source of susta-inable competitiveness; they only reproduce existing process weaknesses in another technical form.

The problem investigated in this paper concerns the managerial assessment of digital transformation in livestock business process management. The main difficulty is not only whether an enterprise has digital tools, but whether these tools change process controllability, data reliability, decision speed and sustainability-oriented outcomes. This problem is important because livestock enterprises ope-rate under biological, logistical, veterinary and market constraints; therefore, digital transformation that is not connected with process management may increase technical complexity without improving managerial results.

The scientific discussion on digital transformation has moved from technology adoption to organizational capability and value creation. Digital transformation is interpreted as an organizational change process produced by digital technologies and strategic responses (Vial, 2019). It affects business models, organizational structures and strategic adaptation (Verhoef et al., 2021). In the context of business process management, technologies such as IoT, analytics and automation requi-re a process-oriented integration logic rather than isolated technological implementation (Butt, 2020). Process maturity is also important because stable improvement depends on the transition from fragmented routines to managed and measured processes (Van Looy, 2014), while business processes in digital transformation require explicit measurement strategies (Debnath et al., 2022).

A separate research stream links digital maturity with sustainability. The potential of digital transformation can be assessed through structured maturity dimensions (Corejova and Chinoracky, 2021). At the same time, maturity models need explicit sustainable performance indicators to avoid a narrow technological interpretation of progress (Fortier et al., 2025). Empirical evidence from agrifood supply chains confirms that digital maturity is related to sustainable performance (Barbosa et al., 2025). For Ukraine, ICT-driven advancement is important for implementing sustainable development values (Lysenko and Makovoz, 2025), while digitalization, risk management and artifi-cial intelligence are increasingly relevant for improving agricultural business processes (Zelisko et al., 2024).

Despite these contributions, two gaps remain important. First, many studies assess digital transfor-mation at the level of technologies, general organizational readiness or supply-chain performance, while livestock enterprises require a process-level diagnostic instrument that connects maturity domains with feed logistics, veterinary control, productivity accounting and traceability. Second, maturity assessment often ends with a diagnostic score and does not show how managers should select a limited portfolio of digital initiatives under budget, time and risk constraints. This is problematic because digital initiatives in livestock enterprises must be staged and coordinated with production cycles, data quality, staff readiness and cyber-resilience.

The scientific problem of the paper is formulated as follows: how can digital maturity diagnostics be connected with KPI-based managerial effects, sustainability outcomes and an operational algorithm for prioritizing digital initiatives in livestock business process management. The working hypothesis is that digital maturity becomes operationally useful when maturity domains, process bottlenecks, KPI baselines, sustainability links and risk-adjusted portfolio logic are treated as one dia-gnostic and decision-support system.

The purpose of the paper is to conduct a diagnostic assessment of digital maturity in livestock business process management and to suggest a portfolio prioritization algorithm that connects matu-rity gaps, KPI effects, sustainability orientation, costs, implementation duration and risks. The object of the research is business process management in livestock enterprises. The subject of the research is the set of digital tools, data governance routines, maturity domains and managerial mechanisms that influence process controllability, transparency, resilience and sustainable performance.

The benefit of the research for science is the transformation of a general BPM-digital transformati-on-sustainability idea into a livestock-specific maturity and prioritization mechanism. The benefit for business is an operational sequence for choosing digital initiatives when financial and organizational resources are limited. The benefit for studies is a replicable case that demonstrates how digital transformation can be connected with measurable process effects and sustainable rural development.

Literature survey

The literature survey is organized around five research streams: BPM and process maturity, digital transformation capability, digital maturity diagnostics, agri-food digitalization and sustainability, and multi-criteria decision-making. The review is not limited to listing previous studies. Its purpose is to identify where the existing literature remains insufficient for livestock enterprises: many studies explain the importance of digital transformation, but fewer studies show how to convert maturity diagnostics into process-level priorities, KPI effects and sustainability-oriented managerial decisi-ons.

BPM and process maturity literature provides the first foundation, but it is not sufficient on its own. Process maturity supports the transition from fragmented routines to managed and optimized processes (Van Looy, 2014). This logic is especially important for digital transformation because di-gital tools should be embedded into process redesign rather than simply attached to existing workflows (Butt, 2020). Business processes in digital transformation also require explicit measurement strategies that make it possible to evaluate the effects of transformation rather than only describe technological adoption (Debnath et al., 2022). However, this stream usually remains sector-neutral and does not explain how maturity should be interpreted for livestock-specific process chains such as feed logistics, veterinary control, productivity accounting and traceability.

Digital transformation literature clarifies why technology adoption alone cannot be treated as transformation. Digital transformation should be understood as organizational change triggered by digital technologies (Vial, 2019). Value creation depends on strategic adaptation, organizational structure and business model change (Verhoef et al., 2021). In addition, organizational performance is linked not only to digital technology itself, but also to digital capability and digital innovation, which mediate the transformation of technology into business value (Khin and Ho, 2019). These studies are conceptually strong, but they remain too broad for enterprise-level process diagnostics. They justify why digital transformation matters, but they do not provide a sufficient operational mechanism for ranking concrete initiatives under cost, time and risk constraints in rural livestock enterprises.

Digital maturity models offer diagnostic structure, but they often stop at scoring. Multi-attribute maturity assessment can be used to diagnose the digital maturity of SMEs (Kljajic Borstnar and Pucihar, 2021), while composite approaches help assess the potential of digital transformation (Corejova and Chinoracky, 2021). Organizational readiness is also important before AI-based transformation because digital transformation requires not only technology but also appropriate structures, competences and implementation conditions (Aldoseri et al., 2024). These resources help assess specific areas, but they do not provide a comprehensive answer to the management que-stion that arises after the assessment: which digital initiatives should be implemented first, and why? The weakness is especially relevant for livestock enterprises because limited budgets require a defensible portfolio logic rather than a general digitalization checklist.

Agri-food and sustainability-oriented studies bring the problem closer to the sector of this article, but a gap still remains. Fuzzy AHP can be used to structure criteria for digital transformation in agribusiness, which supports the idea that transformation priorities must be weighed (Martins et al., 2023). Digital maturity is also related to sustainable performance in agri-food supply chains (Barbosa et al., 2025). Digital technologies and risk management are relevant for agricultural supply cha-ins because they influence coordination, resilience and decision-making (Lumbantobing et al., 2025). The relationship between digital transformation and sustainability in agribusiness is increa-singly treated as a separate research agenda (Annosi et al., 2024). Digitalization, artificial intelligen-ce and risk considerations are also important for improving agricultural business processes (Zelisko et al., 2024). Nevertheless, these studies mainly address agribusiness, supply chains or general ag-ricultural processes, whereas livestock enterprises need a more specific link between maturity do-mains, operational bottlenecks, KPI changes and sustainability effects.

The MCDM literature is useful because digital transformation decisions are multi-criteria by nature. AHP provides a methodological basis for deriving criteria weights (Saaty, 1980), while TOPSIS, VIKOR and related methods support the comparison of alternatives with conflicting criteria (Opricovic and Tzeng, 2004; Zavadskas et al., 2014). However, MCDM models are often applied after alternatives have already been formulated. The missing element is the bridge between maturity dia-gnosis and the construction of alternatives: the enterprise must first identify which bottlenecks are strategically important, then translate them into initiatives, and only then rank them.

The literature survey, therefore, reveals a clear research gap. Existing studies either focus on broad digital transformation theory, general maturity scoring, agribusiness digitalization, sustainability performance or MCDM ranking. The decision-making framework for digital transformation in livestock farming remains underdeveloped; it should combine: areas of maturity -> process bottle-necks -> impact on key performance indicators (KPIs) -> interpretation from a sustainability per-spective -> prioritised digital initiatives. This gap explains the importance of the present study. It transforms digital maturity from a descriptive score into a decision-support system for sustainable livestock business process management.

According to the results of the literature survey presented in Table 1, it is clear that contemporary

Research stream	Key contribution to the paper	Critical limitation addressed in this paper
BPM and process maturity	Explains ownership, standardization, measurement and continuous improvement	Insufficient adaptation to livestock process chains and operational bottlenecks
Digital transformation capability	Shows that technology creates value through organizational and strategic change	Too broad for enterprise-level ranking of concrete initiatives
Digital maturity diagnostics	Provides domain-based scoring and readiness assessment	Often stops at diagnosis and does not guide initiative selection
Agri-food digitalization and sustainability	Links digitalization with sustainability, risk and supply-chain performance	Weak process-level explanation for livestock BPM and KPI effects
MCDM	Supports weighted comparison of alternatives with conflicting criteria	Needs to be connected with maturity gaps before alternatives are ranked

Source: Compiled by the authors

Table 1: Literature streams and their role in the proposed approach.

theories are positioned not merely as an indicator of new technology adoption, but as a mechanism to support DT decision-making. It connects process maturity with digital transformation priorities and sustainability-oriented outcomes in a way that is comprehensible to managers of livestock enterprises.

Materials and methods

The research period covers February 2025 - April 2026. The unit of observation was the enterprise-level digital maturity profile rather than an individual employee. This distinction is important becau-se the paper does not present a representative sociological survey of employees; it presents a pilot expert-documentary diagnostic assessment of enterprise-level process management and digitalizati-on evidence.

The empirical basis includes 20 livestock-related enterprises and organizations. The sample was purposive and structurally balanced: 10 entities represented the state sector and 10 represented pri-vate business structures. The selection logic combined the availability of sufficient information, regional diversity and relevance to livestock-related process management. The balanced structure allows comparison between state-sector and private-sector profiles, but the sample should not be interpreted as statistically representative of all Ukrainian livestock enterprises.

The evidence base was grouped into four blocks. The first block covered process context: process maps, roles, regulations, AS-IS and TO-BE descriptions and internal documents. The second block covered digital tools and architecture: ERP, CRM, BI dashboards, digital journals, cloud services, data integration and system interoperability. The third block covered KPI performance: reporting timeliness, data quality, transparency, automation and coordination indicators. The fourth block covered managerial effects and risks: decision speed, process visibility, quality of control, cyber-resilience and continuity risks.

The expert-documentary assessment used a five-domain maturity model: D1 - strategy and digital change governance; D2 - processes and automation; D3 - data, analytics and transparency; D4 - digital architecture and integration; D5 - cybersecurity, compliance and resilience. Each domain was assessed on a 0-4 scale, where 0 means absence of a formal routine, 1 means episodic use, 2 means partial formalization, 3 means systematic use and 4 means integration, measurement and regular managerial review.

The calculation procedure included three steps (formulas 1-10).

First, the score of each indicator was normalized:

$$r_{ij} = \frac{x_{ij}}{4}, \quad x_{ij} \in [0; 4], \quad r_{ij} \in [0; 1], \tag{1}$$

where x_{ij} is the expert-documentary score of indicator i in domain j.

Second, the domain score was calculated as follows:

$$S_j = \sum_{i=1}^{n_j} w_{ij} r_{ij}, \quad \sum_{i=1}^{n_j} w_{ij} = 1, \tag{2}$$

where w_{ij} is the weight of indicator i within domain j.

Third, the integral digital maturity index was calculated as follows:

$$I = \sum_{j=1}^{m} W_j S_j, \quad \sum_{j=1}^{m} W_j = 1, \tag{3}$$

where I is the integral digital maturity index, m is the number of maturity domains, and W_j is the weight of domain j.

In the pilot version, equal domain weights were used:

$$W_j = \frac{1}{m}. \tag{4}$$

Since the model includes five domains, the equal weight of each domain is:

$$W_j = \frac{1}{5} = 0.20. \tag{5}$$

If it is necessary to present the integral maturity index on the original 0–4 scale, the normalized index can be transformed as follows:

$$I_{0-4} = 4I. \tag{6}$$

The prioritization of digital initiatives used six criteria: expected maturity gain, KPI effect, transparency and controllability, implementation duration, cost and risk. Benefit criteria were normalized so that a larger value increased the utility of an initiative.

$$N_b(z_k) = \frac{z_k - \min(z)}{\max(z) - \min(z)}. \tag{7}$$

Cost, duration and risk criteria were normalized in the opposite direction:

$$N_c(z_k) = \frac{\max(z) - z_k}{\max(z) - \min(z)}. \tag{8}$$

The general utility of initiative k was expressed as follows:

$$U_k = \alpha N_b(\Delta I_k) + \beta N_b(\Delta KPI_k) + \\ + \gamma N_b(T_k) + \rho N_c(L_k) + \\ + \delta N_c(C_k) + \varepsilon N_c(R_k), \quad (9)$$

where U_k is the total utility score of initiative k;

ΔI_k is the expected maturity gain;

ΔKPI_k is the expected KPI effect;

T_k is the transparency and controllability score;

L_k is implementation duration;

C_k is implementation cost;

R_k is risk level;

N_b is benefit normalization;

N_c is cost normalization.

The weights of the prioritization criteria satisfy the following condition:

$$\alpha + \beta + \gamma + \rho + \delta + \varepsilon = 1. \quad (10)$$

The higher the value of U_k, the higher the priority of the corresponding digital initiative.

Sustainability was operationalized through proxy indicators linked with process management. Economic sustainability was represented by reporting timeliness, cost control, losses and resource-use efficiency. Environmental sustainability was represented by feed losses, energy use, traceability, microclimate control and waste-related process discipline. Social sustainability was represented by labor safety, staff digital competence, reduction of routine manual work, veterinary compliance and the reliability of responsibilities in process execution. This proxy design is appropriate for a pilot maturity assessment, while a larger field study should collect direct environmental and social mea-surements.

Data in Table 2 clarify the methodological boundary of the paper. The results should be treated as a pilot diagnostic application that demonstrates the logic of assessment and prioritization. The next stage of research should increase the number of enterprises, collect direct system logs and validate the weight system with a larger expert panel.

Results and discussion

The starting point of the results is a process-oriented interpretation of digital transformation. A digital tool is not assessed by its technical novelty alone. It is assessed by the degree to which it im-proves process execution, data availability, managerial visibility, decision speed and sustainability-related outcomes. For livestock enterprises, the core process chain can be generalized as planning, supply and feed logistics, livestock operations, veterinary and biosecurity control, productivity ac-counting, sales and logistics. Each process generates data and creates decision points; therefore, digital maturity should be measured where process execution and managerial control meet. The diagnostic model consists of five connected domains. D1 captures whether digital transformation is governed as a managed change portfolio. D2 captures whether core livestock processes are described, standardized and automated. D3 captures whether data are reliable, timely

Element	Specification in the paper	Reasoning
Research period	February 2026 -April 2026	Defines the temporal boundary for collecting, harmonizing and verifying the empirical evidence
Observation unit	Enterprise-level digital maturity profile	The study assesses process management systems and enterprise-level evidence, not personal attitudes
Sample	20 livestock-related enterprise profiles: 10 state sector and 10 private-sector	A structurally balanced pilot sample enables group comparison
Data sources	Documents, regulations, reporting forms, digital records, expert-documentary scoring	Combines observable evidence with diagnostic judgement
Assessment method	0-4 maturity scale across five domains	Provides an interpretable maturity profile for enterprises with limited digital data
Decision method	Multi-criteria utility function with benefit and cost-risk normalization	Transforms diagnostic gaps into a ranked portfolio of digital initiatives
Sustainability logic	Economic, environmental and social proxy indicators connected with BPM effects	Links digital maturity with sustainable rural business outcomes

Source: Compiled by the authors

Table 2: Research design and methodological decisions.

and available for managerial analytics. D4 captures whether digital systems are integrated rather than isolated. D5 captures whether digital infrastructure is protected through access control, backup, incident procedures and cyber-resilience. These domains correspond to the full managerial cycle from strategic intent to operational continuity.

The diagnostic results show that the private-sector group had a higher baseline maturity than the state-sector group. This difference can be explained by stronger investment flexibility, faster managerial decisions and a higher probability of using ERP, CRM, BI or integrated accounting tools. At the same time, both groups improved after digital changes, which confirms that maturity diagnostics can support operational prioritization rather than only describe the current state.

The information in Table 3 indicates three patterns. First, all domains improved in both groups. Second, private enterprises remained more digitally mature both before and after the changes. Third, the strongest state-sector improvement related to data and analytics, while the strongest private-sector improvement was connected with processes and automation. This means that digital transformation should be managed as a differentiated domain portfolio rather than as a single technology program.

Subsequent to the analysis of the data in Table 4, the managerial effect is visible not only in technological readiness but also in the quality of management. The state-sector group gained 16-18 per-centage points across indicators, whereas the private-sector group gained 15-17 percentage points. This means that even when initial maturity is different, a process-oriented digitalization programme can improve reporting, data quality, transparency and routine coordination.

The diagnostic profile indicates that the main bottlenecks are not limited to lack of equipment or software. In many livestock enterprises the key restrictions are delayed data collection, duplicate manual entry, weak connection between operational records and managerial dashboards, unclear ownership of data and insufficient cyber-resilience of operational systems. Such restrictions reduce the managerial value of digital tools because managers receive information too late or cannot trust its completeness. For this reason, the proposed solution is not a single technology implementation. It is a set of transformation scenarios that can be combined into a staged roadmap. The scenarios translate maturity gaps into manageable initiatives and prevent the common situation in which an enterprise invests in a complex ERP or IoT solution before basic data and process rules are ready.

Data in Table 5 reveal that sustainability is not an additional slogan placed after digital transformation. It is embedded into the process effects of each scenario. The first scenario improves compliance and reduces errors; the second supports resource and feed-loss control; the third creates the basis for energy and productivity optimization;

Domain	State before	State after	Change	Private before	Private after	Change
D1 Strategy and governance	1.83	2.31	+0.48	2.19	2.80	+0.61
D2 Processes and automation	1.87	2.37	+0.50	2.31	3.01	+0.70
D3 Data and analytics	1.79	2.37	+0.58	2.08	2.65	+0.57
D4 Architecture and integration	1.75	2.20	+0.45	2.18	2.80	+0.62
D5 Cybersecurity and resilience	1.77	2.31	+0.54	2.07	2.67	+0.60
Average	1.80	2.31	+0.51	2.17	2.79	+0.62

Source: Compiled by the authors

Table 3: Average digital maturity scores by domains in the pilot sample, 0-4 scale.

Indicator	Meaning	State before	State after	Change	Private before	Private after	Change
E1	Timeliness of reporting and decision support	45	62	+17	59	75	+16
E2	Data quality and error control	49	66	+17	53	70	+17
E3	Process transparency and KPI coverage	47	63	+16	57	73	+16
E4	Automation and coordination of routine actions	48	66	+18	58	73	+15

Source: Compiled by the authors

Table 4: Managerial effects of digitalization in the pilot sample, percentage points.

Scenario	Main content	Domains	Managerial effect	Sustainability effect
S1 Quick effects	Digital checkpoints, electronic journals, checklists, basic data standardization	D1-D3	Early transparency, fewer manual errors, clearer responsibility	Less waste from data errors, better compliance and staff discipline
S2 Integration and controllability	ERP/accounting integration, reference data, warehouse batches, feed logistics control	D2-D4	Less duplication, shorter cycles, better traceability	Lower feed losses, better resource planning and financial sustainability
S3 Analytics and optimization	BI dashboards, forecasting, microclimate monitoring, resource optimization	D3-D4	Faster decisions, better planning and resource use	Energy efficiency, productivity control and early warning of operational deviations
S4 Cyber-resilience and compliance	Access control, backup, audit, incident management and recovery procedures	D5	Lower downtime and lower risk of data loss	Continuity of rural business infrastructure and protection of sensitive operational data

Source: Compiled by the authors

Table 5: Digital transformation scenarios and sustainability-oriented process effects.

and the fourth protects the continuity of digital in-frastructure. Therefore, the sustainability dimension is operationalized through measurable process consequences rather than only through general environmental declarations.

The prioritization algorithm proposed in the paper connects maturity gaps with KPI effects, transparency, cost, duration and risk. It intentionally remains more transparent than a full AHP-TOPSIS procedure because the first managerial need of livestock enterprises is a reproducible ranking that can be discussed with managers, accountants, veterinarians and IT coordinators. AHP can be used at the next stage to justify weights, while TOPSIS or VIKOR can be used for robustness testing.

Subsequent to the analysis of the data in Table 6, the first implementation priority should be a data warehouse and BI dashboard layer. This initiative has the highest utility because it affects the most critical bottleneck: the absence of reliable and timely managerial data. The ERP contour for feed logistics ranks second because it connects planning, procurement and warehouse accounting. Mobi-le veterinary control and IoT monitoring are important operational initiatives, but their effect be-comes stronger after the data and integration foundation is prepared. The cyber-resilience initiative has a lower direct KPI effect in the table, but it should be treated as a mandatory control embedded into each implementation wave.

Sustainability is a central dimension of the proposed approach, not a secondary effect. In livestock process management, sustainability cannot be reduced to environmental reporting alone. It should be interpreted as the ability of an enterprise to maintain economic viability, reduce resource losses, protect animal and product traceability, support labor safety, increase transparency and preserve the continuity of rural business infrastructure.

Economic sustainability is linked with the ability to plan resources, reduce avoidable losses and make timely decisions. When reporting is delayed or feed movements are not traceable, managers cannot react to deviations in time. Digital maturity in D2, D3 and D4 improves economic sustaina-bility because process automation, data quality and integration reduce duplicate operations and increase the reliability of planning. In the pilot results, the improvement of reporting timeliness and data quality by 16-18 percentage points in the state-sector group and 15-17 percentage points in the private-sector group demonstrates this managerial effect.

Environmental sustainability is linked with feed logistics, energy use, microclimate control, waste reduction and traceability of production. The approach does not claim that environmental outcomes were directly measured in the pilot stage. Instead, it identifies digital process proxies that prepare the basis for direct environmental measurement. For example, ERP-based feed logistics can reduce unexplained feed losses; IoT microclimate monitoring can support more efficient energy use; and BI dashboards can make deviations visible before they become systemic resource losses.

Social sustainability is linked with staff competence, reduction of routine manual work, labor safety, veterinary compliance and accountability of process roles. Digital transformation changes not only the technical system but also the way people work. Therefore, D1 and D2 are

Initiative	Process link	Scenario	Key scores	Utility / rank
Data warehouse and BI dashboards	Productivity accounting and management reporting	S2	Maturity gain 0.12; KPI effect 0.20; transparency 0.85; risk 0.30	0.69 / 1
ERP contour for feed logistics and warehouse batches	Feed planning, procurement, warehouse and write-off	S2	Maturity gain 0.10; KPI effect 0.15; transparency 0.75; risk 0.35	0.34 / 2
Mobile module for veterinary incidents and biosecurity protocols	Veterinary control, vaccination and quarantine measures	S1	Maturity gain 0.06; KPI effect 0.10; transparency 0.65; risk 0.25	0.09 / 3
IoT microclimate monitoring with automated alerts	Animal keeping conditions and operational alerts	S3	Maturity gain 0.08; KPI effect 0.12; transparency 0.60; risk 0.40	0.08 / 4
OT/IoT cyber-resilience: network segmentation and IAM	IT infrastructure and operational systems protection	S4	Maturity gain 0.09; KPI effect 0.08; transparency 0.70; risk 0.45	0.08 / 5

Source: Compiled by the authors

Table 6: Pilot prioritization matrix for digital initiatives.

Maturity domain	Economic sustainability	Environmental sustainability	Social and resilience sustainability
D1 Strategy and governance	Investment prioritization, roadmap discipline, KPI ownership	Strategic attention to resource-saving goals	Change governance, training and responsibility
D2 Processes and automation	Lower process time, fewer manual errors, better cost control	Reduced losses from process deviations	Less routine manual work and clearer execution rules
D3 Data and analytics	Timely reporting and better management decisions	Evidence for resource-use and waste indicators	Transparency and accountability of operational decisions
D4 Architecture and integration	Less duplicated input, better planning across systems	Traceability of feed, batches and operational flows	Stable information flow between units and managers
D5 Cybersecurity and resilience	Lower risk of downtime and financial disruption	Continuity of monitoring and control systems	Protection of sensitive data and operational continuity

Source: Compiled by the authors

Table 7: Sustainability interpretation of maturity domains and process indicators.

important be-cause strategy, governance, training and process ownership determines whether employees understand digital procedures and whether managers use data responsibly. Cyber-resilience in D5 also has a social dimension because data loss or system downtime can disrupt work, veterinary control and reporting obligations.

The sustainability contribution of the paper is the proposal to evaluate digital transformation through a maturity-to-effect chain: maturity domain -> process bottleneck -> digital initiative -> KPI effect -> sustainability proxy. This chain prevents abstract sustainability statements and forces each digital initiative to show how it affects economic, environmental or social performance. In future empirical work, this proxy structure should be expanded with direct measurements such as feed loss percentage, energy consumption per production unit, veterinary incident frequency, number of manual reporting hours, staff digital training coverage and downtime caused by digital incidents. The information in Table 7 (above) shows how the sustainability component can be assessed without turning the review into a general ESG report. Each maturity domain has a specific sustainability interpretation. This structure also explains why cyber-resilience is included in the maturity model: in modern rural infrastructure, sustainability depends not only on environmental indicators but also on reliable and secure digital operations.

Conclusion

The paper confirms that digital maturity in livestock enterprises should be assessed through business process management rather than through technology adoption alone. A five-domain dia-gnostic model was formed: strategy and governance, processes and automation, data and analytics, digital architecture and integration, and cybersecurity and resilience. These domains cover the full management cycle from strategic coordination to operational data use and continuity of digital infrastructure.

The pilot sample of 20 Ukrainian livestock enterprises shows positive but uneven maturity dyna-mics. The average state-sector maturity increased from 1.80 to 2.31 points on the 0-4 scale, or by 0.51 points. The average private-sector maturity increased from 2.17 to 2.79 points, or by 0.62 points. The strongest state-sector improvement was observed in data and analytics, while the stron-gest private-sector improvement was observed in processes and automation. This means that state enterprises need stronger data governance and integration support, whereas private enterprises can move faster toward advanced process automation.

The managerial effects also improved. In the state-sector group, the increase across reporting timeliness, data quality, transparency and automation indicators ranged from 16 to 18 percentage points. In the private-sector group, the increase ranged from 15 to 17 percentage points. These results show that digitalization is most valuable when it reduces delays in reporting, improves the quality of primary records and increases the visibility of process execution.

The prioritization algorithm connects maturity gaps with KPI effects, sustainability proxies, costs, duration and risks. In the pilot ranking, the highest-priority initiative is the creation of a data ware-house and BI dashboards with utility 0.69. ERP-based feed logistics integration ranks second with utility 0.34. Mobile veterinary control, IoT microclimate monitoring and OT/IoT cyber-resilience should be implemented as staged or cross-cutting initiatives because their effectiveness depends on the quality of the data and integration foundation.

The sustainability contribution of the paper is the connection between digital maturity domains and economic, environmental and social process effects. Economic sustainability is supported through cost control, reporting timeliness and resource planning. Environmental sustainability is supported through feed-loss control, traceability, energy-related monitoring and reduction of operational devi-ations. Social and resilience sustainability is supported through staff competence, reduction of rou-tine manual work, veterinary compliance and secure continuity of digital operations.

For livestock enterprise management, the recommended roadmap is to start with data inventory, KPI ownership and BI dashboards; then integrate feed logistics and warehouse accounting; then pilot mobile veterinary and IoT tools; and finally formalize cyber-resilience as a permanent control mechanism. For science, the proposed model can be used for further empirical testing of the relati-onship between digital maturity, BPM quality and sustainability outcomes in rural business. For studies, the model provides a clear case for teaching how digital transformation can relate to mea-surable process effects and managerial decision-making.

Corresponding author:
Sergii Lysenko
National Technical University "Kharkiv Polytechnic Institute"
Kyrpychova St, 2, Kharkiv, Ukraine
E-mail: Serhii.Lysenko@emmb.khpi.edu.ua

References

[1] Aldoseri, A., Al-Khalifa, K. N. and Hamouda, A. M. (2024) "Methodological approach to assessing the current state of organizations for AI-based digital transformation", *Applied System Innovation*, Vol. 7, No. 1, Article 14. ISSN 2571-5577. DOI 10.3390/asi7010014.

[2] Annosi, M. C., Appio, F. P., Brenes, E. R. and Brunetta, F. (2024) "Exploring the nexus of digital transformation and sustainability in agribusiness: Advancing a research agenda", *Technological Forecasting and Social Change*, Vol. 206, p. 123587. ISSN 0040-1625, E-ISSN 1873-5509. DOI 10.1016/j.techfore.2024.123587.

[3] Barbosa, M. W., Salume, P. K., Pinto, M. R. and de Sousa, P. R. (2025) "The effects of digital maturity on the sustainable performance of agri-food supply chains in two emerging economies", *The International Journal of Logistics Management*, Vol. 36, No. 3, pp. 848-873. ISSN 0957-4093, E-ISSN 1758-6550. DOI 10.1108/IJLM-03-2024-0144.

[4] Butt, J. (2020) "A conceptual framework to support digital transformation in manufacturing using an integrated business process management approach", *Designs*, Vol. 4, No. 3, p. 17. ISSN 2411-9660. DOI 10.3390/designs4030017.

[5] Corejova, T. and Chinoracky, R. (2021) "Assessing the potential for digital transformation", *Sustainability*, Vol. 13, No. 19, p. 11040. ISSN 2071-1050. DOI 10.3390/su131911040.

[6] Debnath, N., Salgado, C. H., Peralta, M. G., Riesco, D., Baigorria, L., Montejano, G. and Mazzi, M. (2022) "A measurement strategy for business processes in digital transformation", *EPiC Series in Computing*, Vol. 89, pp. 11-20. ISSN 2398 7340. DOI 10.29007/lcpk.

[7] Fortier, J., Gamache, S. and Fonrouge, C. (2025) "Integrating sustainable performance into the digital maturity models for SMEs in manufacturing", *Applied Sciences*, Vol. 15, No. 7, p. 4041. ISSN 2076-3417. DOI 10.3390/app15074041.

[8] Khin, S. and Ho, T. C. (2019) "Digital technology, digital capability and organizational performance: A mediating role of digital innovation", *International Journal of Innovation Science*, Vol. 11, No. 2, pp. 177-195. ISSN 1757-2223, E-ISSN 1757-2231. DOI 10.1108/IJIS-08-2018-0083.

[9] Kljajic Borstnar, M. and Pucihar, A. (2021) "Multi-attribute assessment of digital maturity of SMEs", *Electronics*, Vol. 10, No. 8, p. 885. ISSN 2079-9292. DOI 10.3390/electronics10080885.

[10] Lumbantobing, R. D. H., Ratnayake, R. M. C., Simatupang, T. M., Okdinawati, L. and Mulyono, N. B. (2025) "Exploring the digital transformation of risk management in agriculture supply chains through system dynamics", *Agriculture & Food Security*, Vol. 14, p. 41. E-ISSN 2048-7010. DOI 10.1186/s40066-025-00560-1.

[11] Lysenko, S. and Makovoz, O. (2025) "ICT-driven advancement as a key element in implementing EU sustainable development values in Ukraine", *European Journal of Sustainable Development*, Vol. 14, No. 2, pp. 677-698. ISSN 2239-5938, E-ISSN 2239-6101. DOI 10.14207/ejsd.2025.v14n2p677.

[12] Martins, F. S., Fornari Junior, J. C. F. B., Mazieri, M. R. and Gaspar, M. A. (2023) "A fuzzy AHP analysis of potential criteria for initiatives in digital transformation for agribusiness", *Revista de Administração Mackenzie*, Vol. 24, No. 1, pp. 1-35. ISSN 1518-6776, E-ISSN 1678-6971. DOI 10.1590/1678-6971/eRAMR230055.en.

[13] Opricovic, S. and Tzeng, G. H. (2004) "Compromise solution by MCDM methods: A comparative analysis of VIKOR and TOPSIS", *European Journal of Operational Research*, Vol. 156, No. 2, pp. 445-455. ISSN 0377-2217, E-ISSN 1872-6860. DOI 10.1016/S0377-2217(03)00020-1.

[14] Saaty, T. L. (1980) "*The analytic hierarchy process: Planning, priority setting, resource allocation*", New York, McGraw-Hill. ISBN 978-0-07-054371-3.

[15] Van Looy, A. (2014) "*Business process maturity: A comparative study on a sample of business process maturity models*", Cham, Springer. ISBN 978-3-319-04201-5. DOI 10.1007/978-3-319-04202-2.

[16] Verhoef, P. C., Broekhuizen, T., Bart, Y., Bhattacharya, A., Dong, J. Q., Fabian, N. and Haenlein, M. (2021) "Digital transformation: A multidisciplinary reflection and research agenda", *Journal of Business Research*, Vol. 122, pp. 889-901. ISSN 0148-2963, E-ISSN 1873-7978. DOI 10.1016/j.jbusres.2019.09.022.

[17] Vial, G. (2019) "Understanding digital transformation: A review and a research agenda", *Journal of Strategic Information Systems*, Vol. 28, No. 2, pp. 118-144. ISSN 0963-8687, E-ISSN 1873-1198. DOI 10.1016/j.jsis.2019.01.003.

[18] Zavadskas, E. K., Turskis, Z. and Kildiene, S. (2014) "State of art surveys of overviews on MCDM/MADM methods", *Technological and Economic Development of Economy*, Vol. 20, No. 1, pp. 165-179. ISSN 2029-4913, E-ISSN 2029-4921. DOI 10.3846/20294913.2014.892037.

[19] Zelisko, N., Raiter, N., Markovych, N., Matskiv, H. and Vasylyna, O. (2024) "Improving business processes in the agricultural sector considering economic security, digitalization, risks, and artificial intelligence", *Ekonomika APK*, Vol. 31, No. 3, pp. 10-21. ISSN 2221-1055, E-ISSN 2413-2322. DOI 10.32317/2221-1055.2024030.10.

Volume XVIII Number 3, 2026

Trends in Household Food Expenditures and Income Elasticity in Slovakia in the Pandemic and Post-pandemic Period

Jozef Palkovič

Institute of Statistics, Operations Research and Mathematics, Faculty of Economics and Management, Slovak University of Agriculture in Nitra, Slovakia

Abstract

This paper examines the evolution of food security in Slovakia during the pandemic and post-pandemic years (2020-2023), with a focus on economic access to food. Using household data from the Statistical Office of the Slovak Republic, the study examines changes in income, the share of food expenditures, and income elasticity across regions and demographic groups. The findings show that while household incomes remained relatively stable during the pandemic, the post-pandemic period was marked by rising food prices that outpaced wage growth. As a result, the share of income spent on food increased from 14% in 2020 to 17% in 2023. Data suggests differences between regions: Bratislava recorded the lowest share (12%), whereas Nitra reached 20%. Vulnerability was greatest among retirees, while rural households demonstrated slightly better resilience, likely due to their higher levels of self-sufficiency. Although income elasticity improved overall, disparities widened in 2023 between urban and rural areas. These results underline the need for targeted policies to address affordability and regional inequalities in order to strengthen food security under conditions of economic stress.

Keywords

Food security, food expenditure, income elasticity, regional disparities, demographic vulnerability.

Palkovič, J. (2026) "Trends in Household Food Expenditures and Income Elasticity in Slovakia in the Pandemic and Post-pandemic Period", *AGRIS on-line Papers in Economics and Informatics*, Vol. 18, No. 3, pp. 41-50. ISSN 1804-1930. DOI 10.7160/aol.2026.180304.

Introduction

Food security has become an increasingly important topic, particularly in recent years, amid the pandemic and the ongoing war in Ukraine. The years 2020-2023 offer a unique context for examining this issue, as they encompass two distinct phases: the COVID-19 pandemic and the subsequent economic recovery. It is also a period of escalating war conflict in Ukraine. During the pandemic years, household income patterns remained relatively stable due to government interventions. Uncertainty and restrictions still substantially affected consumer behaviour. On the other hand, the post-pandemic period was marked by inflationary pressures, with faster increases in food prices than in wages, leading to a rising share of household expenditures allocated to food. This period offers a unique opportunity to investigate the impact of income dynamics, regional disparities, and demographic factors on food security. Results could help to identify vulnerable groups and evaluate the application of traditional economic principles in conditions of price instability.

Food security is defined by the Food and Agriculture Organisation (FAO) as a situation in which all people, at all times, have physical, social, and economic access to sufficient, safe, and nutritious food that meets their dietary needs and preferences for an active and healthy life. According to the FAO concept, food security is built on four pillars: availability, access, utilisation, and stability (FAO, 1996; World Food Summit). The presented study primarily focuses on economic access, which reflects households' ability to purchase food. The analysis concentrates on the relationship between income and the proportion of food expenditures.

Engel formulated a law in 1857 describing the relationship between income and food expenditure. It states that as household income rises, the proportion of income spent on food declines, even if absolute food spending increases (Engel, 1857). However, recent economic shocks —such as the COVID-19 pandemic and subsequent inflationary pressures—challenge the stability of this relationship. When food prices increase faster than wages, households may allocate a growing share of income to food, contradicting the expected downward trend predicted by Engel's Law.

Food security of the population is significantly influenced by price dynamics, which affect affordability; regional disparities, reflecting differences in wage levels and living costs; and demographic factors, such as age, household composition, and urbanisation. Studies show that vulnerability tends to be higher among low-income households, retirees, and families with multiple children, as these groups have limited flexibility in adjusting consumption patterns (Codjia and Saghaian, 2022; Rashid et al., 2024). Rural households may exhibit greater resilience due to partial self-sufficiency, while urban households rely more heavily on market purchases.

This study adopts a conceptual model in which the income elasticity of food expenditure serves as an indicator of economic access to food. A faster decline in the share of food spending with rising income suggests improved food security; a slower decline indicates a worse situation. By examining this elasticity across regions and demographic groups during the period 2020-2023, the analysis captures both short-term crisis effects and longer-term structural changes in household behaviour. An analogical approach was used in studies by Zereyesus et al. (2025) and Bogmans et al. (2024).

Food security in Slovakia has been widely studied in relation to household consumption patterns, income elasticity, and dietary diversity. Rizov et al. (2014) investigated food security using household budget survey data from 2004-2010, estimating food demand systems and diet diversity models. Their findings show that all major food groups are normal goods, with rural and low-income households being more price- and expenditure-sensitive than urban and high-income households. Demand for meat, fish, and fruits and vegetables was found to be both expenditure and own-price elastic, indicating vulnerability to economic shocks.

Further research by Pokrivčák et al. (2014) applied the Quadratic Almost Ideal Demand System (QUAIDS) to Slovak data, confirming that cereals and meat are necessities, while dairy and fruits are perceived as luxuries. The study also highlighted that food security improved after EU accession, but disparities persisted between rural and urban households. A related analysis by Cupák et al. (2016) found that dietary diversity increased steadily from 2004 to 2010. However, economic crises temporarily reduced diversity, particularly among rural and low-income households.

Recent studies have shifted focus to crisis periods. Watter (2025) investigated the impact of COVID-19 and the war in Ukraine on household expenditure structures. He found that food expenditures rose to 30.7% of household income in 2021, signalling increased financial pressure and potential deterioration in food security. These conclusions align with broader European research under the FP7 FOODSECURE project, which emphasised affordability as the key challenge for Central and Eastern European countries, including Slovakia.

Overall, previously cited works confirm that income elasticity, price shocks, and demographic factors strongly influence food security in Slovakia. Vulnerable groups—particularly low-income households, retirees, and families with multiple children—face heightened risks during periods of economic instability. These insights provide a foundation for the present study, which extends the analysis to the most recent period (2020-2023), characterised by pandemic recovery and inflationary pressures.

The presented paper aims to analyse the development of food security in Slovakia between 2020 and 2023, focusing on economic access to food, regional disparities, and demographic factors. It also evaluates the impact of income changes on the share of food expenditures across different population groups.

Material and methods

The analysis of access to food for Slovak households and individual population groups was based on data from the Statistical Office of the Slovak Republic on household accounts for the period 2020-2023. For 2020 and 2021, the database contains information on 4,633 households; for 2022, on 4,991 households; and for 2023, on 4,881 households. Table 1 presents the list and descriptions of variables used to characterise Slovakia's food security.

Data from the Slovak household account statistics were cleaned, and extreme values were adjusted using the Winsorization method, replacing high values with the 99th percentile and very low values with the 1st percentile. The data were characterised using box plots stratified by population group. The impact of income on the share of food

Variable	Description	Units	Data source
Income	Net monthly household income after taxes	€	Statistical office of the Slovak Republic, HSBC Survey 2020-2023
Share of food expenditure	Share is calculated as the sum of food expenses divided by total expenses	Share	Statistical office of the Slovak Republic, HSBC Survey 2020-2023
Region	Regional affiliation to individual regions of the Slovak Republic: BA-Bratislava, BB-Banská Bystrica, TN-Trenčín, TT-Trnava, NR-Nitra, PO-Prešov, KE-Košice, ZA-Žilina	-	Statistical office of the Slovak Republic, HSBC Survey 2020-2023
Gender	Gender of the reference person: M-male, F-female	-	Statistical office of the Slovak Republic, HSBC Survey 2020-2023
Urbanization	City - an area with medium to high population density, Village - an area with sparse population	-	Statistical office of the Slovak Republic, HSBC Survey 2020-2023
Age category	Classification of the reference person by age category: retiree - age over 64 years, young adult - adult up to 40 years old, middle-aged - age between 40 and 64 years	-	Statistical office of the Slovak Republic, HSBC Survey 2020-2023

Source: Author´s work based on data from the Statistical Office of the Slovak Republic

Table 1: Analysed variables.

expenditures was quantified using Engel's expenditure function, as in the Equation 1. To avoid omitted-variable bias, control variables were used: gender, level of urbanisation, region, and age category.

$$p_i = c_0 + c_1 LnI_i + \beta_1 Gender_i + \beta_2 Urban_i + \\ + \sum_{i=1}^{7} \gamma_r Region_{ri} + \sum_{a=1}^{2} \delta_a \, AgeCategory_i + u_i \quad (1)$$

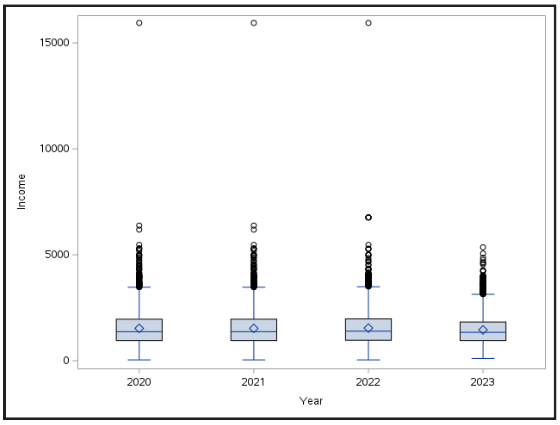
Where p_i is the share of food expenditures in total household expenditures, c_0 is a constant, LnI_i is the logged value of household income, and c_1 is the semi-elasticity expressing the impact of income on the share of food expenditures. Gender, Urban, Region, and Age are dummy variables representing these qualitative factors. In the case of gender and urbanisation, there were just two categories (1 dummy variable). The region included 8 categories, represented by 7 dummy variables, and the age category included 3 categories, represented by 2 dummy variables. Coefficients β_1, β_2, γ_r, and δ_a express the difference between the category with the dummy variable equal to 1 and the reference category. In equations estimated separately for different categories, dummy variables expressing the same category were omitted. Results include only the analysed value of income semi-elasticity influencing the share of food expenditures, as this was the main focus of the conducted analysis. Functional parameters were estimated using the Ordinary Least Squares method for different time periods. Assumptions about normality of residuals (Shapiro-Wilk test), homogeneity of variance (Breusch-Pagan test), and collinearity of variables (Variance inflation factor) were checked using standard procedures. In case of violation of these assumptions, White´s robust standard errors were used to ensure the reliability of the results.

Based on economic theory, it can be assumed that its value will be negative. This parameter was estimated for individual time periods and population groups, and the values were compared. A lower value means a faster average decrease in expenditures on basic needs with a 1% increase in income. Conversely, a higher value means that, as income increases, expenditures on the population's basic needs decrease more slowly, which may indicate a worse food security situation. The mentioned analytical procedures were carried out using SAS Enterprise Guide 7.15.

Results and discussion

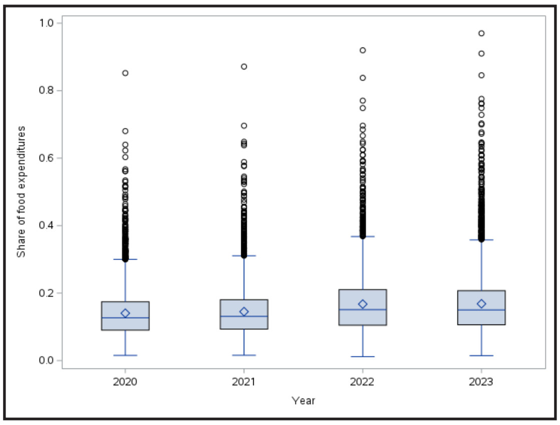
Household statistics from the Statistical Office provides a closer look at how food security is developing in Slovakia. Given the development of factors influencing food security in Slovakia at the end of the analysed period, data from 2020-2023 were used in this case, as they represent the most up-to-date available information. Figures 1 and 2 show how income and the share of expenditures on food developed during this period, that is, variables that significantly affect the population's economic access to food and their subjective perception of food security.

Figure 1 shows the development of total monthly household income, which includes both monetary and non-monetary income after tax, for the period 2020-2023. The development indicates a stable level over the first two years, averaging 1,527 €. The pandemic years 2020 and 2021 were similar in many respects, and there were no significant changes across several variables. This could have been significantly influenced by anti-pandemic measures, which significantly slowed down the dynamics of change in many areas. This concerns not only average income but also income distribution. Figure 2 shows the development of food expenditure shares over the period 2020-2023.



Source: Statistical office of the Slovak Republic

Figure 1: Household income.

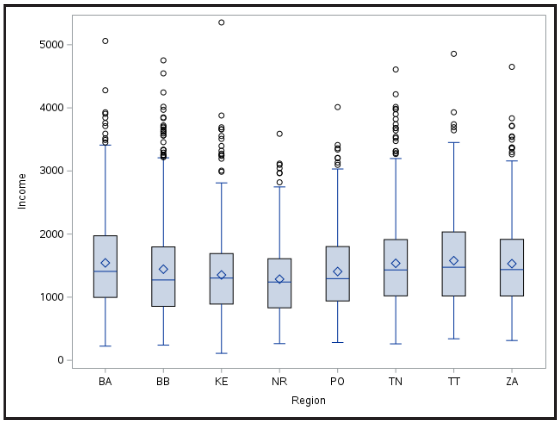


Source: Statistical office of the Slovak Republic

Figure 2: Share of food expenditures.

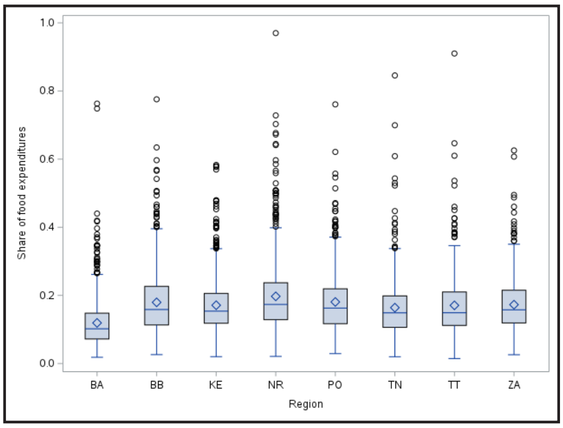
In 2022, the average income increased slightly to 1,544 €, with no significant changes in the income distribution, but in 2023, the average income declined to 1,457 €. The median income also fell from 1,396€ to 1,342 €, and the variability, as measured by the coefficient of variation, decreased from 51% to 46%. This tendency was also reflected in the share of food expenses, as shown in Figure 2. In 2020, residents spent an average of 14% of their income on food. The share increased to 15% in 2021, and despite the rise in average income in 2022, it reached 17% and remained at that level in 2023. The presented development indicates a worsening situation for households in their economic approach to food, as food prices rose faster than wages.

The distribution of household incomes in individual regions in 2023 is shown in Figure 3. Figure 4, located next to it, shows the distribution of food expenditure shares in individual regions. The total net monthly household income in Slovakia declined slightly in 2023. When comparing incomes across regions, not only are the average and median values interesting, but also the minimum, maximum, and quartiles.



Source: Statistical office of the Slovak Republic

Figure 3: Household income in regions in 2023.



Source: Statistical office of the Slovak Republic

Figure 4: Share of food expenditures in regions in 2023.

The lowest average income in 2023 was in the Nitra region, at 1,285 €, while this region also recorded the lowest income variability at 42%. Other regions with low average incomes were the Košice region (1,352 €) and the Prešov region (1,406 €). The Košice region also recorded the lowest minimum income and, along with the Nitra region, the lowest maximum income. Although the Banská Bystrica region had a higher average income, its median income was approximately equal to that of the Nitra region. The highest average income in 2023 was recorded in the Trnava region (1,578 €), followed by the Bratislava region (1,543 €) and the Trenčín region (1,536 €).

Figure 4 follows up on the distribution of food expenditure shares in Figure 3. The lowest share of food expenditure in 2023 was recorded in the Bratislava region, where households spent, on average, 12% of their income on food, with the median at 10%. The highest share of income spent on food was in the Nitra region, where it accounted for 20%, followed by the Banská Bystrica and Prešov regions, where it was 18%. The data suggest that the most vulnerable regions in Slovakia are Nitra, Banská Bystrica, and Prešov.

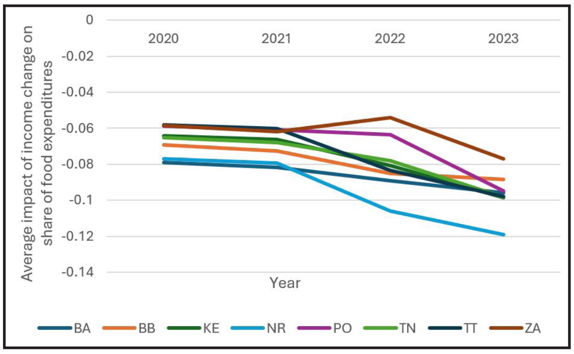
The impact of income on the share of food expenditure in Slovakia

Based on theoretical assumptions, it is expected that, as income increases, the share of expenditure on basic goods, including food, will decrease. It can be assumed that, for a part of the population in a worse food security situation, an increase in income will result in a smaller decrease in the share of food expenditure than for the population in a better economic situation. Table 2 shows the development of the impact of income on the share of food expenditure from 2020 to 2023.

A faster decline in the share of food expenditures will therefore indicate better economic access to food. The development shown in Table 2 thus suggests that, after the pandemic years, there was a slight improvement in access to food, and the share of expenditures on basic needs decreased more rapidly with higher income. Comparing confidence limits for estimated parameters suggests significant improvement in 2022 and 2023 compared with the pandemic years 2020 and 2021.

Figure 5 shows a summary comparison of average impact across regions over time. Table 3 presents estimates of the impact of income on the share of food expenditures across Slovak regions from 2020 to 2023. At the beginning of the observed period in 2020, the Bratislava region recorded the highest average income elasticity of food expenditures. On the other hand, the slowest decline in the share of food expenditures with increasing income was in the Trnava, Prešov, and Žilina regions. Over the following years, the situation slightly improved, and with higher income, the decline in food expenditures was faster.

In 2023, with rising income, spending decreased the slowest in the Žilina and Banská Bystrica regions. The value in the Bratislava region is no longer the highest, and the fastest decline in the share of spending on food was in the Nitra region. This may be due to the agricultural focus of the Nitra region and the population's ability, to some extent, to supply itself with food from its own production.



Source: Author´s work based on data from the Statistical Office of the Slovak Republic (HBS survey)

Figure 5: Impact of income increase on the share of food expenditures in the Slovak regions.

Year	Estimate	Pr > t	95% CL Lower	95% CL Upper	R-Square	Significance F
2020	-0.0675	<.0001	-0.0719	-0.0631	0.2185	<.0001
2021	-0.0702	<.0001	-0.0747	-0.0656	0.2199	<.0001
2022	-0.0823	<.0001	-0.0879	-0.0767	0.2234	<.0001
2023	-0.0969	<.0001	-0.1023	-0.0914	0.2509	<.0001

Source: Author´s work based on data from the Statistical Office of the Slovak Republic (HBS survey)

Table 2: Effect of income increase on the share of food expenditures on total expenditures.

Year	Region	Estimate	Pr > t	95% CL Lower	95% CL Upper	R-Square	Significance F
2020	BA	-0.0789	<.0001	-0.0906	-0.0672	0.276	<.0001
2020	BB	-0.0694	<.0001	-0.0794	-0.0594	0.2793	<.0001
2020	KE	-0.0641	<.0001	-0.0739	-0.0543	0.245	<.0001
2020	NR	-0.0771	<.0001	-0.0918	-0.0625	0.1836	<.0001
2020	PO	-0.0589	<.0001	-0.0717	-0.0461	0.1587	<.0001
2020	TN	-0.0653	<.0001	-0.0784	-0.0522	0.2046	<.0001
2020	TT	-0.0582	<.0001	-0.075	-0.0414	0.1283	<.0001
2020	ZA	-0.0585	<.0001	-0.0725	-0.0445	0.16	<.0001
2021	BA	-0.0817	<.0001	-0.0938	-0.0696	0.276	<.0001
2021	BB	-0.0726	<.0001	-0.083	-0.0623	0.2844	<.0001
2021	KE	-0.0664	<.0001	-0.0766	-0.0562	0.2447	<.0001
2021	NR	-0.0795	<.0001	-0.0946	-0.0644	0.1841	<.0001
2021	PO	-0.061	<.0001	-0.0743	-0.0478	0.1592	<.0001
2021	TN	-0.0681	<.0001	-0.0816	-0.0546	0.207	<.0001
2021	TT	-0.0602	<.0001	-0.0775	-0.0428	0.1291	<.0001
2021	ZA	-0.0618	<.0001	-0.0763	-0.0473	0.1639	<.0001
2022	BA	-0.0891	<.0001	-0.1027	-0.0755	0.2316	<.0001
2022	BB	-0.0851	<.0001	-0.0989	-0.0714	0.2802	<.0001
2022	KE	-0.0806	<.0001	-0.0937	-0.0674	0.204	<.0001
2022	NR	-0.106	<.0001	-0.1238	-0.0883	0.1931	<.0001
2022	PO	-0.0636	<.0001	-0.0801	-0.0471	0.1511	<.0001
2022	TN	-0.0779	<.0001	-0.094	-0.0618	0.1961	<.0001
2022	TT	-0.0833	<.0001	-0.1042	-0.0624	0.1488	<.0001
2022	ZA	-0.0541	<.0001	-0.0713	-0.0368	0.1321	<.0001
2023	BA	-0.0957	<.0001	-0.1068	-0.0847	0.2975	<.0001
2023	BB	-0.0883	<.0001	-0.1031	-0.0735	0.2109	<.0001
2023	KE	-0.0984	<.0001	-0.1123	-0.0845	0.2418	<.0001
2023	NR	-0.1192	<.0001	-0.1392	-0.0992	0.1827	<.0001
2023	PO	-0.095	<.0001	-0.1119	-0.0782	0.191	<.0001
2023	TN	-0.0981	<.0001	-0.113	-0.0833	0.2451	<.0001
2023	TT	-0.0978	<.0001	-0.1146	-0.0811	0.2022	<.0001
2023	ZA	-0.077	<.0001	-0.0927	-0.0612	0.1699	<.0001

Source: Author´s work based on data from the Statistical Office of the Slovak Republic (HBS survey)

Table 3: Effect of income increase on the share of food expenditures on total expenditures.

The advantage of household data is that it provides a more detailed view of individual segments of the population, which may be in different situations regarding food security. Tables 4 and 5 compare the impact of income across age groups and urbanisation levels. Across all groups, a similar tendency can be observed as in Table 2.

The comparison between age categories shows that the slowest decline in the food expenditure share is among the pensioner group, for whom food security is the worst. The categories of young adults (ages 18-40) and the middle generation (ages 40-64) fare better. The differences between these two categories in terms of the impact of income on the share of food expenditures decreased in 2022 and 2023. Comparison of interval estimates suggests significant differences between pensioners and other categories, with the worst situation in the pensioners' segment. Young adults and the middle-aged generation did not differ substantially, and initial differences in 2020 changed to almost identical results in 2023.

An interesting comparison is the development of income's influence on the share of food expenditures between urban and rural areas, as shown in Table 5. From similar values in 2020

Year	Age category	Estimate	Pr > t	95% CL Lower	95% CL Upper	R-Square	Significance F
2020	Pensioner	-0.0522	<.0001	-0.0618	-0.0425	0.0995	<.0001
2020	Young adults	-0.0816	<.0001	-0.0909	-0.0723	0.2601	<.0001
2020	Middle age	-0.0695	<.0001	-0.0754	-0.0636	0.2532	<.0001
2021	Pensioner	-0.0546	<.0001	-0.0645	-0.0446	0.1005	<.0001
2021	Young adults	-0.0848	<.0001	-0.0944	-0.0752	0.2625	<.0001
2021	Middle age	-0.0722	<.0001	-0.0782	-0.0661	0.2551	<.0001
2022	Pensioner	-0.0704	<.0001	-0.0823	-0.0585	0.1233	<.0001
2022	Young adults	-0.0928	<.0001	-0.1048	-0.0808	0.2384	<.0001
2022	Middle age	-0.0837	<.0001	-0.0912	-0.0761	0.2438	<.0001
2023	Pensioner	-0.0797	<.0001	-0.0892	-0.0702	0.1944	<.0001
2023	Young adults	-0.1067	<.0001	-0.1197	-0.0937	0.3044	<.0001
2023	Middle age	-0.1041	<.0001	-0.1119	-0.0963	0.2826	<.0001

Source: Author´s work based on data from the Statistical Office of the Slovak Republic (HBS survey)

Table 4: Impact of income change on share of food expenditures - age.

Year	Urbanisation	Estimate	Pr > t	95% CL Lower	95% CL Upper	R-Square	Significance F
2020	City	-0.0662	<.0001	-0.0716	-0.0607	0.2208	<.0001
2020	Rural area	-0.0705	<.0001	-0.0781	-0.0630	0.2310	<.0001
2021	City	-0.0687	<.0001	-0.0743	-0.0630	0.2218	<.0001
2021	Rural area	-0.0735	<.0001	-0.0813	-0.0657	0.2336	<.0001
2022	City	-0.0785	<.0001	-0.0854	-0.0716	0.2210	<.0001
2022	Rural area	-0.0900	<.0001	-0.0996	-0.0803	0.2441	<.0001
2023	City	-0.0924	<.0001	-0.0990	-0.0858	0.2450	<.0001
2023	Rural area	-0.1058	<.0001	-0.1153	-0.0963	0.2619	<.0001

Source: Author´s work based on data from the Statistical Office of the Slovak Republic (HBS survey)

Table 5: Impact of income change on share of food expenditures - urbanisation.

and 2021, the differences between them gradually increase. An income increase in rural areas leads to a larger decrease in the share of food expenditures than in cities. The difference between average values in urban and rural areas widened further in 2023. However, according to the confidence limits, this difference remains insignificant at the 0.05 level.

The results in Table 5 indicate that self-sufficiency options in rural areas can be strengthened with increasing income and ultimately lead to a faster decline in the share of spending on basic food than in cities. Table 6 compares the effect of income changes on the share of food expenditures by gender. The trend follows that of the previous figures and tables, and an improving situation can be observed over the years 2020-2023. Interestingly, the average parameter values suggested a slightly better situation for female households during the pandemic period. However, the difference between genders fluctuated until 2023, when the opposite result emerged, with values suggesting a better situation in households with a male as the reference person. However, the interval estimates do not show any potentially significant differences between the parameters estimated for males and females.

Based on the results presented, it can be concluded that during the period 2020-2023, there was a slight improvement in the situation, and that, with an increase in income, the decline in the share of spending on necessities for the population was faster. On the other hand, 2023 is marked by increasing differences in the rate of this decline, particularly between urban and rural areas and between men and women. It can be assumed that, in terms of food access, low-income residents in cities, especially pensioners, are at greater risk. From a territorial perspective, the Žilina, Banská Bystrica, and Prešov regions are the most vulnerable, where, with a higher average share of the population's spending on food, the decline in spending is slower at higher income levels. In the Nitra region, on the other

Year	Age category	Estimate	Pr > t	95% CL Lower	95% CL Upper	R-Square	Significance F
2020	Male	-0.0647	<.0001	-0.0707	-0.0586	0.2193	<.0001
2020	Female	-0.0726	<.0001	-0.0791	-0.0660	0.2212	<.0001
2021	Male	-0.0675	<.0001	-0.0737	-0.0613	0.2230	<.0001
2021	Female	-0.0752	<.0001	-0.0820	-0.0685	0.2213	<.0001
2022	Male	-0.0798	<.0001	-0.0876	-0.0719	0.2212	<.0001
2022	Female	-0.0874	<.0001	-0.0956	-0.0792	0.2284	<.0001
2023	Male	-0.1017	<.0001	-0.1094	-0.0940	0.2590	<.0001
2023	Female	-0.0936	<.0001	-0.1013	-0.0858	0.2535	<.0001

Source: Author´s work based on data from the Statistical Office of the Slovak Republic (HBS survey)

Table 6: Impact of income increase on share of food expenditures - gender categories.

hand, an interesting contrast emerged. Despite a higher average share of food expenditures in total expenses, the decline in this share was steepest among higher-income households.

Discussion

The paper's results confirmed the rising share of food expenditures and the income elasticity. It shows that the share of food expenditures in Slovakia increased from 14% in 2020 to 17% in 2023 despite income growth. It could be caused by food price inflation. This confirms Engel´s observation that the share of food expenditures decreases with increasing income, a pattern that may be disrupted by shocks such as higher inflation. Similar results were reported by Kubicová and Kádeková (2011) and Nagyová et al. (2013), who also found the relationship to be vulnerable to income shocks.

The study also identified significant disparities among Slovak regions in the share of food expenditures, with the highest values in Nitra, Banská Bystrica, and Prešov region (up to 20%). In comparison, in the Bratislava region it was only 12%. This aligns with Michálek's (2023) conclusion, which also confirmed persistent regional inequalities in Slovakia, and with the International Monetary Fund's report (2024).

Analysis focused on COVID and post-COVID years and confirmed findings by Watter (2025), who observed similar trends in spending by Slovak households and concluded that spending on essentials increased after the pandemic. This also aligns with global conclusions (e.g., Gebeyehu et al., 2023; Beckman et al., 2021), which confirmed increased food insecurity worldwide after the COVID-19 pandemic. This was caused especially by income shocks and rising prices.

The groups most affected by food insecurity during the COVID years were pensioners, women, and large families, which aligns with Sumsion et al. (2024). The study also confirms the findings of Millard et al. (2022) regarding rural households' greater resilience, driven by higher self-sufficiency.

Conclusion

The paper presents several main trends in Slovakia's food security. The pandemic years were characterised by relative stability in expenditure and income patterns. The reason was government measures that minimised economic shocks. However, the post-pandemic years brought new challenges. Despite a temporary increase in average income in 2022, in 2023, a decline in both median and average household income followed, accompanied by a continued increase in the share of food expenditures. This indicates that food price inflation outpaced wage growth, which reduced economic access to food for many households.

Based on the paper´s results, significant regional disparities can be concluded. At the same time, Bratislava maintained the lowest share of food spending, in Nitra, Banská Bystrica, and Prešov, the highest values were recorded, with food expenditures reaching up to 20% of household income. Interestingly, the Nitra region, despite its high food expenditure share, recorded the fastest decline in this share as income rose. The reason could be the local agricultural resources, which may provide some resilience to food security in this region.

The demographic analysis focused on identifying vulnerable groups. Results showed that retirees allocate a larger share of their income to food, making them more vulnerable to inflation and price fluctuations. Age and urbanisation were also found to be significant factors. Rural households are better positioned to reduce their food expenditure share as income rises, possibly due to self-sufficiency options.

These findings emphasise two main conclusions. While income growth generally improves access to food across regions, significant structural inequalities persist. Policy makers should focus on supporting the most vulnerable groups, which are low-income urban households and pensioners. Important issues in Slovak food security also led to significant disparities in regional vulnerability.

Acknowledgements

The paper presents the results of research supported by the VEGA grant, Climate Change and Food Security in Europe - Economic and Environmental Challenges (project number: 1/0256/26).

Corresponding author:
Ing. Jozef Palkovič, PhD.
Institute of Statistics, Operations Research and Mathematics, Faculty of Economics and Management
Slovak University of Agriculture in Nitra, Tr. A. Hlinku 2, 94976 Nitra , Slovakia
Phone: +421 376 414 162, E-mail: jozef.palkovic@uniag.sk

References

[1] Beckman, J., Baquedano, F. and Countryman, A. (2021) "The impacts of COVID-19 on GDP, food prices, and food security", *Q Open*, Vol. 1, No. 1, p. qoab005. ISSN 2633-9048. DOI 10.1093/qopen/qoab005.

[2] Bogmans, C., Pescatori, A. and Prifti, E. (2024) "How do Economic Growth and Food Inflation Affect Food Insecurity?", *IMF Working Papers*, 2024(188), p. A001. ISSN 1020-792X. DOI 10.5089/9798400287336.001.A001.

[3] Codjia, C. O. and Saghaian, S. H. (2022) "Determinants of Food Expenditure Patterns: Evidence from U.S. Consumers", *Sustainability*, Vol. 14, No. 13, p. 8156. ISSN 2071-1050. DOI 10.3390/su14138156.

[4] Cupák, A., Pokrivčák, J. and Rizov, M. (2016) " "Diverzifikácia spotreby potravín na Slovensku (Diversity of Food Consumption in Slovakia)", *Politická Ekonomie*, Vol. 64, No. 5, pp. 608-626. ISSN 0032-3233. DOI 10.18267/j.polek.1082. (In Slovak).

[5] Engel, E. (1857) "Die productions-und consumtionsverhältnisse des Königreichs Sachsen", *Zeitschrift des Statistischen Bureaus des Königlich Sächsischen Ministeriums des Innern*, Vol. 8, pp. 1-54. ISSN 2628-359X. (In German).

[6] FAO (1996) "An Introduction to the Basic Concepts of Food Security", *World Food Summit*. FAO, Rome.

[7] Gebeyehu, D. T., East, L., Wark, S. and Islam, M. S. (2023) "A systematic review of the direct and indirect COVID-19's impact on food security and its dimensions", *BMC Public Health*, Vol. 23, p. 2298. ISSN 1471-2458. DOI 10.1186/s12889-023-17104-6.

[8] International Monetary Fund, European Department. (2024) "Regional Inequality in Slovakia", *IMF Staff Country Reports*, 2024(076), p. A003. ISSN 1934-7685. DOI 10.5089/9798400271021.002.A003.

[9] Kubicová, Ľ. and Kádeková, Z. (2011) "Comparison of the income development and the food demand elasticities of the private households in Slovakia", *Agricultural Economics - Czech*, Vol. 57, No. 8, pp. 404-411. ISSN 0139-570X. DOI 10.17221/36/2010-AGRICECON.

[10] Nagyová, Ľ., Stávková, J. and Kádeková, Z. (2013) "Income disparities and their impact on the level and development of food expenditures of households in Slovakia", *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis*, Vol. 61, No. 7, pp. 2569-2575. ISSN 1211-8516. DOI 10.11118/actaun201361072569.

[11] Michálek, A. (2023) "Income inequalities and poverty in Slovakia: Development and changes", *European Spatial Research and Policy*, Vol. 30, No. 2, pp. 207-233. ISSN 1231-1952. DOI 10.18778/1231-1952.30.2.12.

[12] Millard, J., Sturla, A., Smutná, Z., Duží, B., Janssen, M. and Vávra, J. (2022) "European Food Systems in a Regional Perspective: A Comparative Study of the Effect of COVID-19 on Households and City-Region Food Systems", *Frontiers in Sustainable Food Systems*, Vol. 6. ISSN 2571-581X. DOI 10.3389/fsufs.2022.844170.

[13] Pokrivčák, J., Cupák, A. and Rizov, M. (2014) "Food demand and consumption patterns in the new EU member states: The case of Slovakia", EAAE Congress Paper.

[14] Rashid, F. N., Sesabo, J. K., Lihawa, R. M. and Mkuna, E. (2024) "Determinants of household food expenditure in Tanzania", *Agriculture & Food Security*, Vol. 13. ISSN 2048-7010. DOI 10.1186/s40066-023-00462-0.

[15] Rizov, M., Cupák, A. and Pokrivčák, J. (2014) "Food security and household consumption patterns in Slovakia", LICOS Discussion Paper No. 360.

[16] Sumsion, R. M., June, H. M. and Cope, M. R. (2023) "The impact of COVID-19 on food security: a review", *SN Social Sciences*, Vol. 3, p. 176. ISSN 2662-9283. DOI 10.1007/s43545-023-00762-z.

[17] Watter, P. (2025) "Food Security in Eastern and Central Europe During the COVID-19 Pandemic: A Case Study of Slovakia", *International Scientific Days 2024 - Proceedings*, pp. 454-461. ISBN 978-80-552-2816-7. DOI 10.15414/2024.9788055228167.454-461.

[18] Zereyesus, Y. A., Xia, T., Nava, N. J., Li, X. and Cardell, L. (2025) "*International income and price elasticity estimates: An update*", Economic Research Service, U.S. Department of Agriculture, Report No. TB-1971. DOI 10.32747/2025.9228006.ers.

Volume XVIII Number 3, 2026

Data Management in Agricultural Enterprises: A Framework for Data-Driven Decision Support

Michal Stočes[1], Jan Jarolímek[1], Irena Benešová[2], Pavel Šimek[1], Zdeňka Žáková Kroupová[2], Ludmila Pánková [2], Jiří Vaněk[1], Eva Kánská[1], Martin Havránek[1], Michael Anderle[1]

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

[2] Department of Economics, Faculty of Economics and Management, Czech University of Life Sciences Prague, Czech Republic

Abstract

The growing digitalisation of agriculture is leading to a rapid increase in the volume of data produced by agricultural enterprises; however, its practical use is often limited by fragmented data sources, varying data quality, insufficient metadata, and inconsistent storage and management practices. The aim of this article is to propose a practical framework for the systematic management of data in agricultural enterprises to support its long-term availability, reuse, and integration into decision-making processes. The proposed approach is based on four interlinked steps: identifying data needs and issues, creating and recording a data inventory, drawing up a Data Management Plan (DMP), and implementing it in practice. The framework links technical aspects of data management – such as data formats, metadata, quality, integrity, security, backup and archiving – with the definition of responsibilities for individual staff members and farm (agricultural enterprise) management. For practical implementation, two basic approaches to data storage are distinguished: file systems and data platforms or repositories. The more advanced platform-based approach enables the centralisation of heterogeneous data sources, the automation of their transfer and processing, and subsequently the integration of the acquired data with analytical tools and decision-support systems. The proposed framework thus represents a path from the isolated collection of operational data to its systematic management and utilisation as a long-term information source for both operational and strategic decision-making within an agricultural enterprise.

Keywords

Agricultural data, Data Management Plan, data quality, metadata, precision agriculture, Institutional Data Repositories, Decision Support Systems, FAIR data principles .

Stočes, M., Jarolímek, J., Benešová, I., Šimek, P., Žáková Kroupová, Z., Pánková, L., Vaněk, J., Kánská, E., Havránek, M. and Anderle, M. (2026) "Data Management in Agricultural Enterprises: A Framework for Data-Driven Decision Support", *AGRIS on-line Papers in Economics and Informatics*, Vol. 18, No. 3, pp.51-60. ISSN 1804-1930. DOI 10.7160/aol.2026.180305.

Introduction

The digitalisation of agriculture is leading to a rapid increase in the volume of data generated during routine operational, production and administrative processes. Data is generated by agricultural machinery, sensors, information systems, agricultural enterprise staff and external organisations. Its importance is growing further with the development of precision farming, automation, Business Intelligence (BI), Decision Support Systems (DSS) and artificial intelligence methods. The main sources of data for agricultural enterprises are the enterprises themselves (internal data), technology suppliers, public administration, the market environment and research organisations. The growing availability of these sources creates significant potential for data-driven decision-making (Oral et al., 2023), whilst also placing high demands on systematic data management.

However, the sheer volume of available data does not automatically guarantee its practical usability. Previous research shows that a significant proportion of agricultural data remains in formats unsuitable for subsequent analytical processing. Records often remain isolated in separate applications, on local personal computers or in suppliers' accounts. Retrieving data then depends on a specific employee's personal knowledge, and linking data requires repeated exports and manual adjustments. The most common problems include low consistency, incompleteness, a lack of context and metadata (Stočes et al., 2018), or errors arising from manual entry. Operational and analytical uses of data, however, have different requirements: whilst incomplete data may suffice for immediate machine control or operational intervention, a lack of context is a major obstacle for long-term evaluation, prediction and comparison of production cycles. Without precise knowledge of the location, time, sensors used and data acquisition methodology, it is not possible to reliably compare individual production periods.

Metadata is a key element in preserving the meaning and traceability of data. (Michener et al., 1997) Metadata describes the original source, structure, format, temporal and spatial coverage, and currency of the data, and supports their correct interpretation and integration across different information systems. In an agricultural context, metadata includes, for example, geographical location, time of collection, equipment identification, meteorological conditions, and information on soil and crops. Data that is inadequately described loses its informational value over time, and its further processing becomes unfeasible (Šimek et al., 2013).

Another significant problem is fragmented storage and limited interoperability. Data formats that are dependent on the original software mean that a file that has been technically preserved may not be practically usable in the future (proprietary data format). Another potential problem is that the data is owned by the service provider. In this context, the FAIR principles (Findable, Accessible, Interoperable, Reusable) (Kumar et. al., 2024) provide a relevant general framework not only for scientific data but also for the business environment of agricultural organisations.

This paper builds directly on the authors' previous research. The study "*Information sources in agriculture*" (Jarolímek et al., 2024) identified the main sources of agricultural data and highlighted the importance of a Data Management Plan (DMP) (Michener, 2015). for their effective and secure management. The follow-up study, "*Agriculture Data Platform – Institutional Data Repository – Selected Aspects*" (Stočes et al., 2023), focused on the technological requirements of a data platform, including repository architecture, data types, metadata, sharing, interoperability and the FAIR principles. The results show that the organisational and technological aspects of data management are closely interrelated.

The aim of this article is to propose a comprehensive and practical methodological framework for the systematic management of data in agricultural enterprises, integrating the identification of data sources, their recording, the creation of a Data Management Plan, and its subsequent implementation. The framework covers the entire data lifecycle, from acquisition and storage through to analytical processing and use for both operational and strategic decision-making within the enterprise, whilst assessing two basic options for technological implementation: an organised file system and a data platform.

Materials and methods

The framework was based on the methodology "*Data management in the agricultural sector*" (Stočes et al., 2026) and on related analyses of agricultural information sources. These materials are the result of long-term collaboration and structured interviews with representatives of agricultural, forestry and food businesses, technology suppliers and public administration bodies. The results of the authors' previous research, focusing on the architecture of data repositories and institutional data platforms, were also utilised. To design the framework, the available data sets, methods of their storage and the practical needs of business management were analytically assessed. Legal and contractual frameworks governing the protection, access and use of data were also considered, in particular the General Data Protection Regulation (GDPR) (European Parliament and Council of the European Union, 2016). and the European Data Act (European Parliament and Council of the European Union, 2023).

The methodological approach combined content analysis, qualitative comparison and design synthesis. Content analysis was used to identify and categorise data management requirements across the phases of the data life cycle: acquisition, description, storage, protection, and processing. For each requirement, the operational problem being addressed and a verifiable output were defined (e.g. assigning a metadata record to missing context, or a requirement for verifiable export to avoid dependence on a supplier).

The proposed synthesis subsequently integrates the identified requirements into a comprehensive four-step methodological framework for data management: Identification of data needs and issues – assessment of data origin, traceability and portability. Data registration and inventory creation – compilation of a structured overview of datasets, their location, formats and responsibilities. Preparation of a Data Management Plan – formalisation of rules for description, storage, backup, archiving, access and ethical and legal aspects. Implementation and roll-out – application of DMP rules in day-to-day practice, setting up processes and staff training.

The FAIR principles (Findable, Accessible, Interoperable, Reusable) have been integrated into the methodological framework.

The qualitative comparison assessed two basic architectural options for data storage and management: a controlled file system and a data platform or repository. The evaluation criteria included data organisation, metadata management, processing automation, interoperability, traceability of changes, security and overall operational complexity (Lin et al., 2020).

Metadata requirements were assessed according to their ability to ensure long-term traceability, correct interpretation and repeated processing of data without reliance on the creator's personal memory. The basic scope of a dataset description includes the dataset identifier, source, custodian, purpose of use, and temporal and spatial boundaries. At the attribute level, the framework requires the development of a data dictionary that defines the meaning of items, units of measurement, code lists, known sensor calibration conditions, and the method for recording missing or estimated values.

Drawing on the data governance roles discussed by Abraham et al. (2019), the procedure adapts this framework into four basic organisational roles to support operational discipline and enable staff to cover one another's responsibilities. To ensure operational discipline and substitutability, the procedure defines four basic organisational roles: *Originator*: creates the primary record and provides details of the circumstances of its creation. *Administrator*: checks the structure, quality, storage and documentation of the dataset. *User*: assesses the suitability of the source data for a specific decision-making task. *Company management*: provides the necessary resources, approves the DMP rules and contractual terms.

The continuity of processing was assessed through the controlled separation of three data layers: *Raw data layer*: represents data in its as-is or unmodified form from devices, applications or external services. *Cleansed data layer*: includes the results of data cleansing, the removal of duplicates, the harmonisation of units of measurement and the standardisation of identifiers. *Analytical layer*: contains aggregated datasets, spatial layers, calculated indicators and data for visualisation in analytical and decision-making tools (e.g. Microsoft Excel, Power BI, QGIS, R, Python, Grafana) (Grafana Labs, 2024; Negi et al., 2025).

A key validation criterion was the ability to unambiguously link the derived analytical output to a specific version of the input data and the calculation method used. Routine transformations and data cleansing must not overwrite the original records, whilst the retention of both the originals and the processed datasets is subject to specified retention rules and disposal periods.

This work is of a conceptual and design nature. The result is a methodological data management model adaptable to the technical, personnel and economic capabilities of a specific agricultural enterprise. As part of the assessment, a model data process was created, linking both automated (e.g. IoT data sources) and manual data collection, through the storage of primary data, to analytical models and support for both operational and strategic management decision-making.

The internal consistency of the proposal was assessed based on the logical sequence of individual steps – from the identification and recording of sources, through metadata description and secure storage, to the final analytical reports. Economic return on investment and direct operational benefits were not empirically measured at this stage and represent a direction for follow-up research in the real-world operations of agricultural enterprises. Saiz-Rubio and Rovira-Más, 2020).

Results and discussion

Classification of sources and data formats

The data overview for an agricultural holding distinguishes between three main categories of sources: agricultural enterprise data e.g. operational records, agrotechnical operations, sensor measurements, financial and economic data, public sources e.g. LPIS (Land Parcel Identification System), land registry, meteorological data, and open statistical data) and external non-public sources (paid consultancy services, supplier databases, measurements from contracted machinery).

The origin of the data, access regime and technical format are distinct characteristics. The choice of format has a significant impact on portability. Open formats with publicly available specifications protect against dependence on specific software, whilst proprietary formats require licences or conversion tools. (Falcão et al., 2023).

Data representations and requirements for their preservation

In practice, the condition for further processing means that it is not enough simply to store data physically on a disk; it is also necessary to attach a basic technical and content-related description (metadata) to it, without which the file will lose its meaning over time. If this condition is not met, the file remains technically preserved but is practically unusable for analysis. Put simply, this is a minimum set of instructions that enables both humans and software tools to correctly retrieve and accurately interpret the data in the future. Table 1 provides an overview of the most common data representations found on agricultural holdings, together with their typical uses and specific conditions for further processing (Nurseitov et al., 2009; Butler et al., 2016).

Practical implementation of the framework:

- *Identification and registration of data sources:* The organisation shall compile a register of all available datasets. For each dataset, the source, format, location, intended use, historical scope and dependence on an external supplier shall be recorded. The register will reliably identify duplicates and gaps in coverage.
- *Creation and updating of the Data Management Plan:* The DMP formalises file-naming conventions, object identifiers (e.g., unique land parcel code), data dictionaries, backup rules, retention periods, and contingency procedures in the event of a supplier service outage.
- *Defining organisational roles and responsibilities:* Effective governance requires a clear division of responsibilities:
    - *Data originator:* Responsible for the initial collection of the record and for providing details of the circumstances of its creation (e.g. machine operator, agronomist).
    - *Data manager:* Checks the structure, validity and completeness of the metadata description and the storage of the dataset.

Data Format	Typical Uses	Conditions for further processing
CSV	Spreadsheet exports, sensor logs, time series	Describe the delimiter, character encoding (UTF-8), units of measurement and column data types.
XLSX, ODS	Spreadsheet records, operational calculations and working analyses	Carefully separate the input data from the formulas; describe the structure of the worksheets and the meaning of the data cells.
GeoPackage, GeoTIFF	Vector (block boundaries) and raster (yield maps, orthophotos) spatial data	Keep the reference coordinate system (CRS, e.g. S-JTSK, WGS84), extent and resolution.
JSON, XML	Structured exchange of records between systems, e.g. REST API	Keep the data schema (XSD/JSON Schema), the meanings of the attributes and the interface version number.
Proprietary formats	Exports of machine terminals and specialised software	To provide technical documentation, a verified conversion to an open format and access to the original application.
Image formats and images (JPEG, PNG, TIFF)	Photographic documentation of vegetation, images from unmanned aerial vehicles (UAVs), sensors, remote sensing data	Retain EXIF metadata, GPS location, capture time, spectral bands, resolution and sensor calibration.
IoT data streams and time series (Parquet, InfluxDB, MQTT)	Continuous monitoring of the microclimate, soil moisture and stable operations	Define the sampling frequency, calibration curves, UTC timestamp and rules for measurement failures.

Source: Own processing

Table 1: Overview of data representations in agricultural practice.

- *Data user:* Uses the data for analytical evaluation and assesses its suitability for a given decision.
- *Company management*: Ensures the necessary capacity and investment in IT, and approves the company's data strategy.

Metadata and the separation of data layers

A key architectural principle of the proposed framework is the strict separation of original records from processed layers: *Raw Data:* This is stored in an unaltered form as received from telemetry, sensors or imports. The originals are never overwritten, which guarantees reproducibility. *Cleansed Layer* (cleansed data): Created by applying validations, standardising units, removing errors and adding identifiers. The cleansing process generates a log of the changes made. *Analytical Layer* (Processed Data): Contains aggregated datasets, yield maps, cost calculations and key performance indicators (KPIs) prepared for visualisation and as input for DSS (Wieder and Nolte, 2022).

A comparison of file-based and platform-based solutions

In the practical implementation of data management on an agricultural enterprise, a distinction is made between two basic technological levels: a managed file system and a data platform or repository. The choice of a specific architecture depends on the volume and diversity of the data generated, the size of the agricultural enterprise, the technical infrastructure, and the organisation's human and financial resources. A managed file system represents a technically less complex solution, suitable primarily for a manageable range of sources, administrative tasks and less frequent processing. Its reliability relies on strict adherence to a uniform directory structure, file-naming conventions, the maintenance of a set register and the use of accompanying metadata files or a data dictionary, which requires a high degree of discipline. In contrast, a data platform (repository) enables the centralised storage of heterogeneous data sources from machines, sensors and applications, the automation of data flows via APIs, and direct integration with analytical and decision-making tools. The platform-based approach replaces manual steps with controlled formation processes; upon ingestion, it automatically checks the structure and mandatory attributes and ensures precise traceability of data provenance across layers. However, it requires a higher level of technical management, continuous monitoring and dedicated administrative support. The transition to a data platform alone will not resolve incomplete or erroneous records; its successful implementation is contingent upon the prior harmonisation of identifiers and the standardisation of metadata descriptions as early as the file stage (Fountas et al. 2015; Wilson et al. 2017).

When selecting the technological infrastructure, the organisation is considering two main options. Table 2 presents a comparative analysis based on five defined qualitative criteria (Sawadogo and Darmont, 2021).

The criterion	Managed File System	Data Platform / Repository
Data organisation	Structured directories, naming conventions and manual set registration	Data catalogue, unique identifiers and relationships between objects
Metadata Management	Accompanying text and file metadata records and dictionaries	A centralised description integrated directly with the dataset and checks
Data Processing	Manual imports/transfers or batch scripts	Managed automated ETL/ELT (Extract, Transform, Load or Extract, Load, Transform) processes and integration channels
Traceability	File versioning and manual change logs	Automatic tracking of runs, data lineage and transformation history
Operational Complexity	Low technical complexity, high demands on discipline at work	It needs professional administration, monitoring and technical support

Source: Own processing

Table 2: Comparison of file-based and platform-based solutions.

Integration with analytical software

Cleaned and structured data from a data platform or managed repository can be used in various types of analysis and decision-support tools, depending on users' needs (Wolfert et al., 2017), for example:

- *Spreadsheet programmes* (e.g. Microsoft Excel): These are used for quick data checks, creating simple reports and pivot tables, or analysing smaller data sets.
- *Reporting and decision-support tools* (e.g. Microsoft Power BI): These enable data from different areas of the business – such as agronomic and financial data – to be linked together to create interactive reports for business management.
- *Geographic Information Systems* (e.g. QGIS): These enable the analysis of spatial data and the combined display of, for example, soil analysis results, yield maps and plot boundaries. To ensure these data are correctly linked, their coordinate systems must be aligned.
- *Programming languages and data analysis tools* (e.g., R and Python): These are used for advanced statistical analysis, data processing, machine learning, and the automation of repetitive procedures.
- *Tools for real-time monitoring and data visualisation* (e.g. Grafana): These are used to create dashboards that display trends in measured values over time and the current status of equipment, for example, based on data from sensor networks. They also allow automatic alerts to be set up when specified thresholds are exceeded or abnormal conditions occur.

The legal assessment of data management emphasises that the requirement to "own the data" necessitates a contractual definition of specific user rights. Contracts with technology suppliers should guarantee the availability of complete and unaltered data exports, including the necessary metadata. Attention must also be paid to proprietary formats, whose use may be restricted to a specific supplier's software. To mitigate the risk of vendor lock-in (dependence on a single supplier), it is advisable to contractually ensure the ability to export data to open or widely supported formats that enable transfer to other systems and further use. The framework also takes into account compliance with the GDPR when handling employees' personal data, as well as the provisions of the European Data Act, which, among other things, regulates users' rights to access data generated during the use of connected agricultural machinery and equipment and to share such data with third parties. (Wiseman et al., 2019).

It is recommended that the proposed framework be implemented gradually, considering the agricultural enterprise size, available technical equipment, and staff capacity. A suitable starting point is to select a single high-value-added area, such as monitoring costs and yields for individual land parcels. This task can be used to verify the availability and quality of the necessary data, the possibilities for linking it, and the usefulness of the resulting reports for decision-making. Experience gained from the initial implementation can then be used to refine data management rules and to gradually extend the framework to other areas of the agricultural enterprise. The framework is designed as a generally applicable methodological guide, the specific form of which must be tailored to the needs and capabilities of the individual agricultural enterprise. The framework itself does not constitute proof of economic return or an increase in operational efficiency. Verification of these benefits will be the subject of follow-up research based on the practical implementation of the framework and an evaluation of the results achieved. The evaluation should include, for example, the costs of implementation and operation, the time required to prepare data and reports, the error rate of the data, and its availability for decision-making. A comparison of the baseline situation with the situation following the implementation of the framework will enable an assessment of its actual benefits and limitations under the specific conditions of the agricultural enterprise (Žáková Kroupová et al., 2025).

Conclusion

The proposed approach demonstrates that the effective use of data on an agricultural enterprise requires more than just its collection and technical storage. A key prerequisite is the systematic management of data throughout its entire lifecycle, linking the identification of data sources, the creation of a data inventory, the definition of operational rules within the Data Management Plan, and the subsequent implementation of these rules into the agricultural enterprise's day-to-day operations. A process setup in this way creates the conditions for the long-term availability, traceability, quality, and reusability of data in accordance with the international FAIR principles.

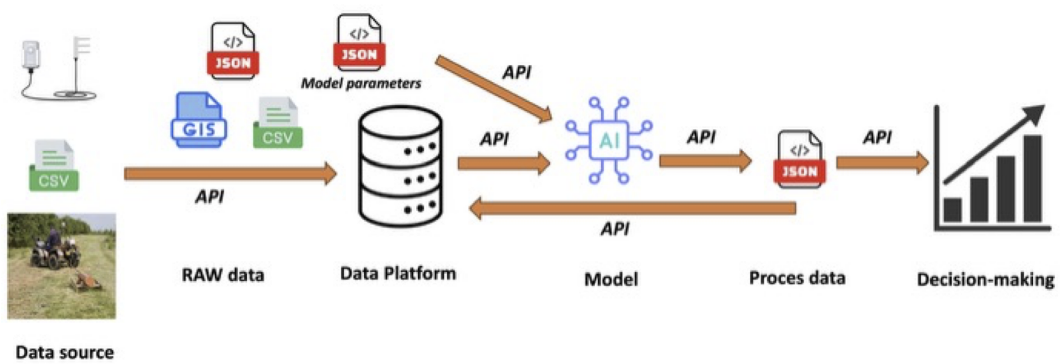
The results confirm the particular importance of data quality and standardised data description. Data that is inconsistent, insufficiently complete or lacking appropriate metadata has limited value for advanced analytics, decision support systems (DSS) and the application of machine learning and artificial intelligence methods. Data quality is not merely a technical issue but is fundamentally influenced by human factors and established working procedures (Milleret el., 2024). For this reason, the Data Management Plan should be viewed as a continuously updated management tool, rather than a one-off document. Regular reviews, updates, staff training and adaptation to new technologies and legislative requirements (e.g. the GDPR and the Data Act) are essential for its long-term effectiveness.

A key benefit of the framework is the comprehensive integration of organisational data management with its technological storage and automated analytical processing. As illustrated by the model data process (Figure 1), the entire chain links together the individual stages from primary collection through to decision support:

- *Acquisition of primary data* (Data Source → RAW Data): Data from automated IoT sensors, mobile measuring devices or structured files (CSV, GIS) is transferred via an application programming interface (API) to the raw data layer.
- *Centralised storage* (Data Platform): The data platform stores raw records in their unprocessed form (RAW data), manages metadata structures and model parameters, and strictly separates the original data from processed datasets.
- *Model processing and AI* (Model / AI → Processed Data): The platform communicates via a bidirectional API with analytical and predictive models utilising artificial intelligence algorithms, which transform the raw data into validated and structured derived data (Processed Data).
- *Decision-making support:* The processed data is delivered to management dashboards, GIS systems or analytical tools for both operational and strategic business management.

Practical implementation also confirms that no single technological solution is suitable for all agricultural businesses. The choice between a managed file system and an advanced data platform must be tailored to the agricultural enterprise's size, the nature and volume of the data being processed, the available infrastructure, and the agricultural enterprise's human and financial resources. In both cases, however, contractual control over the data, format portability and the avoidance of vendor lock-in are essential.

The main benefit of the proposed framework is therefore the possibility of a gradual transition for an agricultural enterprise from a situation in which data is used only once for a specific operational task to one in which it constitutes a long-term managed information resource. Further research should focus on the practical validation of the entire framework in enterprises of various sizes and on quantifying its impact on data quality, data management costs and the effectiveness of decision-making processes.



Source: Own processing

Figure 1: Automated model data process.

Acknowledgements

The results and knowledge included herein have been obtained owing to support from the following institutional grant: Internal grant agency of the Faculty of Economics and Management, Czech University of Life Sciences Prague, grant no. 2026A007.

Corresponding author:
doc. Ing. Pavel Šimek, Ph.D.
Department of Information Technologies, Faculty of Economics and Management
Czech University of Life Sciences Prague
Kamýcká 129, 165 00 Prague-Suchdol, Czech Republic
Phone: +420 224386680, E-mail: simek@pef.czu.cz

References

[1] Abraham, R., Schneider, J. and vom Brocke, J. (2019) "Data governance: A conceptual framework, structured review, and research agenda", *International Journal of Information Management*, Vol. 49, pp. 424-438. ISSN 0268-4012. DOI 10.1016/j.ijinfomgt.2019.07.008.

[2] Butler, H., Daly, M., Doyle, A., Gillies, S., Hagen, S. and Schaub, T. (2016) "*The GeoJSON Format*", RFC 7946, Internet Engineering Task Force (IETF), Aug. 2016. ISSN 2070-1721. DOI 10.17487/RFC7946.

[3] European Parliament and Council of the European Union. (2016) "Regulation (EU) 2016/679 of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)", *Official Journal of the European Union*, L 119, 4 May 2016, pp. 1-88. ISSN 1977-0677 [Online]. Available: https://eur-lex.europa.eu/eli/reg/2016/679/oj. [Accessed: 15 Jun. 2026].

[4] European Parliament and Council of the European Union. (2023) "Regulation (EU) 2023/2854 of 13 December 2023 on harmonised rules on fair access to and use of data and amending Regulation (EU) 2017/2394 and Directive (EU) 2020/1828 (Data Act)", *Official Journal of the European Union*, L, 2023/2854, 22 Dec. 2023. E-ISSN 1977-0677 [Online]. Available: https://eur-lex.europa.eu/eli/reg/2023/2854/oj. [Accessed: 15 Jun. 2026].

[5] Falcão, R., Matar, R., Rauch, B., Elberzhager, F. and Koch, M. (2023) "A Reference Architecture for Enabling Interoperability and Data Sovereignty in the Agricultural Data Space", *Information*, Vol. 14, No. 3, p. 197. ISSN 2078-2489. DOI 10.3390/info14030197.

[6] Fountas, S., Carli, G., Sørensen, C. G., Tsiropoulos, Z., Cavalaris, C., Vatsanidou, A., Liakos, B., Canavari, M., Wiebensohn, J. and Tisserye, B. (2015) "Farm management information systems: Current situation and future perspectives", *Computers and Electronics in Agriculture*, Vol. 115, pp. 40-50. E-ISSN 1872-7107. DOI 10.1016/j.compag.2015.05.011.

[7] Grafana Labs. (2024) "*Grafana Alerting*", Grafana documentation. [Online]. Available: https://grafana.com/docs/grafana/latest/alerting/. [Accessed: 17 Jun. 2025].

[8] Jarolímek, J., Samek, J., Šimek, P., Stočes, M., Vaněk, J. and Pavlík, J. (2024) "Information sources in agriculture", *Plant, Soil and Environment*, Vol. 70, No. 11, pp. 712-718. E-ISSN 1805-9368. DOI 10.17221/361/2024-PSE.

[9] Kumar, P., Hendriks, T., Panoutsopoulos, H. and Brewster, C. (2024) "Investigating FAIR data principles compliance in Horizon 2020 funded agri-food and rural development multi-actor projects", *Agricultural Systems*, Vol. 214, p. 103822. E-ISSN 1873-2267. ISSN 0308-521X. DOI 10.1016/j.agsy.2023.103822.

[10] Lin, D., Crabtree, J., Dillo, I., Downs, R. R., Edmunds, R., Giaretta, D., De Giusti, M., L'Hours, H., Hugo, W., Jenkyns, R., Khodiyar, V., Martone, M. E., Mokrane, M., Navale, V., Petters, J., Sierman, B., Sokolova, D. V., Stockhause, M. and Westbrook, J. (2020) "The TRUST Principles for digital repositories", *Scientific Data*, Vol. 7, p. 144. ISSN 2052-4463. DOI 10.1038/s41597-020-0486-7.

[11] Michener, W. K., Brunt, J. W., Helly, J. J., Kirchner, T. B. and Stafford, S. G. (1997) "Nongeospatial Metadata for the Ecological Sciences", *Ecological Applications*, Vol. 7, No. 1, pp. 330-342. E-ISSN 1939-5582. DOI 10.1890/1051-0761(1997)007[0330:NMFTES]2.0.CO;2.

[12] Michener, W. K. (2015) "Ten Simple Rules for Creating a Good Data Management Plan", *PLOS Computational Biology*, Vol. 11, No. 10, p. e1004525. ISSN 1553-7358. DOI 10.1371/journal.pcbi.1004525.

[13] Miller, R., Whelan, H., Chrubasik, M., Whittaker, D., Duncan, P. and Gregório, J. (2024) "A Framework for Current and New Data Quality Dimensions: An Overview", *Data*, Vol. 9, No. 12, p. 151. ISSN 2306-5729. DOI 10.3390/data9120151.

[14] Negi, A., Tripathi, L., Prabha, S. and Azad, C. (2025) "Charting the Future: An In-Depth Review of Cuttingedge Big Data Visualization Tools for 2025", *2025 12th International Conference on Computing for Sustainable Global Development (INDIACom), Delhi, India, IEEE,* pp. 1-6. ISSN 0973-7529. DOI 10.23919/INDIACom66777.2025.11115697.

[15] Nurseitov, N., Paulson, M., Reynolds, R. and Izurieta, C. (2009) "Comparison of JSON and XML Data Interchange Formats: A Case Study", *Proceedings of the ISCA 22nd International Conference on Computer Applications in Industry and Engineering (CAINE), San Francisco, USA, ISCA,* pp. 157-162. ISBN 9781615676668. [Online]. Available: https://www.cs.montana.edu/izurieta/pubs/caine2009.pdf. [Accessed: 20 Jun. 2025].

[16] Oral, E., Chawla, R., Wijkstra, M., Mahyar, N. and Dimara, E. (2023) "From Information to Choice: A Critical Inquiry Into Visualization Tools for Decision Making", *IEEE Transactions on Visualization and Computer Graphics*, Vol. 30, No. 1, pp. 359-369. ISSN 1077-2626. DOI 10.1109/TVCG.2023.3326593.

[17] Saiz-Rubio, V. and Rovira-Más, F. (2020) "From Smart Farming towards Agriculture 5.0: A Review on Crop Data Management", *Agronomy*, Vol. 10, No. 2, p. 207. ISSN 2073-4395. DOI 10.3390/agronomy10020207.

[18] Sawadogo, P. and Darmont, J. (2021) "On data lake architectures and metadata management", *Journal of Intelligent Information Systems*, Vol. 56, pp. 97-120. E-ISSN 1573-7675. DOI 10.1007/s10844-020-00608-7.

[19] Stočes, M., Jarolímek, J., Benešová, I., Cihelka, P., Čechura, L., Havránek, M., Žáková Kroupová, Z., Pánková, L. and Vaněk, J. (2026) "*Správa dat v zemědělském sektoru*", Prague, Faculty of Economics and Management, Czech University of Life Sciences Prague. DOI 10.83124/7yfg4-bqs05. (In Czech).

[20] Stočes, M., Šilerová, E., Vaněk, J., Jarolímek, J. and Šimek, P. (2018) "Možnosti využití otevřených dat v sektoru cukr-cukrová řepa", *Listy cukrovarnické a řepařské*, Vol. 134, No. 3, pp. 117-121. ISSN 1210-3306. (In Czech).

[21] Stočes, M., Vaněk, J., Jarolímek, J., Novák, V., Masner, J., Šimek, P., Kánská, E., Havránek, M., Kubata, K. and Voral, V. (2023) "Agriculture Data Platform-Institutional Data Repository-Selected Aspects", *AGRIS on-line Papers in Economics and Informatics*, Vol. 15, No. 4, pp. 127-133. ISSN 1804-1930. DOI 10.7160/aol.2023.150409.

[22] Šimek, P., Vaněk, J., Jarolímek, J., Stočes, M. and Vogeltanzová, T. (2013) "Using metadata formats and AGROVOC thesaurus for data description in the agrarian sector", *Plant, Soil and Environment*, Vol. 59, No. 8, pp. 378-384. E-ISSN 1805-9368. DOI 10.17221/261/2013-PSE.

[23] Wieder, P. and Nolte, H. (2022) "Toward data lakes as central building blocks for data management and analysis", *Frontiers in Big Data*, Vol. 5, p. 945720. E-ISSN 2624-909X. DOI 10.3389/fdata.2022.945720.

[24] Wilson, G., Bryan, J., Cranston, K., Kitzes, J., Nederbragt, L. and Teal, T. K. (2017) "Good enough practices in scientific computing", *PLOS Computational Biology*, Vol. 13, No. 6, p. e1005510. ISSN 1553-7358. DOI 10.1371/journal.pcbi.1005510.

[25] Wiseman, L., Sanderson, J., Zhang, A. and Jakku, E. (2019) "Farmers and their data: An examination of farmers' reluctance to share their data through the lens of the laws impacting smart farming", *NJAS-Wageningen Journal of Life Sciences*, Vol. 90-91, p. 100301. E-ISSN 2212-1307. DOI 10.1016/j.njas.2019.04.007.

[26] Wolfert, S., Ge, L., Verdouw, C. and Bogaardt, M.-J. (2017) "Big Data in Smart Farming-A review", *Agricultural Systems*, Vol. 153, pp. 69-80. ISSN 0308-521X. DOI 10.1016/j.agsy.2017.01.023.

[27] Žáková Kroupová, Z., Aulová, R., Rumánková, L., Bajan, B., Čechura, L., Šimek, P. and Jarolímek, J. (2025) "Drivers and barriers to precision agriculture technology and digitalisation adoption: Meta-analysis of decision choice models", *Precision Agriculture*, Vol. 26, p. 17. E-ISSN 1573-1618. DOI 10.1007/s11119-024-10213-1.

Agris on-line Papers in Economics and Informatics

Volume XVIII Number 3, 2026

Agri25: Value Creation Rating Model for Agricultural Enterprises - Comparison of Logistic and Discriminant Analysis in Czech Conditions

Jindřich Špička[1], Zdeňka Náglová[2]

[1] Department of Statistic, Faculty of Economics and Management, Czech University of Life Sciences Prague, Czech Republic

[2] Department of Economics of Agriculture and Food Industry, Institute of Agricultural Economics and Information, Prague, Czech Republic

Abstract

This study develops a sector-specific value creation rating model for Czech agricultural enterprises using Economic Value Added (EVA) as the classification criterion. Based on accounting data from the CRIBIS database covering period 2019-2023, we compare logistic regression and discriminant analysis for distinguishing value-creating from value-destroying firms. Results demonstrate that logistic regression provides superior predictive accuracy, with efficient return on invested capital and adequate working capital relative to long-term financing emerging as the primary determinants of positive EVA. The findings confirm that traditional rating models developed for industrial enterprises exhibit limited applicability in agriculture due to the sector's unique characteristics, including Common Agricultural Policy (CAP) subsidies that suppress bankruptcy rates and create distinct financial dynamics. This research highlights the necessity of sector-adapted analytical tools that shift focus from bankruptcy prediction to value creation assessment, providing practical frameworks for financial institutions, farm managers, and policy-makers evaluating agricultural financial performance and sustainability.

Keywords

Agriculture, creditworthiness, logistic regression, discriminant analysis, economic value added, financial health, Czech farms.

Špička, J. and Náglová, Z. (2026) "Agri25: Value Creation Rating Model for Agricultural Enterprises - Comparison of Logistic and Discriminant Analysis in Czech Conditions", *AGRIS on-line Papers in Economics and Informatics*, Vol. 18, No. 3, pp. 60-77. ISSN 1804-1930. DOI 10.7160/aol.2026.180306.

Introduction

Predicting financial condition and bankruptcy remains an important topic in finance and risk management. For the agricultural sector, however, the applicability of traditional models remains a matter of controversy. Agriculture is highly influenced by structural, climatic, and policy risks (Dai, 2025), making financial results volatile and often poorly captured by models originally designed for industrial or other types of firms (Wu et al., 2022).

Agricultural enterprises in the European Union exhibit substantially lower bankruptcy risk compared to firms in other sectors due to operational subsidies provided under the Common Agricultural Policy (CAP). These direct payments stabilize farm incomes and create a financial safety net that mitigates insolvency risk, as regular public funding ensures baseline liquidity regardless of market performance (Mazancová et al., 2025). Consequently, traditional bankruptcy prediction models, which are designed to identify firms at risk of failure in competitive market environments, have limited applicability in the agricultural sector. Instead, rating models that assess the overall financial health and operational efficiency of farms provide more meaningful analytical insights for EU agricultural holdings.

The primary objective of this study is to develop and validate a sector-specific value creation rating model for Czech agricultural enterprises that effectively distinguishes between value-creating and value-destroying firms. To achieve this overarching goal, the following specific objectives are pursued:

- To construct a comprehensive dataset of Czech agricultural enterprises using accounting data from the CRIBIS database for the period 2019-2023, ensuring adequate representation of the sector's financial characteristics.
- To implement Economic Value Added (EVA) as the key performance criterion for classifying enterprises, thereby shifting focus from bankruptcy prediction to economic value creation assessment.
- To develop and estimate both logistic regression and discriminant analysis models using relevant financial ratios that capture profitability, liquidity, capital structure, and operational efficiency.
- To systematically compare the predictive accuracy and reliability of these two statistical approaches through appropriate validation techniques.
- To identify the most significant financial determinants that differentiate value-creating from value-destroying agricultural enterprises in the Czech context. Finally, to evaluate the practical applicability of the developed model for financial institutions, agricultural policy-makers, and farm managers in assessing financial performance and sustainability of agricultural operations.

Literature review

This literature review synthesizes key approaches to bankruptcy prediction and credit rating assessment, with a focus on European agriculture, and highlights the limitations of existing models when applied to Czech agricultural enterprises.

To traditional bankruptcy models applied to agriculture belongs Altman's Z-score and its variants. The Altman's Z-score (1968) and its subsequent modifications have been widely applied in agriculture. Novotná and Svoboda (2010) tested Altman's model alongside Czech-specific indices (IN95, IN99, IN01), the Kralicek quick test, and the Bonity index on 150 Czech farms between 2003 and 2007, differentiated by less-favored area (LFA) classifications. Their findings showed inconsistencies, as profitable farms were often misclassified into the grey zone due to the mismatch between model assumptions and agricultural realities. Similarly, Baryshnikov et al. (2019) confirmed that Altman's model does not provide sufficient assessment for agricultural enterprises.

In the international literature, Altman's model has also been applied by Milić et al. (2022) in Serbian agriculture and food-processing enterprises of varying sizes, by Srebro et al. (2021) on Serbian farms, and by Kiaupaite-Grushniene (2016) in Lithuania. Results range from partial usefulness to serious misclassification problems. This underlines the fact that the Z-score, though historically important, cannot be reliably generalized to agriculture.

The Bonity index (Grünwald, 2001) and Kralicek's quick test are popular in Czech and Slovak literature. Their applications to agriculture (Novotná and Svoboda, 2010; Vavrek et al., 2021a) have revealed mixed results. Vavrek et al. (2021) applied Altman's, Taffler's, and Bonity indices to Slovak agricultural enterprises and showed that each model produced divergent classifications—Altman pointed to future risk, Bonity was neutral, and Taffler suggested no near-term threats.

In the Czech and Slovak Republic, there were also developer-specific models. Neumaier and Neumaierová (2002, 2005) developed the IN indices, which were tailored for Czech firms (mostly industrial companies) in transition economies. The progression from IN95 to IN05 sought to improve accuracy and adaptability. Applications in agriculture, however, remain a subject of controversy. For example, Kopta (2009) tested IN99 alongside Gurčík's index and others on Czech agricultural enterprises (between 1996 and 2006) and concluded that while there is a statistical relationship between index values and future performance, predictive power for bankruptcy was limited.

Gurčík (2002) developed a proprietary discriminant index specifically for Slovak agricultural enterprises, based on a sample of 60 farms. The model was intended not only for bankruptcy prediction but also for assessing future growth potential. Similarly, Chrastinová (1998) introduced the Ch-index as a tool for predicting bankruptcy. Both have been tested in Czech contexts (Kopta, 2009), where they are often classified into the "grey zone" with limited discriminatory power.

Špička et al. (2018) criticized Gurčík's index for overemphasizing profitability (four out of five indicators), while neglecting liquidity, capital structure, and working capital management.

Likewise, the Ch-index was found to misclassify many enterprises into neutral categories, limiting its practical usability.

Karas et al. (2017) tested Altman's model and IN05 on Czech active companies and bankrupt firms (2011-2014). They concluded that traditional models developed for industrial and processing enterprises demonstrate lower predictive accuracy when applied to agriculture. This confirms the need for sector-specific models.

Recent advances suggest that statistical and AI-based methods may outperform traditional indices. Gajdosikova et al. (2025) employed logistic regression, decision trees, machine learning, and artificial neural networks across farms in Central and Eastern Europe. These methods demonstrated high predictive accuracy but also revealed strong heterogeneity across economies, emphasizing the importance of localized models and continuous monitoring of financial risks.

There are also studies beyond Central Europe, for example, from Ukraine, where Dorohan-Pysarenko (2021) combined discriminant analysis with expert scoring to include non-financial risk factors such as environmental threats. Tyshchenko et al. (2023) tested several models (Altman, Springate, Tereshchenko, Matviychuk) on Ukrainian farms, with mixed outcomes—some models successfully flagged bankruptcies but also produced many false positives. These findings align with the central and eastern European experience: the structural volatility of agriculture reduces the transferability of generic models.

Meanwhile, alternative approaches, such as Data Envelopment Analysis (DEA), were explored by Horváthová and Mokrišová (2018) in the tourism sector, showing potential as a methodological alternative that is also adaptable to agriculture.

Across the reviewed studies, one consistent observation emerges: traditional bankruptcy models (Altman, Taffler, Bonity, IN) provide non-uniform and often contradictory results when applied to agricultural enterprises. Profitable farms are often categorized as being at risk (the grey zone), while long-term distressed firms may escape early warning signs.

The evidence suggests that no single existing model—whether Altman's Z-score, Czech IN indices, or Slovak Gurčík and Ch-index —adequately captures the financial risk of agricultural enterprises in the Czech Republic. Results are often contradictory, with many firms categorized in the grey zone, regardless of their actual performance. Comparative studies confirm that industrial-based models have lower predictive validity in agriculture.

Over the past decades, a wide spectrum of analytical, statistical, and machine learning techniques has been developed to predict financial distress and corporate bankruptcy. Early studies relied primarily on traditional statistical models, which laid the foundation for understanding the relationship between financial ratios and business failure. Beaver (1966) introduced univariate ratio analysis as one of the first quantitative approaches to assess financial distress, while Altman (1968) or McKee (1976) expanded this concept through multivariate discriminant analysis (MDA). These models were simple and interpretable but limited by their assumption of linearity and inability to capture.

Among the early approaches stated above, McKee (2000) provided additional methods for assessing and predicting financial distress, including subsequent advancements such as linear probability models (Meyer and Pifer, 1970) and multivariate conditional probability models, such as logit and probit frameworks (Ohlson, 1980). Other methodological innovations encompass recursive partitioning models (Marais et al., 1984; Frydman et al., 1985), survival analysis using proportional hazard models (Lane et al., 1986), expert systems (Messier and Hansen, 1988), and mathematical programming approaches (Gupta et al., 1990). More recent developments have incorporated artificial intelligence techniques, including neural networks (Tam and Kiang, 1992; Bell et al., 1990), as well as the rough sets approach (Slowinski and Zopoudnis, 1995). These methods collectively represent a diverse set of quantitative tools for evaluating corporate financial health and predicting potential financial distress.

As computational technologies evolved, traditional statistical methods such as discriminant analysis and univariate models were gradually replaced by data-driven algorithms capable of uncovering complex patterns in financial data—a transformation often referred to as the advent of data mining (Papíková and Papík, 2022). Techniques such as support vector machines (SVM), decision trees (DT), and random forests (RF) have demonstrated an improved ability to capture nonlinear relationships (Xie et al., 2011; Geng et al., 2015), although challenges related to overfitting and interpretability persist.

The use of neural networks (NNs) further enhances predictive performance, allowing for nonlinear modelling of financial variables; however, issues with class imbalance remain (Tumpach et al., 2020). More recently, ensemble learning and hybrid artificial intelligence methods —including AdaBoost, XGBoost, and CatBoost —have become state-of-the-art, achieving outstanding accuracy and robustness even in imbalanced datasets (Lahmiri et al., 2020; Jabeur et al., 2021). Among these, CatBoost has proven particularly effective for small and medium-sized enterprises, reaching an AUC close to 99% (Papíková and Papík, 2022).

This motivates the development of a new model for predicting financial situations tailored to Czech agriculture on a larger scale, encompassing multiple farms. Such a model should combine the strengths of existing indices (profitability and solvency) with additional sector-specific indicators (liquidity, subsidies, working capital, and market volatility). Furthermore, it should strike a balance between statistical accuracy and practical transparency, providing a reliable decision-support tool for farmers, policymakers, and creditors alike.

Research gap

Previous studies have predominantly focused on bankruptcy prediction rather than comprehensive value creation rating model, which is more relevant for evaluating the overall financial health and operational efficiency of subsidized agricultural holdings. The Czech agricultural sector lacks a sector-specific model that accounts for its unique characteristics and employs appropriate performance criteria beyond mere bankruptcy risk. While recent advances in statistical and artificial intelligence methods show promise, there is a lack of research comparing traditional statistical approaches in the specific context of Czech agriculture using value creation metrics. This study addresses these gaps by developing and validating a predictive model specifically tailored to Czech agricultural enterprises, employing EVA as the classification criterion rather than bankruptcy, and systematically comparing logistic regression with discriminant analysis to identify the most reliable methodological approach for this sector.

Materials and methods

Materials

Data cleaning

The accounting data of agricultural enterprises were obtained from the financial statements in the CRIBIS database (2025). The CRIBIS database, developed by CRIF-Czech Credit Bureau, a.s., is a comprehensive information system that contains detailed financial, credit, and identification data on businesses operating in the Czech Republic and abroad. It integrates data from public registers, financial statements, and payment behaviour analyses to assess the risk profile of enterprises. The data were processed and cleaned prior to the application of statistical methods in the following manner. The initial dataset contained agricultural enterprises with reported land area in the Land Parcel Identification System (LPIS). The cleaning procedure began by removing observations with missing values for the key variable LAND_AREA, ensuring that only complete cases with valid agricultural area measurements were retained for analysis.

Subsequently, a comprehensive filtering process was implemented to eliminate nonsensical or economically implausible values that could distort the statistical analysis. This involved removing enterprises with non-positive values for critical financial statement items, as negative values for these variables would indicate data entry errors or reporting inconsistencies. Specifically, observations were excluded if they reported non-positive values for total liabilities, fixed assets, tangible fixed assets, current assets, inventories, short-term receivables, cash and cash equivalents, external financing sources, short-term liabilities, product sales revenue, and material and energy consumption.

Additional filters were applied to remove observations with negative values for long-term receivables, loans and credits, long-term liabilities, depreciation, netbook value of assets, and interest expenses, as these negative values would be economically meaningless. Enterprises with missing registered capital were also excluded from the analysis. Importantly, over-indebted firms with non-positive equity were removed, as assessing the financial health of such enterprises would be economically irrelevant (it biases profitability ratios).

Following the initial data cleaning, twenty-seven financial health indicators were calculated covering four main categories: profitability ratios (including return on assets, return on equity, return on sales, return on fixed assets, return on capital employed, and return on invested capital), activity ratios (such as total asset turnover, fixed asset turnover, inventory turnover, and receivables turnover), liquidity ratios (comprising current, quick, and cash ratios, along with working capital ratios), and leverage ratios (including various debt ratios and coverage ratios).

To address the presence of extreme values that could unduly influence the statistical analysis, a trimming approach was employed (Dehnel, 2014). The first and ninety-ninth percentiles were calculated for each financial indicator, and observations falling outside this range were systematically removed from the dataset. This trimming procedure helps ensure that the subsequent statistical tests are not dominated by outliers while preserving the central distribution characteristics of the financial health indicators.

This multi-stage cleaning process yielded a refined dataset comprising only enterprises with economically plausible financial data and indicator values within reasonable bounds, thereby providing a solid foundation for the subsequent statistical analysis of the financial health of agricultural enterprises.

Calculation of Economic Value Added (EVA)

The value creation is defined as positive EVA for the purpose of this research. The EVA methodology was employed to assess whether agricultural enterprises generated value beyond the expectations of equity investors. This approach compares the actual Return on Equity (ROE) achieved by each enterprise with the expected return on equity derived from the Capital Asset Pricing Model (CAPM) (Elbannan, 2015).

The calculation of EVA involved a two-step process. First, the ROE was computed for each agricultural enterprise using the standard formula: *ROE = Net Income / Shareholders' Equity* × *100*. This metric represents the actual return generated on shareholders' invested capital during the reporting period.

Second, the expected return on equity (r_e) was determined using the CAPM framework, which establishes the minimum return that equity investors should require given the systematic risk associated with the investment. The CAPM-based expected return on equity was calculated using several key parameters obtained from Professor Damodaran's database[1]. The risk-free rate (r_f) served as the baseline return for risk-free investments. The unlevered beta captures the systematic risk of the agricultural sector, excluding the influence of financial leverage. The effective tax rate (t) and debt-to-equity ratio (D/E) were used to calculate the levered beta (beta levered), which reflects the additional financial risk imposed by the enterprise's capital structure through formula: *beta levered = beta unlevered* × [1 + (1-*t*) × *D/E*]. The Czech Republic equity risk premium (Total Equity Risk Premium) represented the additional return required by investors for bearing market risk above the risk-free rate. Finally, the expected return on equity (r_e) was computed using the CAPM equation: $r_e = r_f$ + *beta levered* × *Total Equity Risk Premium*. Table 1 provides an overview of re computed in this study.

Indicator	2019	2020	2021	2022	2023
r_e	4.74 %	3.87 %	5.46 %	10.42 %	9.06 %

Source: Own processing

Table 1: Expected Return on Equity (r_e) in Czech agriculture based on the CAPM framework.

The EVA calculation then compared these two measures: *EVA* = (*Actual ROE* - r_e) × *Shareholders' Equity*. A positive EVA indicates that the enterprise generated returns above investor expectations, thereby creating economic value, while a zero or negative EVA suggests value destruction, as the enterprise failed to meet the minimum return requirements dictated by its systematic risk profile.

This methodology provides a more sophisticated assessment of enterprise performance than traditional accounting-based profitability measures, as it explicitly accounts for the cost of equity capital and the risk-return relationship inherent in financial markets. By incorporating market-based expectations through the CAPM framework, the EVA approach offers insights into whether agricultural enterprises are generating genuine economic value for their shareholders beyond mere accounting profits.

An approach based on the EVA indicator was therefore adopted, and the choice of the value creation rating model was made accordingly. This decision was also motivated by the fact that, in the agricultural sector (Agriculture, Forestry and Fishing, NACE A), the bankruptcy rate —defined as the number of bankruptcies per year relative to the number of economically active entities—is very low compared to other industries. According to the study conducted by Insolcentrum (n.d.), the bankruptcy rate in Czech agriculture, forestry, and fishing was 2 % in 2021. Consequently, the sample of bankrupt enterprises would be very limited.

[1] https://pages.stern.nyu.edu/~adamodar/

Data description

Our analysis covers the accounting years 2019-2023, as the database used contains the necessary accounting data. One-fifth of the enterprises in the panel data specialize in crop production, one-fifth specialize in livestock production, and more than half engage in a combination of crop and livestock production (Table 2). The average farm size in the sample is 1,062 hectares, with the smallest enterprises having zero area and the largest enterprise having more than 10,000 hectares (Table 3).

Type of farming	Relative frequency (%)
Specialized crop production	20.6 %
Specialized livestock production	21.1 %
Mixed crop and livestock production	58.2 %

Source: Own processing

Table 2: Distribution by type of farming.

Mean	Standard Deviation	Minimum	Maximum
1,062.08	1,105.42	0	10,160.26

Source: Own processing

Table 3: Distribution by land area (hectares).

Based on the frequency table (Table 4), the results demonstrate the distribution of enterprises achieving positive versus negative Economic Value Added (EVA) across the analyzed sample.

Table 4 shows that, in most years, farms with negative EVA (denoted as EVA-) predominated. A combination of rapidly increasing capital costs (due to high interest rates of the Czech National Bank resulting from inflation), a sharp rise in operating expenses, and a simultaneous decline in revenues created a situation in which operating profit was insufficient to cover capital costs. This led to negative EVA in a substantial proportion of agricultural enterprises in 2022 and 2023.

These findings suggest that a substantial number of agricultural enterprises in the sample failed to meet the minimum return expectations of equity investors, as determined by the CAPM model. The enterprises with zero or negative EVA values (EVA-) represent firms that, despite potentially generating accounting profits, failed to provide returns sufficient to compensate shareholders for the systematic risk associated with their investments. Conversely, the enterprises achieving positive EVA (EVA+) successfully created economic value by generating returns above the risk-adjusted cost of equity capital. These firms demonstrated superior performance in converting shareholder investments into returns that exceeded market expectations, given their respective risk profiles.

The substantial presence of both value-creating and value-destroying enterprises within the agricultural sector suggests significant heterogeneity in financial performance and management effectiveness. This distribution pattern indicates that while some agricultural enterprises effectively manage their resources to generate superior returns, a notable portion struggles to meet the fundamental requirement of covering their cost of equity capital, highlighting potential inefficiencies or structural challenges within parts of the agricultural sector.

Methods

Comparison of discriminant analysis and logistic regression

In the development of agricultural financial performance models, researchers frequently encounter the need to classify enterprises into distinct performance categories based on multiple financial indicators. Two statistical methodologies that have gained considerable attention for such classification tasks are discriminant analysis (e.g., Neumaier and Neumaierová, 2005) and logistic regression (e.g. Ohlson, 1980; Zmijewski, 1984). While both techniques serve the fundamental purpose of predicting group membership based on predictor variables, they differ substantially in their underlying assumptions, mathematical frameworks, and practical applications within the field of agricultural economics research.

Indicator	Unit	2019	2020	2021	2022	2023
Sample size	number	1260	1340	1353	1325	983
Share of companies with EVA+	%	43.17	54.63	47.60	35.62	18.31
Share of companies with EVA-	%	56.83	45.37	52.40	64.38	81.69

Source: Own processing

Table 4: Sample size by year and EVA.

Theoretical foundations

Discriminant analysis operates under the premise of identifying linear combinations of predictor variables that maximize the separation between predefined groups. In the context of agricultural financial performance modeling, this technique assumes that the predictor variables follow a multivariate normal distribution within each performance category and that the covariance matrices across groups are homogeneous. The method seeks to construct discriminant functions that optimize the ratio of between-group variance to within-group variance, thereby creating decision boundaries that effectively distinguish between different levels of financial performance (Hatcher, 2013).

Logistic regression, by contrast, employs a probabilistic approach that models the log-odds of group membership as a linear function of predictor variables. This methodology does not require the stringent distributional assumptions inherent in discriminant analysis, making it particularly appealing for agricultural data that often violate normality assumptions due to seasonal variations, extreme weather events, or market volatilities. The logistic function ensures that predicted probabilities remain bounded between zero and one, providing intuitive interpretations of the likelihood that a particular agricultural enterprise belongs to a specific performance category (Pampel, 2011).

Assumptions and data requirements

The assumption structure represents perhaps the most critical distinction between these methodologies. Discriminant analysis requires that predictor variables exhibit multivariate normality within each group, expressed formally as:

$$x_i|G = g \sim N(\mu_g, \sum) \text{ for all } g = 1,2,\ldots,G \tag{1}$$

This condition may prove challenging in agricultural applications where financial ratios often display skewed distributions or contain outliers resulting from extraordinary circumstances such as crop failures or commodity price spikes. Additionally, the homoscedasticity assumption demands equal covariance matrices across performance groups:

$$\sum_1 = \sum_2 = \cdots = \sum_G = \sum \tag{2}$$

This equality may not hold when comparing small-scale family farms with large agricultural corporations, as the variance structures of financial indicators may differ systematically across farm types (Klecka, 1980).

Logistic regression demonstrates considerably more flexibility in its assumption requirements. The technique does not mandate specific distributional forms for predictor variables, though it does assume a linear relationship between predictors and the log-odds of the outcome variable:

$$E\left[\ln\left(\frac{P(Y=1|x}{1-P(Y=1|x}\right)\right] = \beta_0 + \sum_{i=1}^{p} \beta_i x_i \tag{3}$$

This linearity assumption can be validated through various diagnostic procedures and, if violated, can often be addressed through variable transformations or the inclusion of polynomial terms. The logit model also assumes independence of observations:

$$P(Y_1, Y_2, \ldots, Y_n|X) = \prod_{i=1}^{n} P(Y_i|x_i) \tag{4}$$

Furthermore, logistic regression accommodates both continuous and categorical predictor variables without additional modifications, a feature particularly valuable in agricultural research where factors such as farm type, geographic region, or crop diversity may serve as important classification variables (Timming, 2022).

Model estimation and implementation

The estimation procedures for these methodologies reflect their different theoretical orientations. Discriminant analysis typically employs maximum likelihood estimation under the assumption of multivariate normality, calculating sample means and pooled covariance matrices to construct discriminant functions. The classification rule assigns observations to the group with the smallest Mahalanobis distance or, equivalently, the highest posterior probability based on Bayes' theorem (Hatcher, 2013).

Logistic regression utilizes iterative maximum likelihood estimation, as closed-form solutions generally do not exist. The iterative reweighted least squares algorithm or similar optimization procedures are employed to determine parameter estimates that maximize the likelihood function. This process continues until convergence criteria are satisfied, typically involving minimal changes in parameter estimates or log-likelihood values between iterations (Pampel, 2011).

Performance evaluation and diagnostic capabilities

In evaluating the quality of a predictive model that classifies firms into value-creating and value-destroying categories, we employ concordant and discordant pairs. The fundamental concept involves comparing all possible pairs of firms in which one firm belongs to the value-creating category (EVA+; denoted by the value 1) and the other to the value-destroying category (EVA-; denoted by the value 0). For each such pair, we then examine whether the model assigned a higher predicted probability to the correct firm. Specifically, we refer to a concordant pair when the model assigns a higher probability to the firm that is actually value-creating. Conversely, a discordant pair arises when the model assigns a higher probability to the firm that is actually value-destroying. In the case of discriminant analysis, rather than comparing probabilities, we work with discriminant scores, or values of the discriminant function (Pampel, 2011). Discriminant analysis calculates a discriminant score for each firm, which expresses its position along the discriminant axis. This score determines the group to which the firm is classified—higher values typically indicate membership in the value-creating category, while lower values indicate membership in the value-destroying category. The score itself is not a probability in the classical sense but serves as a measure of the firm's "distance" from the centroids of the respective groups (Brown and Wicker, 2000).

Key model evaluation metrics are derived from the ratio of concordant and discordant pairs. The most important of these is the C-statistic, which is calculated as the ratio of concordant pairs (plus half of the tied pairs) to the total number of pairs. A pair is concordant if the model assigns a higher predicted probability (or discriminant score) to the firm that is actually profitable, and discordant otherwise. The C-statistic (or concordance index) is calculated as the share of concordant pairs (plus half of tied pairs) relative to all possible pairs, with values ranging from 0.5 (random prediction) to 1.0 (perfect prediction).

Both methodologies provide comprehensive frameworks for model evaluation, although their diagnostic capabilities differ in significant ways. Discriminant analysis provides canonical correlation coefficients that indicate the strength of association between discriminant functions and group membership, while Wilks' lambda tests assess the statistical significance of group differences. The technique also generates structure coefficients that reveal the contribution of individual variables to group separation (Brown and Wicker, 2000).

Logistic regression provides a comprehensive suite of diagnostic tools, including goodness-of-fit statistics such as the Hosmer-Lemeshow test, pseudo R-squared measures for evaluating model explanatory power, and residual analyses for identifying influential observations or model inadequacies. The availability of odds ratios provides direct interpretation of the multiplicative effect of predictor variables on the odds of group membership, facilitating communication of results to agricultural practitioners and policymakers (Pampel, 2011).

Practical considerations for agricultural applications

The choice between discriminant analysis and logistic regression in modeling agricultural financial performance often depends on the specific characteristics of the dataset and research objectives. Discriminant analysis may prove advantageous when predictor variables approximate multivariate normality and when the primary goal is to understand the underlying structure that distinguishes between performance groups. The technique's ability to generate multiple discriminant functions can provide valuable insights into the dimensionality of group separation in complex agricultural systems.

Logistic regression is typically the preferred approach when dealing with non-normal data distributions, unequal group sizes, or when the research emphasis is on prediction accuracy rather than understanding group structure. Its robustness to assumption violations and flexibility in handling various data types make it particularly suitable for agricultural datasets that incorporate diverse financial metrics, operational characteristics, and environmental factors.

Limitations and considerations

Despite their widespread application, both methodologies possess inherent limitations that researchers must acknowledge. Discriminant analysis's sensitivity to outliers can prove problematic in agricultural contexts where extreme events may generate atypical financial observations. The technique's performance may deteriorate significantly when sample sizes are small relative to the number of predictor variables, a common scenario in specialized agricultural sectors (Klecka, 1980).

Logistic regression, while more robust to distributional violations, can experience difficulties with complete or quasi-complete separation, particularly when dealing with perfectly predictive variables or small sample sizes. Additionally, the technique's reliance on large-sample theory for statistical inference may prove inadequate in studies with limited observations, necessitating alternative approaches such as exact logistic regression or bootstrap procedures (Pampel, 2011).

Results and discussion

Statistically significant predictors of firm profitability in at least one of the analyzed years are shown in Tables 5 (Logistic Regression) and 6 (Discriminant Analysis). A positive sign indicates a direct relationship, while a negative sign denotes an inverse effect of the given financial indicator on EVA. To ensure the temporal stability of the models, predictors were selected based on their frequent occurrence across the observation period and low correlation with other selected variables.

According to logistic regression results,

Indicator	Symbol	2019	2020	2021	2022	2023	Note
Return on Assets	ROA	+				-	
Return on Tangible Assets	ROTA					+	
Return on Capital Employed	ROCE		+	+	+	+	
Return on Invested Capital	ROIC	+	+	+	+	+	Selected
Cash Ratio	CR					-	
Working Capital to Long-term Capital	WCLTC	-			+	+	Selected
Working Capital to Assets	WCA	+			-	-	
Net Debt Payback Period	NDPP	-			-		
Equity Ratio	ER	-			-		
Debt to Equity Ratio	DER				+	+	
Short-term Debt Ratio	SDR				-		
Acid-Test Ratio	ATR	-					
Total Debt Ratio	TDR					-	
Fixed Assets Coverage	FAC	-	-			-	Selected
Debt to Sales Ratio	DSR	+					
Fixed Assets Turnover	FAT					-	

Note: "+" = direct significant relationship (significance level 0.01), "-" inverse significant relationship (significance level 0.01)
Source: Own processing

Table 5: Statistically significant predictors - logistic regression.

Indicator	Symbol	2019	2020	2021	2022	2023	Note
Return on Assets	ROA	+	+	+	+	+	Selected
Return on Sales	ROS	+	+	+			
Return on Tangible Assets	ROTA	-	-	-	-		
Return on Invested Capital	ROIC	+			+	+	
Working Capital to Sales Ratio	WCS				+		
Net Debt Payback Period	NDPP	-	-	-	-	-	Selected
Equity Ratio	ER	-			-	-	Selected
Debt to Equity Ratio	DER			+		+	
Total Debt Ratio	TDR			+			
Loan Burden Coverage	LBC				+	+	
Long-term Debt Ratio	LDR		+				
Interest Coverage Ratio	ICR			+			
Fixed Assets Turnover	FAT		+	+			

Note: "+" = direct significant relationship (significance level 0.01), "-" inverse significant relationship (significance level 0.01)
Source: Own processing

Table 6: Statistically significant predictors - discriminant analysis.

the statistically significant predictors of firm profitability are Return on Invested Capital (ROIC), Working Capital to Long-term Capital (WCLTC), and Fixed Assets Coverage (FAC). The resulting equation for the entire period is as follows:

$$Agri25L = -1.828 + 0.5198\,ROIC + 0.0087\,WCLTC - 0.0113\,FAC \quad (5)$$

Return on Invested Capital (ROIC) = Net Profit / (Fixed Assets + Non-Cash Net Working Capital) * 100

Working Capital to Long-term Capital (WCLTC) = Net Working Capital / (Equity + Long-term Liabilities) * 100

Fixed Assets Coverage (FAC) = Equity / Fixed Assets * 100

A Max-rescaled R-Square value of 0.6542 (Nagelkerke R-Square) means that your model explains approximately 65% of the variability in the data (according to this metric). This is typically considered a reasonably good model fit.

The Return on Invested Capital (ROIC) exhibits the strongest positive effect on EVA, with a coefficient of 0.5198, making it the dominant factor in the predictive model. This result aligns with fundamental principles of financial management—firms that more efficiently transform invested capital into profit generate greater value for their owners and exhibit a higher capacity to produce cash flows (Brealey et al., 2025). A high ROIC reflects not only operational efficiency but also the quality of managerial investment decisions. Firms with higher ROIC are better able to cover capital costs, contributing to long-term financial stability and lowering the likelihood of financial distress (Gibson, 2013).

Working Capital to Long-term Capital (WCLTC) also exerts a positive, though substantially weaker, influence on EVA (coefficient 0.0087). This indicator reflects a firm's liquidity relative to its long-term financing sources (Gibson, 2013). The positive relationship can be interpreted as the importance of maintaining sufficient financial flexibility—firms with higher net working capital possess greater operational maneuverability and are better prepared to absorb short-term financial fluctuations. However, the small coefficient value suggests that, while liquidity contributes to profitability, its impact is secondary to investment returns. Excessively conservative working capital management may even lead to suboptimal resource utilization and reduced overall profitability.

Fixed Assets Coverage (FAC) exhibits a negative relationship with EVA (coefficient -0.0113). This implies that enterprises with extremely high fixed asset coverage by equity may paradoxically be less value-creating. This can be explained by several mechanisms. First, an overly conservative capital structure, with minimal debt utilization, forgoes the tax benefits of debt and reduces the return on equity through the loss of leverage effects. Second, excessive reliance on equity may indicate limited access to external financing, signaling lower perceived creditworthiness (Brealey et al., 2025). A very high FAC may thus reflect insufficient expansion and investment activity, where firms accumulate equity without utilizing growth opportunities. The low absolute coefficient value, however, suggests that this effect is relatively moderate and likely occurs only in cases of excessive capitalization.

Descriptive statistics of the variables included in equation (5) and the resulting Agri25L profitability scores (Table 7) were used to determine the boundary of the value-creating zone, defined by the 5th and 95th percentiles.

Based on the logistic regression results, the boundary of the value-creating zone was set at Agri25L > 0.707 (the 95th percentile of Agri25L for firms with negative EVA). The boundary

Indicator	N	Mean	Std Dev	5th pctl	95th pctl
ROIC	6261	5.622	10.535	-5.403	23.075
WCLTC	6261	22.248	34.962	-26.155	66.431
FAC	6261	121.069	97.081	28.100	272.501
Agri25L	6261	-0.089	5.234	-6.113	8.515
Agri25L (EVA-)	3689	-2.541	2.861	-7.500	0.707
Agri25L (EVA+)	2572	3.427	5.829	-1.008	13.834

Source: Own processing

Table 7: Descriptive statistics of indicators in the logistic regression (2019 - 2023).

of the value-destroying zone was set at Agri25L < -1.008 (5th percentile of Agri25L for firms with positive EVA). Values between -1.008 and 0.707 represent the "gray zone," where the model cannot reliably classify firms.

According to discriminant analysis, the statistically significant predictors are ROA, NDPP, and ER. The resulting equation for the entire period is:

$$Agri25D = 0.4162 + 0.0515\ ROA - 0.0044\ NDPP - 0.0042\ \mathrm{ER} \qquad (6)$$

Return on Assets (ROA) = Operating Profit / Total Assets * 100

Net Debt Payback Period (NDPP) = Net debt / (Net Profit + Depreciation)

Equity Ratio (ER) = Equity / Total Assets * 100

The discriminant function demonstrated statistically significant separation between groups (Wilks' $\lambda = 0.560$, $F(3, 6257) = 1636.0$, $p < 0.0001$). The eigenvalue of 0.784 indicated a moderate to good ratio of between-group to within-group variance. The canonical correlation of 0.663 revealed that the discriminant function accounted for approximately 44% of the variance in group membership (canonical $R^2 = 0.44$), representing a medium to strong effect size.

These findings indicate that the groups are meaningfully distinguishable based on the predictor variables included in the model. The discriminant function demonstrates adequate discriminatory power, though 56% of the variance in group classification remains unexplained. The highly significant F-statistic, combined with the substantial sample size, provides strong evidence for reliable group separation.

Return on Assets (ROA) demonstrates a positive effect on EVA (coefficient 0.0515), making it the most significant discriminant factor. This finding reflects the fundamental principle that a firm's ability to generate operating profit from its total assets is a key indicator of performance and financial stability (Gibson, 2013). Unlike equity-based measures, ROA provides a more comprehensive view of the efficiency of all resources, regardless of financing structure. Higher ROA indicates both operational effectiveness and competitiveness. Firms with higher ROA can better cover financial costs, reinvest in growth, and build reserves to withstand downturns.

Net Debt Payback Period (NDPP) has a negative relationship with EVA (coefficient -0.0044). This metric measures how long it would take a firm to repay its net debt from internally generated funds (net profit and depreciation). A longer payback period indicates higher financial burden and lower debt-servicing capacity, signaling over-leverage and elevated refinancing risk.

Equity Ratio (ER) also shows a negative impact on EVA (coefficient -0.0042). A higher share of equity, typically seen as a sign of stability, here correlates with lower profitability. This can result from an overly conservative capital structure with limited leverage, which reduces return on equity and overall capital efficiency (Brealey et al., 2025). A very high ER may also indicate restricted access to external financing or a cautious strategy that limits growth opportunities and tax advantages from debt use. This effect mirrors that of FAC in the logistic regression model.

Descriptive statistics of the variables used in equation (6) and the resulting Agri25D scores (Table 8) were used to determine the boundary of the value-creating zone, defined by the 5th and 95th percentiles.

In all analyzed years, both models effectively distinguished between value-creating and value-destroying firms, as the Cohen's d values were well above 0.8, indicating a large effect size.

To further assess model accuracy, we compared the effectiveness of each model in classifying firms into their correct zones (value-creating or value-

	Logistic Regression (Agri25L)				Discriminant Analysis (Agri25D)			
Year	Mean(0)	Mean(1)	Std Dev	Cohen's d	Mean(0)	Mean(1)	Std Dev	Cohen's d
2019	-2.9682	2.5952	4.0424	1.376262	0.1891	0.5980	0.2189	1.867976
2020	-3.054	2.4009	4.2292	1.289818	0.1763	0.5798	0.2352	1.715561
2021	-2.5618	3.2756	4.6169	1.264355	0.2164	0.6656	0.2368	1.896959
2022	-1.4451	5.5581	4.2773	1.637295	0.3285	0.8535	0.2383	2.203105
2023	-2.9255	5.2232	3.9845	2.04510	0.2156	0.7856	0.2245	2.53898

Source: Own processing

Table 9: Practical significance of classification by Cohen's d.

destroying). This was evaluated using concordant and discordant pairs (Table 10).

Logistic Regression		Discriminant Analysis	
Percent Concordant	94.4 %	Percent Concordant	84.0 %
Percent Discordant	5.6 %	Percent Discordant	16.0 %
C	0.944	C	0.840

Source: own processing

Table 10: Comparison of the predictive power of logistic regression and discriminant analysis models

The comparison indicates that both models exhibit high predictive accuracy, correctly classifying over 80% of cases. Logistic regression outperformed discriminant analysis.

The superior performance of logistic regression stems primarily from the differing methodological assumptions underlying the two approaches. Discriminant analysis imposes strict requirements on the data—multivariate normality of explanatory variables and equality of covariance matrices across groups (Klecka, 1980). These conditions are rarely satisfied in corporate financial data, which typically exhibit skewness, contain outliers, and display heterogeneous characteristics between financially sound and distressed firms. This is corroborated by Morris (1998), who demonstrates that financial ratios calculated from accounting statements tend to follow skewed distributions, with the dispersion of values around the mean varying substantially across industries. Furthermore, the practice of maintaining balanced samples of healthy and bankrupt firms, while common in empirical studies, is not mathematically necessary from a statistical perspective (Rencher and Christensen, 2012). When these parametric assumptions are violated, discriminant analysis experiences a significant deterioration in predictive accuracy. Nevertheless, despite these limitations, discriminant analysis remains the most frequently employed method for constructing bankruptcy prediction models according to Čámská (2014) and De Laurentis et al. (2010).

In contrast, logistic regression is more robust. It does not require normality or homogeneity of variance and naturally constrains extreme values within the (0,1) probability range. It models relationships through the logit transformation, effectively capturing non-linear patterns between predictors and the dependent variable (Pampel, 2011). This flexibility is especially valuable in financial analysis, which often involves heterogeneous enterprises of varying sizes and sectors.

Finally, discriminant analysis assumes that groups are linearly separable in predictor space. If the relationship between financial indicators and EVA involves non-linear effects or interactions that the discriminant function cannot capture, logistic regression will model these relationships more effectively (Hatcher, 2013).

Discussion of results at the national level within the agricultural sector is feasible only through comparison with the IN99 index (Neumaierová and Neumaier, 2002), which is classified as a predictive model based on whether a firm creates value in terms of positive economic value added (EVA), consistent with the approach adopted by the authors. The IN99 index comprises four financial ratios: the total assets to external financing ratio (i.e., total debt ratio), EBIT to assets, revenues to assets, and the current ratio (the proportion of current assets to short-term liabilities). Within the agricultural sector, the profitability indicator (EBIT/assets) and current ratio carry the greatest weight. All indicators demonstrate a positive effect on creditworthiness. When comparing the estimated Agri25L model with IN99, both models incorporate return on capital in narrower and broader conceptions. Agri25L focuses solely on return on investment, whereas IN99 encompasses return on total capital through two distinct indicators. IN99 does not consider the coverage of fixed assets by equity capital; instead, it directly assesses liquidity. Agri25L does not directly measure liquidity; instead, it addresses it indirectly through working capital. In the case of Agri25D, both models utilize the return on assets metric. The Agri25D model includes an equity risk indicator that reduces creditworthiness, while the IN99 model employs total indebtedness, which enhances creditworthiness. Both models incorporate debt servicing capacity: in Agri25D, the debt maturity period has a negative effect on creditworthiness, whereas in IN99, the firm's ability to meet its obligations contributes positively to its creditworthiness.

Additional models, such as IN01 and IN05 (Neumaierová and Neumaier, 2005), are constructed on the same principle as the preceding model, with the addition of the EBIT-to-interest coverage ratio. The developed Agri25 models do not incorporate this variable.

Kralicek's Quick Test (Kralicek, 1991) represents another established creditworthiness assessment tool that evaluates financial health through four key indicators divided into two dimensions:

financial stability (equity ratio and debt repayment period from cash flow) and profitability (cash flow to sales ratio and return on assets). The resulting creditworthiness index combines these dimensions to provide an overall assessment of firm viability. Comparing Kralicek's approach with the Agri25 models reveals both convergence and divergence: both methodologies incorporate return on assets as a profitability measure, and both address capital structure, though from different perspectives —Kralicek through the equity ratio (positively affecting creditworthiness) and Agri25D through equity risk (negatively affecting creditworthiness). A distinctive feature of Kralicek's Quick Test is the explicit incorporation of cash flow metrics, particularly the debt repayment capacity measured through cash flow, which is not directly captured in the Agri25 models. However, debt servicing capability is indirectly reflected in Agri25D through the debt maturity period indicator.

Grünwald (2001) presents a method for evaluating corporate financial health without reference to production orientation (creditworthiness index), which employs six indicators (ROA, ROE, quick ratio, working capital-to-inventory ratio, cash flow-to-external financing ratio, and EBIT-to-interest ratio). Liquidity is assessed indirectly in the Agri models, and debt is also approached differently. The variables defining creditworthiness are thus considerably differentiated, focusing primarily on profitability and debt servicing capacity.

The system of balance sheet analysis developed by Rudolf Doucha (1996)—also not a sector-specific model, but of Czech origin—focuses on creditworthiness assessment. The model incorporates indicators from the domains of leverage, liquidity, activity, and profitability. This balance sheet analysis corresponds to some extent with previous models, utilizing the self-financing coefficient (as does Agri25D), the ratio of output to liabilities (as in IN99), ROE (Grünwald 2001, IN99), and the quick ratio (also employed by Grünwald 2001, and conceptually related to the current ratio in IN99). These models were developed during comparable periods.

Existing models specifically oriented toward agriculture include the Chratinová index and Gurčík's index. However, both indices are bankruptcy prediction models and, therefore, comparison with creditworthiness models is not methodologically appropriate.

Conclusion

This study developed and validated a sector-specific value creation rating model for Czech agricultural enterprises using Economic Value Added (EVA) as the classification criterion. Based on accounting data from 2019 to 2023, the analysis demonstrates that logistic regression provides superior predictive accuracy compared to discriminant analysis in distinguishing between value-creating and value-destroying agricultural firms. The most significant determinants of positive EVA were identified as efficient return on invested capital and adequate working capital relative to long-term financing.

From a methodological perspective, this research contributes to statistical modeling in agricultural finance by demonstrating that the relaxed distributional assumptions of logistic regression enhance model robustness when dealing with heterogeneous agricultural financial data. The adoption of EVA as a classification criterion, rather than bankruptcy status, represents an innovation that better captures the unique risk profile of subsidized agricultural sectors, where bankruptcy rates are artificially suppressed by CAP interventions. The study validates the principle that financial risk assessment frameworks must be adapted to the structural, regulatory, and operational characteristics of the agricultural sector rather than relying on generic industrial models.

The practical implications are substantial for multiple stakeholders. Financial institutions can utilize the model as a decision-support tool for risk assessment and loan approval processes, enabling more accurate evaluation of agricultural borrowers beyond traditional collateral-based approaches. For agricultural enterprises, the model provides a diagnostic framework for self-assessment, enabling farm managers to benchmark their performance and identify areas that require strategic adjustment. Policy-makers can employ the model to evaluate the financial sustainability of agricultural support programs and design targeted interventions for distressed farms. The model's emphasis on value creation aligns with contemporary agricultural policy objectives focused on competitiveness and sustainable growth.

Several limitations must be acknowledged. The analysis is restricted to Czech agricultural enterprises from 2019 to 2023, which limits its generalizability to other countries and may be potentially affected by COVID-19 disruptions. Reliance on commercial databases introduces selection bias, potentially underrepresenting smaller family farms and specific subsectors. The binary classification approach fails to capture the magnitude of value creation, and the model overlooks non-financial factors such as management quality, technological adoption, and environmental practices. Additionally, static financial ratios fail to capture the temporal dynamics and seasonal variations of agriculture.

Future research should address these limitations through cross-country comparative studies to understand how different policy frameworks affect model applicability, expand samples to include underrepresented farm types, and incorporate machine learning techniques to capture non-linear relationships. Developing dynamic models using time-series data would better account for agricultural volatility. Integrating non-financial variables, such as environmental performance and digital adoption, would create more comprehensive assessment frameworks. Finally, longitudinal validation studies tracking predictions against actual performance over extended periods would provide empirical support for continuous model refinement and enhanced reliability of value creation rating model for the agricultural sector.

Acknowledgements

The research was Supported by the Ministry of Agriculture of the Czech Republic, internal research Project (No. 1150/2025) of the Institute of Agricultural Economics and Information.

Corresponding author:
Ing. Zdeňka Náglová, Ph.D.
Institute of Agricultural Economics and Information
Mánesova 75, Prague, Czech Republic
Email: naglova.zdenka@uzei.cz

References

[1] Altman, E. I. (1968) "Financial ratios, discriminant analysis and the prediction of corporate bankruptcy", *Journal of Finance*, Vol., 23, No. 4, pp.589-609. ISSN 00221082. DOI 10.2307/2978933.

[2] Barashnikov, N., Samygin, D. and Murzin, D. (2019) "Forecasting bankruptcy models for agrarian business", *IOP Conference Series: Earth and Environmental Science*, Vol. 274. DOI 10.1088/1755-1315%2F274%2F1%2F012048.

[3] Beaver, W. (1966) "Financial ratios as predictors of failure", *Journal of Accounting Research*, Vol. 4, pp. 71-111. ISSN 00218456. DOI 10.2307/2490171.

[4] Bell, T. B., Ribar, G. S. and Verchio, J. (1990) "Neural nets versus logistic regression", *Proceedings of the 1990 Deloitte & Touche/University of Kansas Symposium on Auditing Problems*, pp. 29-54.

[5] Boďa, M. and Úradníček, V. (2019) "Predicting financial distress of Slovak agricultural enterprises", *Ekonomický časopis*, Vol. 67, No. 4, pp. 426-452. ISSN 2729-7470.

[6] Brealey, R., Myers, S., Allen F., and Edmans, A. (2025) "*Principles of Corporate Finance*", McGraw Hill. ISBN 9781266288548.

[7] Brown, M. T. and Wicker, L. R. (2000) "Discriminant Analysis", In: Tinsley, H. E. A. and Brown, S. D. (eds.) "*Handbook of Applied Multivariate Statistics and Mathematical Modeling"*, Academic Press, pp. 209-235. ISBN 978-0-12-691360-6.

[8] Chrastinová, Z. (1998) "*Metódy hodnotenia ekonomickej bonity a predikcie finančnej situácie poľnohospodárskych podnikov*", VÚEPP: Bratislava, pp. 34. ISBN 80-8058-022-7. (In Slovak).

[9] Cribis. (2025) "Informace o firmách", CRIF - Czech Credit Bureau, a.s. [Online]. Available: https://rejstrik-firem.kurzy.cz/26212242/crif-czech-credit-bureau-as/ [Accessed: March 10, 2025].

[10] Čámská, D. (2014) "Financial distress prediction models with an emphasis on transition economies", University of Economics, Prague, PhD thesis.

[11] Dai, J. (2025) "Impact of climate policy uncertainty on agriculture development: Multidimensional analysis from land use, food structure, and carbon emissions", *Land Degradation & Development*, Vol. 36, No.3, pp. 4545-4561. ISSN 1099-145X. DOI 10.1002/ldr.5652.

[12] De Laurentis, G., Maino, R. and Molteni, L. (2010) "*Developing, Validating and Using Internal Ratings: Methodologies and Case Studies*", 1st ed., John Wiley & Sons: Singapore, pp. 324. ISBN 978-0-470-71149-1. DOI 10.1002/9780470971901.

[13] Dehnel, G. (2014) "Winsorization Methods in Polish Business Survey", *Statistics in Transition New Series*, Vol. 1, pp. 97-110. ISSN 1234-7655. DOI 10.59170/stattrans-2014-007.

[14] Dorohan-Pysarenko, L., Rebilas, R., Yehorova, O., Yasnolob, I., Kononenko, Z. (2021) "Methodological Peculiarities of Probability Estimation of Bankruptcy of Agrarian Enterprises in Ukraine", *Agricultural and Resource Economics*, Vol. 7, No. 2, pp. 20-39. ISSN 2414-584X. DOI /10.51599/are.2021.07.02.02.

[15] Doucha, R. (1996) "*Finanční analýza podniku: praktické aplikace*", 1st ed., Prague: Vox Consult. ISBN 80-902111-2-7. (In Czech).

[16] Elbannan, M. A. (2015) "The Capital Asset Pricing Model: An Overview of the Theory", *International Journal of Economics and Finance*, Vol. 7, No. 1, pp. 216-228. ISSN 1916-9728. DOI 10.5539/ijef.v7n1p216.

[17] Frydman, H., Altman, E. I. and Kao, D. (1985) "Introducing Recursive Partitioning for Financial Classification: The Case of Financial Distress", *Journal of Finance*, Vol. 40, No. 1, pp. 269-291. ISSN 1540-6261. DOI 10.2307/2328060.

[18] Gajdosikova, D., Michulek, J. and Tulyakova, I. (2025) "AI-Based Bankruptcy Prediction for Agricultural Firms in Central and Eastern Europe", *International Journal of Financial Studies*, Vol. 13, No. 3, pp. 133. ISSN 2227-7072. DOI 10.3390/ijfs13030133.

[19] Geng, R., Bose, I. and Chen, X. (2015) "Prediction of financial distress: An empirical study of listed Chinese companies using data mining", *European Journal of Operational Research*, Vol. 241, No. 1, pp. 236-247. ISSN 1872-6860. DOI 10.1016/j.ejor.2014.08.016.

[20] Gibson, C. H. (2013) "*Financial Statement Analysis*", 13th ed., Mason, Ohio: South-Western; Andover: Cengage Learning.

[21] Grünwald, R. (2001) "*Analýza finanční důvěryhodnosti podniku",* 1st ed. Prague: Ekopress. ISBN 80-86119-47-5. (In Czech).

[22] Gurčík, Ľ. (2002) "G-index - the financial situation prognosis method of agricultural enterprises", *Agricultural Economics - Czech*, Vol. 48, No. 8, pp. 373-378. ISSN 1805-9295. DOI 10.17221/5338-AGRICECON (In Slovak).

[23] Gupta, Y. P., Rao, R. P. and Baggi, P. K. (1990) "Linear goal programming as an alternative to multivariate discriminant analysis: a note", *Journal of Business Finance and Accounting*, Vol. 17, No. 4, pp. 593-598. ISSN 1468-5957. DOI 10.1111/j.1468-5957.1990.tb01146.x.

[24] Hatcher, L. (2013) "*Advanced Statistics in Research: Reading, Understanding, and Writing Up Data Analysis Results*", Shadow Finch Media LLC. ISBN 9780985867003.

[25] Horváthová, J. and Mokrišová, M. (2018) "Risk of Bankruptcy, Its Determinants and Models", *Risks*, Vol. 6, No. 4, pp. 117. ISSN 2227-9091. DOI 10.3390/risks6040117.

[26] Insolcentrum. (n.d.) "*Almost half of companies in bankruptcy will not repay creditors a single cent*", [Online]. Available: https://www.insolcentrum.cz/temer-polovina-firem-v-konkurzu-nevrati-veritelum-ani-korunu/ [Accessed: Nov. 22, 2025]. (In Czech).

[27] Jabeur, S. B., Stef, N. and Carmona, P. (2022) "Bankruptcy Prediction using the XGBoost Algorithm and Variable Importance Feature Engineering", *Computational Economics*, Vol. 61, pp. 715-741. ISSN 0927-7099. DOI 10.1007/s10614-021-10227-1.

[28] Karas, M., Režňáková, M. and Pokorný, P. (2017) "Predicting Bankruptcy of Agriculture Companies: Validating Selected Models", *Polish Journal of Management Studies*, Vol. 15, No. 1. ISSN 2081-7452. DOI 10.17512/pjms.2017.15.1.11.

[29] Kiaupaite-Grushniene, V. (2016) "Altman Z-Score Model for Bankruptcy Forecasting of the Listed Lithuanian Agricultural Companies", *ICAAT 2016*, pp. 222-234. ISBN 978-94-6252-261-9. DOI 10.2991/icaat-16.2016.23.

[30] Kislingerová, E. and Neumaierová, I. (2000) "Vybrané příklady firemní výkonnosti podniku", VŠE: Praha. ISBN 80-7079-641-3. (In Czech).

[31] Klecka, W. R. (1980) "*Discriminant Analysis*", Sage University Paper series, No. 19, Beverly Hills and London: SAGE Publications. ISBN 0-8039-1491-1. DOI 10.4135/9781412983938.

[32] Kopta, D. (2009) "Possibilities of financial health indicators used for prediction of future development of Agricultural enterprises", *Agricultural Economics - Czech*, Vol. 55, No. 3, pp. 111-125. ISSN 1805-9295. DOI 10.17221/589-AGRICECON.

[33] Kralicek, P. (1991) "*Grundlagen der Finanzwirtschaft*", Wien: Ueberreuter. ISBN 3-8000-3415-8.

[34] Lahmiri, S., Bekiros, S., Giakoumelou, A. and Bezzina, F. (2020) "Performance assessment of ensemble learning systems in financial data classification", *Intelligent Systems in Accounting, Finance and Management*, Vol. 27, No. 1, pp. 3-9. ISSN 1550-1949. DOI 10.1002/isaf.1460.

[35] Lane, W. R., Looney, S. W. and Wansley, W. (1986) "An application of the Cox proportional hazards model to bank failure", *Journal of Banking and Finance*, Vol. 10, pp. 511-531. ISSN 1872-6372. DOI 10.1016/S0378-4266(86)80003-6.

[36] Marais, M., Patell, J. and Wolfson, M. (1984) "The Experimental Design of Classification Models: An Application of Recursive Partitioning and Bootstrapping to Commercial Bank Loan Classifications", *Journal of Accounting Research*, Vol. 22, pp. 87-114. ISSN 1475679X. DOI 10.2307/2490861.

[37] Mazancová, K., Martinát, S., Pícha, K., and Navrátil, J. (2025) "Perceived and "Real” Importance of Subsidies for Agricultural Enterprises in Post-Communist EU Countries", *Sage Open*, Vol. 15, No. 1. ISSN 2158-2440. DOI 10.1177/21582440251324756.

[38] McKee, T. E. (1976) "Discriminant prediction of going concern status: a model for auditors", *Proceedings of the Sixtieth American Accounting Association Annual Meeting*, pp. 294-321. E-ISSN 3047-0579. DOI 10.62885/ekuisci.v3i1.843.

[39] McKee, T. E. (2000) "Developing a Bankruptcy Prediction Model via Rough Sets Theory", *International Journal of Intelligent Systems in Accounting, Finance & Management,* Vol. 9, No. 3, pp. 159-173. ISSN 1055-615X. DOI 10.1002/1099-1174(200009)9:33.0.CO;2-C.

[40] Messier, W. F. Jr. and Hansen, J. V. (1988) "Inducing Rules for Expert System Development: An Example Using Default and Bankruptcy Data", *Management Science*, Vol. 34, No. 12, pp. 1403-1415. ISSN 1526-5501. DOI 10.1287/mnsc.34.12.1403.

[41] Meyer, P. A. and Pifer, H. W. (1970) "Prediction of bank failures", *Journal of Finance*, Vol. 25, pp. 853-868. ISSN 1540-6261. DOI 10.2307/2325421.

[42] Milić, D. M., Tekić, D. D., Novaković, T. J., Zekić, V. N., Popov, M. R. and Mihajlov, Z. G. (2022) "The Impact of Profitability and Productivity on the Risk of Bankruptcy for Agricultural and Food Companies in Vojvodina", *Journal of Agricultural Sciences (Belgrade)*, Vol. 67, No. 1, pp. 47-61. ISSN 2683-3867. DOI 10.2298/JAS2201047M.

[43] Neumaier, I. and Neumaierová, I. (2005) "Index IN05", In: *Evropské finanční systémy 2005,* Masarykova univerzita, Brno, pp. 143-148. ISBN 80-210-3753-9.

[44] Neumaierová, I. and Neumaier, I. (2002) "*Výkonnost a tržní hodnota firmy*," 1st ed., Grada: Prague. ISBN 80-247-0125-1. (In Czech).

[45] Novotná, K., Gurčík, Ľ., Lušňáková, Z. (2023) "Economic determinants of the development and sustainability of family farms in Slovakia", *Agricultural Economics - Czech*, Vol. 69, pp. 291-299. ISSN 1805-9295. DOI 10.17221/143/2023-AGRICECON.

[46] Novotná, M. and Svoboda, J. (2010) "Application of a system of indices to an analysis of agricultural holdings economy in the LFA regions", *Acta Universitatis Agriculturae et Silviculturae Mendelianae. Brunensis*, Vol. LVIII, No. 3, pp. 165-174. ISSN 2464-8310. DOI 10.11118/actaun201058030165.

[47] Ohlson, J. (1980) "Financial ratios and the probabilistic prediction of bankruptcy", *Journal of Accounting Research*, Vol. 18, No. 1, pp. 109-131. ISSN 1475679X. DOI 10.2307/2490395.

[48] Pampel, F. C. (2011) "*Logistic Regression*", Thousand Oaks: SAGE Publications, Inc. ISBN 9781412984805. DOI 10.4135/9781412984805.

[49] Papíková, L. and Papík, M. (2022) "Effects of classification, feature selection, and resampling methods on bankruptcy prediction of small and medium-sized enterprises", *Intelligent Systems in Accounting, Finance and Management*, Vol. 29, No. 4, pp. 254-281. ISSN 2160-0074. DOI 10.1002/isaf.1521.

[50] Slowinski, R. and Zopounidis, C. (1995) "Application of the Rough Set Approach to Evaluation of Bankruptcy Risk", *International Journal of Intelligent Systems in Accounting, Finance and Management*, Vol. 4, No. 1, pp. 27-41. ISSN 2160-0074. DOI /10.1002/j.1099-1174.1995.tb00078.x.

[51] Spicka, J., Hlavsa, T., Soukupova, K. and Stolbova, M. (2019) "Approaches to estimation the farm-level economic viability and sustainability in agriculture: A literature review", *Agricultural Economics - Czech*, Vol. 65, pp. 289-297. ISSN 1805-9295. DOI 10.17221/269/2018-AGRICECON.

[52] Srebro, B., Mavrenski, B., Bogojević Arsić, V., Knežević, S., Milašinović, M. and Travica, J. (2021) "Bankruptcy Risk Prediction in Ensuring the Sustainable Operation of Agriculture Companies", *Sustainability*, Vol. 13, p. 7712. ISSN 2071-1050. DOI 10.3390/su13147712.

[53] Tam, K. Y. and Kiang, M. Y. (1992) "Managerial Applications of Neural Networks: The Case of Bank Failure Predictions", *Management Science*, Vol. 38, No. 7, pp. 926-947. ISSN 1526-5501.

[54] Timming, A. R. (2022) "*Applied Statistics*", Business & Management Research, SAGE Publications, Inc. ISBN 9781529792751.

[55] Tumpach, M., Surovičová, A., Juhászová, Z., Marci, A. and Kubiščková, V. (2020) "Prediction of the bankruptcy of Slovak companies using neural networks with SMOTE", *Ekonomický časopis*, Vol. 68, No. 10, pp. 1021-1039. ISSN 2729-7470. DOI 10.31577/ekoncas.2020.10.03.

[56] Tyshchenko, V., Achkasova, S., Karpova, V., Kanyhin, S. (2023) "Assessment the Influence of Debt Capital on the Bankruptcy of Enterprises in the Agricultural Sector", *Agricultural and Resource Economics*, Vol. 9, No. 2, pp. 183-204. ISSN 2414-584X. DOI 10.22004/ag.econ.337440.

[57] Vavrek, R., Kravčáková Vozárová, I. and Kotulič, R. (2021) "Evaluating the Financial Health of Agricultural Enterprises in the Conditions of the Slovak Republic Using Bankruptcy Models", *Agriculture*, Vol. 11, pp. 242. ISSN 2077-0472. DOI 10.3390/agriculture11030242.

[58] Vavrek, R., Kravčáková Vozárová, I., Kotulič, R., Adamišin, P., Dubrovská, M. and Ivanková, V. (2021) "Assessing the financial health of Agricultural enterprises incorporating the spatial dimension", *International Food and Agribusiness Management Review*, Vol. 25, No.3, pp. 469-488. ISSN 1559-2448. DOI 10.22434/IFAMR2020.0030.

[59] Wu, Y., Li, X., Liu, Q. and Tong, G. (2022) "The Analysis of Credit Risks in Agricultural Supply Chain Finance Assessment Model Based on Genetic Algorithm and Backpropagation Neural Network", *Computional Economics*, Vol. 60, pp. 1269-1292. ISSN 0927-7099. DOI 10.1007/s10614-021-10137-2 .

[60] Xie, C., Luo, C. and Yu, X. (2011) "Financial distress prediction based on SVM and MDA methods", *Quality & Quantity*, Vol. 45, No. 3, pp. 671-686. ISSN 1573-7845. DOI 10.1007/s11135-010-9376-y.

[61] Zmijewski, M. E. (1984) "Methodological Issues Related to the Estimation of Financial Distress Prediction Models", *Journal of Accounting Research*, Vol. 22, pp. 59-82. ISSN 1475679X. DOI 10.2307/2490859.

Volume XVIII Number 3, 2026

Evaluation of AI Predictability on Crop Yield using XAI - An Experimental Approach

Ramesh Varathan[1], P. Kumaresan[2]

[1] School of Computer Science Engineering and Information Systems, Vellore Institute of Technology, Tamil Nadu, India

[2] Centre For Artificial Intelligence Research (C-FAIR), Vellore Institute of Technology, Tamil Nadu, India

Abstract

Accurate crop yield prediction is essential for improving agricultural planning and sustainability. This study explores the integration of explainable artificial intelligence (XAI) with machine learning models to enhance transparency in agricultural decision-support systems. A comparative evaluation of multiple regression algorithms - Linear, Ridge, Lasso, Elastic Net, Decision Tree, Random Forest, AdaBoost, XGBoost, CatBoost, LightGBM, Gradient Boosting, and K-Nearest Neighbour Regression (KNR) is performed for crop yield prediction. Categorical attributes are transformed using One-Hot Encoding, while model reliability is ensured through 5-Fold Cross-Validation and hyperparameter optimization using Grid Search. Among the evaluated approaches, KNR demonstrated the best predictive capability with an R^2 value of 0.9827 and reduced errors (MAE:9.03, RMSE:117.58, MSE:13824.73). To improve interpretability, SHAP, LIME, and ELI5 techniques are employed to explain feature contributions and model decision. The proposed framework provides transparent and reliable insights, supporting data-driven agricultural management and sustainable crop production strategies.

Keywords

XAI, SHAP, LIME, CYP, Agriculture, ML, KNR.

Varathan, R. and Kumaresan P. (2026) "Evaluation of AI Predictability on Crop Yield using XAI - An Experimental Approach", *AGRIS on-line Papers in Economics and Informatics*, Vol. 18, No. 3, pp. 79-101. ISSN 1804-1930. DOI 10.7160/aol.2026.180307.

Introduction

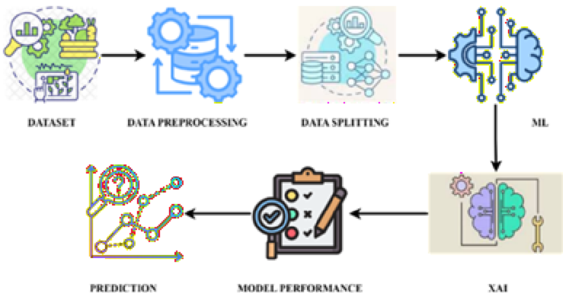
Farming is essential to human society, supplying food and serving as a foundation for employment and economic stability. Although humans have consumed grains and plants for over 10,000 years, the organized cultivation of crops and the management of arable land are comparatively recent advancements. Active land and vegetation management started about approximately 11,000 years ago, during the Neolithic period. Agriculture plays a crucial role in India's economy, providing the majority of the country's food needs and greatly enhancing economic stability. India's rapidly expanding population and significant climate change have made it more important than ever to maintain the agri-food value chain.

Climate change is altering temperature patterns and weather conditions, which significantly influences agricultural productivity. Variations in rainfall, rising temperatures, and unexpected climatic events can disrupt crop growth and reduce overall yield quality. These environmental changes also affect soil health, nutrient availability, and farming practices, making food production more uncertain. Addressing these challenges requires sustainable agricultural strategies and adaptive technologies to maintain stable food supplies and support long-term food security.

Many scientific methods have been applied in agriculture to balance food supply and demand. However, significant climate variability creates challenges for farmers in adopting more flexible and sustainable practices (Sharma et al., 2023). With the use of modern technology and innovative farming methods, agriculture needs to boost production while minimizing resource use. Therefore, precise crop production estimation is crucial to tackling food security issues. In India, farming is responsible for 70 percent of the global water consumption (Khan and Noor, 2019).

ML, a branch of AI, enables more accurate outcome prediction without the need for explicit programming. ML algorithms analyze historical data to forecast predicted outcomes. DL, a specialized subset of machine learning, utilizes algorithms modelled after biological neurons. With the advent of powerful processors, massive datasets, and flexible software libraries, deep learning has become one of the most popular machine learning methods available today. DL is characterized by the use of models with several processing layers that acquire data representations at different abstraction levels (LeCun et al., 2015).

Modern AI-based systems offer sophisticated solutions to a numerous intricate challenge in sector such as agriculture, healthcare, nutrition, energy, and transportation, all of which significantly contribute to enhancing people's quality of life. In this model operate with minimal humanlike involvement and permit only a small degree of inaccuracy. Recent advancements in DL have shown impressive precision in tackling intricate tasks including facial recognition, image segmentation, and object identification. Despite their high precision, it often remains challenging for users to understand how these algorithms arrive at their decisions. Consequently, AI researchers have devoted considerable effort to developing tools, methodologies, and approaches that facilitate user understanding and reliance on the output generated by machine learning models. In this field of study, known as XAI, the goal is to elucidate the rationale and underlying processes responsible for biases in AI-driven models (Arrieta et al., 2020).



Source: Authors´ own work

Figure 1: Workflow of Machine Learning and Explainable AI System for Forecasting agricultural yield.

The use of AI in agriculture is bringing a revolution in predicting agricultural outputs. Through machine learning algorithms, these predictions provide farmers with important insights into anticipated yields and consumption trends. The limited transparency of AI models, often described as black boxes, leads to doubts among agricultural stakeholders about the reliability and credibility of the predictions they generate.

To tackle these issues, XAI has emerged as a key research focus, aiming to enhance the transparency and interpretability of AI decisions by making the reasoning behind the model outputs more understandable to users. The research investigates the interaction of forecasting agricultural output and XAI, emphasizing practical relevance and interpretability in agricultural contexts (Akkem et al., 2024). The outlines the integration of ML and explainable AI in a system developed for forecasting agricultural output (Figure 1). Using XAI for crop yield prediction involves training AI models on agricultural data and then employing XAI techniques to understand which factors most influence the model's yield forecasts. This offers clear insights into the predictions, enabling farmers and researchers to pinpoint the main factors influencing crop yield. In turn, this understanding aids in making more informed decisions to enhance agricultural practices.

This work aims to underscore the critical role of interpretability in the crop yield forecasting and to guide its seamless integration into current predictive models. In pursuit of this goal, we investigate the key issues faced when integrating AI into crop yield forecasting systems. By evaluating sophisticated XAI tools such as SHAP, LIME, and ELI5, the paper outlines a systematic approach for increasing the clarity and trustworthiness of AI-powered crop yield prediction models. As agriculture continues to embrace digital transformation, the integration of AI and XAI empowers farmers to make predictions that are not only accurate but also interpretable. By enhancing decision-making and promoting sustainable farming, this progress also aligns with the larger goal of securing global food supplies. The subsequent sections of this paper delve into the evolving domain of crop yield prediction, its foundations in AI, and the growing role of XAI in enhancing predictive systems.

Gap analysis and contribution

1. Many crop yield prediction studies primarily focus on improving prediction accuracy using machine learning and deep learning models. The proposed framework integrates multiple XAI techniques (SHAP, LIME, and ELI5) to explain model predictions.

2. Several studies rely on a simple train-test split, which may lead to overfitting and unreliable performance estimates. The framework employs 5-Fold Cross-Validation to ensure robust model evaluation. Hyperparameter tuning using Grid

Search further improves model reliability and generalization.

3. Categorical agricultural variables are sometimes improperly handled, reducing model effectiveness. One-Hot Encoding is applied to transform categorical features into machine-readable representations. This improves learning efficiency and predictive performance.
4. Current research predominantly focuses on ensemble methods such as Random Forest, XGBoost, and LightGBM. K-Nearest Neighbour Regression (KNR) is less explored despite its potential effectiveness. The proposed framework combines high predictive performance with explainability. The integration of KNR and XAI techniques provides accurate predictions while maintaining transparency and accountability.

The existing literature on crop yield prediction predominantly emphasizes predictive accuracy while overlooking model interpretability, comprehensive algorithm comparison, and robust validation mechanisms. To address these limitations, this study proposes an explainable machine learning framework that integrates extensive regression model evaluation, hyperparameter optimization, 5-fold cross-validation, and multiple XAI techniques (SHAP, LIME, and ELI5). The framework achieves superior prediction performance using KNR while simultaneously providing transparent explanations of model decisions, thereby supporting trustworthy and sustainable agricultural decision-making.

The contribution of this work for crop yield prediction is specified below:

1. Comprehensive comparative evaluation of twelve ML regression algorithms for crop yield prediction.
2. One-Hot Encoding for Nominal agricultural features without Ordinal Bias.
3. Robust Validation through 5-fold cross-validation combined with GridSearchCV.
4. Identification of KNR as the best performing model with better metrics.
5. XAI framework deployment of SHAP, LIME, and ELI5.
6. Actionable decision support insights mapped to three stakeholder groups.
7. Comparative Performance Analysis of Models for Production Prediction
8. SOTA comparison demonstrating competitive superiority.

The manuscript is further organized in the following manner. Section 2 presents a comprehensive review of related work and discusses the foundational concepts of ML and XAI in agricultural domain. Section 3 delineates the methodological framework, including a detailed description of the dataset and feature engineering processes, and preprocessing techniques. Section 4 reports the experimental results and provides an in-depth analysis and interpretation of the findings. Section 5 discusses outcome of the implementation and the analysis of various ML models and XAI interpretation. Finally, Section 6 concludes the key contributions and outlining potential for future research. The major contributions are finally summarized in Section 6, which also outlines areas that could be explored further.

Related work

In the relevant article, the importance of yield predictions from several key ML and XAI papers is addressed. The relevant work focuses at the importance of ML and XAI in several fundamental research studies pertaining to crop yield forecasts.

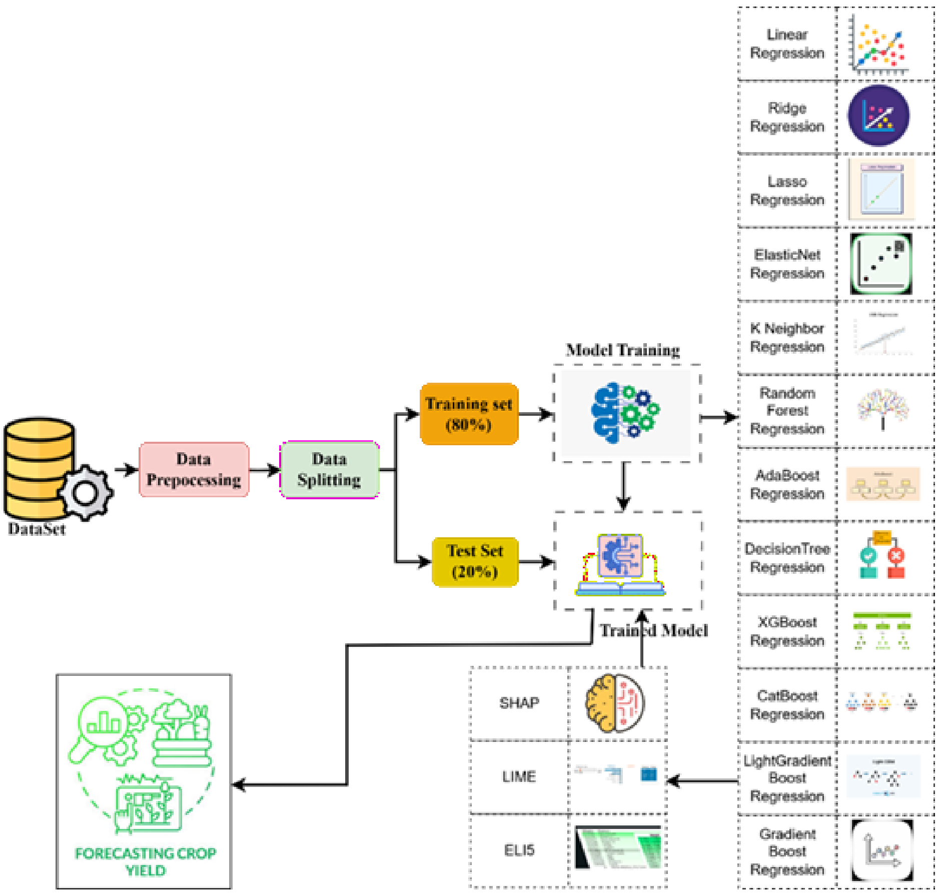
The combined ML with agronomic crop modeling principles to develop a foundational approach for large scale agricultural productivity forecasting. Their methodology followed a system focused on reusability, accuracy, and modularity. Feature extraction was performed using crop simulation results in combination with meteorological, remote sensing, and soil datasets from the MCYFS database (Abasha et al. 2023; Dilli et al., 2023). However, this work focused on large-scale pan-European forecasting and did not incorporate XAI interpretability methods to explain individual model predictions. Khaki and Wang (2019) introduced a deep neural network framework for predicting corn hybrid yield, verifying yield outputs, and evaluating yield variability, utilizing a combination of genotype data and environmental variables, including soil and weather conditions. Incorporating weather data showed that the model performed well, with an RMSE of 12 percent for average yield predictions and 50 percent for the standard deviation. However, the study was limited to a single crop type and did not provide any explainability of feature contributions. Abbas et al. (2020) investigated the effectiveness of four machine learning algorithms: Elastic net, linear regression, k-nearest neighbors, and support vector regression for predicting potato tuber yield across six field datasets in Atlantic Canada. The study focused on data collected during the 2017 and 2018 growing seasons from fields located in Prince Edward Island and New Brunswick, labelled as PE2017, PE2018, NB2017, and NB2018. Among the algorithms tested, SVR consistently outperformed the others across all datasets, achieving RMSE values of 5.97, 4.62, 6.60, and 6.17 t/ha for PE2017, PE2018, NB2017, and NB2018, respectively. A limitation of this study is that only four algorithms were compared and no XAI techniques were applied to explain the predictions. Vijayan and Chowdhary (2025) introduced a hybrid optimization algorithm, WOA_APSO (Whale Optimization Algorithm combined with Adaptive Particle Swarm Optimization), to enhance disease classification in rice crops. This approach integrates bio-inspired optimization techniques with deep learning models, specifically utilizing a Convolutional Neural Network (CNN) for classification tasks. The hybrid algorithm focuses on optimizing feature selection, thereby improving the CNN's performance in identifying various rice crop diseases. The proposed model achieved a remarkable accuracy of 97.5% on the benchmark dataset, outperforming models such as Support Vector Machines and Artificial Neural Networks. However, this work addressed disease classification rather than yield quantity prediction, and model decisions were not made interpretable through XAI. Khosla et al. (2020) concentrated on predicting kharif crop yields in the Visakhapatnam region of Andhra Pradesh. They employed a Modular Neural Network (MNN) to forecast monsoon rainfall, a key factor affecting kharif crop productivity. Following this, Support Vector Regression (SVR) was used to estimate actual yields using regional rainfall data. This integrated MNN-SVR approach supported the implementation of effective agricultural practices, resulting in enhanced crop yields and outperforming other prediction methods. While effective, the study was region-specific and lacked a systematic comparison across multiple ML algorithms or any XAI integration. Fagade and Pawar (2020) applied SVM and ANN models to classify crops based on environmental and soil-related parameters such as rainfall, temperature, soil type, pH, and humidity, utilizing data sourced from Maharashtra's agricultural website. The dataset encompassed nine agricultural zones, and a user-friendly interface was created to allow farmers to input relevant information. The ANN model achieved a high prediction accuracy of 86.80%, demonstrating its effectiveness.

Abekoon et al. (2025) explored the predictive capabilities of soil nitrogen, phosphorus, and potassium concentrations an 85-day cabbage growth cycle using explainable machine learning methods such as SHAP and LIME. Their analysis indicated that plant height and leaf count positively influenced nutrient level predictions, while the number of growth days and average leaf area exhibited a negative impact. According to their analysis, the number of growth days and average leaf area had a negative effect on nutrient level predictions, whereas plant height and leaf count had a positive effect. The model's interpretability and practical utility were validated through experiments with greenhouse-grown cabbage, culminating in the development of a user-friendly application aimed at supporting agricultural decision-making and optimization. However, this study focused on soil nutrient prediction rather than crop yield and was restricted to a single controlled crop environment. In their study, Agarwal et al., (2025) introduced the Successive Integration Hybrid Forecasting Model (SIHFM), a hybrid forecasting framework that leverages both statistical and dynamic techniques to estimate cotton crop yields using historical datasets from 1964 to 2020. The ARIMA model was employed for initial statistical forecasting, followed by LSTM and GRU models for dynamic predictions using the outputs from ARIMA. This hybrid framework successfully captured temporal trends and nonlinear patterns, resulting in improved yield prediction accuracy. A key limitation is that the study was restricted to cotton yield in India did not include XAI techniques for model transparency. Abdel-salam et al., (2024) proposed a novel crop yield prediction framework integrating hybrid feature selection with an optimized Support Vector Regression (SVR) model. The framework uses a new FMIG-RFE hybrid feature selection method after preprocessing with K-means clustering and CFS. SVR hyperparameters were adjusted using an improved Crayfish Optimization Algorithm (ICOA), which increased prediction accuracy. However, the study relied on a single SVR model and did not benchmark against a diverse set of regression algorithms, nor did it provide XAI based interpretability. Joshi et al., (2024) associates Long Short-Term Memory (LSTM), one-dimensional Convolutional Neural Network (1D-CNN), and Bidirectional LSTM (Bi-LSTM) models to create an explainable DL framework for agricultural output forecasting. They used methods like SHAP, Integrated Gradients (IG) (Abasha et al., 2025), and LIME to interpret model decisions. Bi-LSTM performed the best out of all the models, with an R2 score as high as 0.88. According to their analysis, the most important factors influencing winter wheat yield were temperature, precipitation, and enhanced vegetation index (EVI) during later stages of crop growth. While this study applied SHAP and LIME, it exclusively used deep learning models which are inherently less interpretable than ensemble regression methods. The present study addresses all of these limitations by applying 12 diverse regression algorithms with three XAI methods (SHAP, LIME, ELI5) in a leakage free, unified framework on Indian crop data.

Materials and methods

Proposed system

The proposed system architecture (Figure 2) presents an agricultural forecasting framework like ML regression and XAI. The process begins with the acquisition of a historical agricultural dataset that includes features such as crop, crop year, state, season, area, annual rainfall, fertilizer, pesticide, and yield. Data preprocessing is the first step, and it includes managing outliers, scaling numerical features, encoding categorical variables, and handling missing values. Following preprocessing, the dataset is split into two subsets: 20% for testing and 80% for training. LR, RR, Lasso Regression, ENR, KNR, RFR, ABR, DTR, XGBR, CBR, LGBMR, and GBR are among the machine learning regression models that are subsequently created using the training data. Following the training phase, model performance is quantitatively assessed on the test dataset using standard evaluation metrics, including MAE, MSE, RMSE, and R2_Score. To improve model transparency and interpretability, Explainable AI techniques like SHAP, LIME, and ELI5 are applied. These methods allow for a clear comprehension of the decision-making process by providing insights into the significance of features and the logic underlying model predictions. Ultimately, the most accurate and interpretable model is chosen for crop yield forecasting, aiding in informed agricultural planning and management.



Source: Authors´ own work

Figure 2: Predictive System Architecture for Agricultural Yield Forecasting using ML and XAI.

Pseudo Code: Agricultural Yield Forecasting using Supervised Regression Models Integrated with Explainable Artificial Intelligence

Input:

- Dataset (D):
- Features (X): crop, crop year, season, state, annual rainfall, area, fertilizer, pesticides.
- Target (y): yield
- Regression Models (M): Any regression model in ML.
- Explainability Methods: SHAP, LIME, ELI5.
- Evaluate the model with suitable performance metrics.

Algorithm steps:

Step 1: Load and Preprocess the Dataset

- Load the dataset D and handle missing values.
- Encode categorical features (Crop, Season, State) using one-hot encoding.
- Normalize numerical features (Annual_Rainfall, Area, Fertilizer, Pesticides, Crop_Year) using MinMaxScaler or StandardScaler.
- Partition the dataset into training (80%) and testing (20%).

Step 2: Cross-Validation & Hyperparameter Tuning

- Initialize k-fold CV with k splits
- Define parameter grid for model M
- Apply GridSearchCV with pipeline, param_grid, and k-fold CV
- Fit GridSearchCV on {X_train, y_train}
- best_pipeline = BRST_ESTIMATOR from GridSearchCV.

Step 3: Train the Regression Model (M)

- Select an appropriate supervised regression algorithm M for modelling the relationship between input features and the target variable. The relationship between the target variable and the input features is modeled by algorithm M.
- Fit the selected regression model M using the training dataset to learn the underlying mapping from features to target output.

Step 4: Model Evaluation

- Generate predicted target values best_pipeline by applying the trained model M to the test feature set (X_test)
- Compute performance measures:
  - Root Mean Squared Error (RMSE)
  - Mean Squared Error (MSE)
  - Mean Absolute Error (MAE)
  - R2_Score

Step 5: Explainability Using XAI Methods

Step 5.1: Global Explainability Using SHAP

- To measure the global contribution of each input feature to the model's predictions, compute SHAP values.
- Rank input features according to their relative importance in influencing the model's crop yield predictions, as determined by the computed SHAP values.

Step 5.2: Local Explainability Using LIME

- Select a test instance and generate perturbations.
- Train a simple local model to approximate the predictions of M.
- Identify the most important features influencing this specific prediction.

Step 5.3: Model Debugging & Explanation Using ELI5

- Use ELI5 to get feature importance rankings and visual explanations for how predictions are made.
- Generate weight-based feature importance for regression models.

Step 6: Decision Support and Interpretation

- If SHAP highlights “State” and “Season” globally → Useful for regional planning.
- If LIME shows high importance for “Annual_Rainfall” and “Fertilizer” → Farmers should focus on these factors.
- If ELI5 confirms feature importance rankings → Model’s decisions are interpretable.

OUTPUT:

- Predicted Crop Yield (y’) for new inputs.
- Best_Params = best_pipeline parameters.
- Performance evaluation metrics.
- Feature importance ranking from SHAP, LIME, and ELI5.
- Human-readable explanations for decisions using ELI5.

Dataset

Mendeley Data's historical agricultural production crop dataset (http://doi.org/10.17632/ncw2vbcgnk.2), encompasses records from 1997 to 2020, spanning 27 Indian states and 3 Union Territories. It comprises 19,689 instances, each described by ten distinct attributes as outline (Table 1), with detailed attribute definitions provide (Table 2). The dataset captures six agricultural seasons and includes data on 55 unique crop types cultivated across diverse agro-climatic regions of India.

Dataset	Total number of attributes	Number of instances
Crop Yield Dataset	9	19689

Source: Authors´ own analysis

Table 1: Dataset description.

No of Attributes	Description of the attributes
crop	Specifies the crop species or variety cultivated during the recorded agricultural cycle.
crop year	Denotes the specific calendar year corresponding to the cultivation period of the crop.
season	The period of the agricultural cycle during which the crop was planted and harvested
state	The Indian state where the agricultural activity took place.
area	The size of the land (in hectares) used for cultivating the crop.
annual rainfall	The annual rainfall (in millimeters) record in the crop cultivation area.
fertilizer	The crop's fertilizer dosage, expressed in kilograms.
pesticide	Pesticide quantity applied during crop cultivation, measured in kilograms.
yield	The production of the crop, calculated as yield per unit of cultivated area.

Source: Authors´ own analysis

Table 2: Description of the attributes.

Data preprocessing

Data preprocessing is essential to the effectiveness of AI and ML. Given that ML models are highly dependent on data quality, proper preprocessing is necessary to enable efficient learning before the data is fed into the model.

One hot encoding

A preprocessing method is used to translate categorical variables to a number that can be used in a machine learning model is one-hot encoding (Bolikulov et al., 2024). It generates individual binary columns for every category and makes it easy for models to deal with nominal data without causing artificial ranking among the categories. This technique helps make sure that each category is treated individually and avoids bias that comes from the assignment of numerical values to categorical attributes. In case of agricultural datasets, one-hot encoding helps to represent the difference across states, seasons, and crop types which can effectively help the prediction models to capture the regional and seasonal influence on the crop yield.

Training and testing data

The dataset is separated into two sections: the test dataset and the training dataset. Our model trains and tests using the dataset in an 80:20 random state of 42.

Machine learning models

Linear regression

Linear regression seeks to find a straight-line relationship between input features and the output feature. It assumes that the residuals are normally distributed with constant variance and that there is no correlation among the predictor variables (Samira et al., 2022).

Formula for predicting crop yield with nine input features and one output feature can be represented as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \varepsilon \quad (1)$$

where,

Y: The output feature, representing the predicted crop yield.

X_1, X_2, ……, X_8 = The nine input features, which could include factors such as Crop, Crop_Year, State, Season, Area, Annual_Rainfall, Fertilizer, Pesticide.

β_0 = The intercept term, representing the baseline crop yield when all input features are zero.

β_1, β_2, ……, β_8 = The coefficients foe the respective input features, indicating their impact on the crop yield.

ε = The error term, accounting for any variations not explained by the model.

The coefficients are determined using methods like Ordinary Least Squares. By interpreting the coefficients, one can understand the impact of each factor on crop yield. For instance, a positive coefficient for rainfall indicates an increase in yield with more rainfall.

Ridge Regression

Ridge Regression (RR) is a regularized regression technique that improves the stability of prediction models when strong correlations exist among independent variables. It introduces an L2 penalty term that shrinks the regression coefficients, helping to control model complexity and reduce overfitting. This approach stabilizes parameter estimation and improves prediction reliability while retaining all input variables in the model. As a result, ridge regression is widely used in agricultural data analysis and decision-support systems to handle multicollinearity and enhance model robustness (Sudha et al., 2022).

$$J(\beta) = \sum_{i=1}^{n}\left(Y_i - \beta_0 - \sum_{j=1}^{p}\beta_j X_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j{}^2 \quad (2)$$

Where:

$J(\beta)$ = Cost function to minimize

Y_i = Actual crop yield for the ith observation

β_0 = Intercept term

β_j = Coefficients of the j^{th} input feature

X_{ij} = Value of the jth feature for the ith observation

λ = Regularization parameter

n = Number of observations

p = Number of input features

Lasso regression

A linear regression method that combines regularization and variable selection is called the Least Absolute Shrinkage and Selection Operator (LASSO). LASSO is frequently used in agricultural settings. It operates by limiting the sum of the absolute values of the model coefficients by α and minimizing the traditional sum of squared errors. By reducing the coefficients of redundant or irrelevant features to precisely zero, the hyperparameter α enables the model to choose only the most significant predictors. LASSO regression's primary goal is to find a feature subset that reduces the prediction error for a continuous response variable. It is regarded as a simple yet effective linear model that highlights the most important variables in large datasets and has built-in regularization. However, for LASSO to function well, feature scaling and regularization parameter tuning must be done carefully (Sarr and Sultan, 2023).

$$\mathrm{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{\mathrm{n}} X_{\mathrm{n}} + \lambda\sum_{i=1}^{n}|\beta_i| \quad (3)$$

Where,

Y = Predicted crop yield

$X_1, X_2, \ldots, X_n$ = The independent input features

β_0 = Intercept term

$\beta_1, \beta_2, \ldots, \beta_n$ = The feature coefficients

λ = The regularization parameter that controls the strength of the penalty.

The coefficients are estimated by minimizing the lasso loss function. Its ultimately delivers interpretable results by selecting the most influential features affecting crop yield, making it particularly suitable for high-dimensional agriculture datasets.

Elastic Net regression

Elastic Net is a regression technique designed to handle situations where there is a high correlation among the features of a Agri dataset. It addresses the limitations of both Lasso Regression and Ridge Regression, providing a balanced approach to selecting the most relevant predictors while minimizing prediction errors. Unlike Lasso and Ridge individually, Elastic Net applies both L1 (Lasso) and L2 (Ridge) penalties simultaneously, resulting in a more effective and accurate prediction model (Kuruguntu Mohan et al., 2022).

$$Y = \beta_0 + \sum_{i=1}^{n}\beta_i X_i + \lambda_1\sum_{i=1}^{n}|\beta_i| + \lambda_2\sum_{i=1}^{n}\beta_i{}^2 \quad (4)$$

Where:

Y = Agricultural forecasting output

β_0 = Intercept term

β_i = Coefficients of the input features

X_i = Input features

n = Number of features

$\lambda_1\sum_{i=1}^{n}|\beta_i|$ = L1 regularization (Lasso),

$\lambda_2\sum_{i=1}^{n}\beta_i{}^2$ = L2 regularization (Ridge)

λ_1 and λ_2 = Hyperparameters controlling the strength of L1 and L2 regularization

K-Neighbor regression

The K-Nearest Neighbor algorithm is a distribution-free technique used for forecasting of agricultural output. In KNN, the parameter k plays a critical role in determining prediction accuracy. For a given test sample, the algorithm identifies k nearest neighbors from the training dataset and estimates the yield by averaging the response values of these k-neighbors (Maya Gopal and Bhargavi, 2019).

$$Y = \frac{1}{K}\sum_{i=1}^{k} Y_i \quad (5)$$

Where,

Y = Predicted crop yield

K = Number of nearest neighbors

Y_i = Actual crop yield of the ith nearest neighbor

Random Forest Regression

Random Forest Regression (RFR) is a robust ensemble learning technique widely used in crop yield prediction, along with resist the overfitting, capability to non-linear relationships, and efficiency in managing high-dimensional datasets. This method aggregates multiple decision trees to enhance predictive accuracy and stability, making it suitable for complex agricultural datasets that include factors like rainfall, fertilizer usage, Irrigation, Season, Area, and climatic conditions (Liaw and Wiener, 2002). RFR has been successfully applied in crop yield prediction across different regions and crop types. It leverages historical numerical datasets that include environmental, agronomic, and meteorological features.

$$Y = \frac{1}{M}\sum_{m=1}^{M} T_m(X) \quad (6)$$

Where,

Y = Predicted crop yield

M = The total number of estimators (decision trees) in the random forest model

$T_m(X)$ = Output of the m-th decision tree for the input X.

AdaBoost Regression

Adaptive Boosting Regression (ABR) is a ML technique that combines multiple weak learners in a sequential manner to form a strong predictor. The process starts by initializing weights for all training samples. In each iteration, a base estimator is fitted, and greater weight is given to the misclassified or poorly predicted observations, allowing the next learner to focus more on these difficult cases. This process is repeated, with each new learner aiming to correct the errors of its predecessor, ultimately improving the overall prediction performance (Nguyen and Ngo, 2025).

Decision Tree Regression

A decision tree addresses both classification and regression challenges through two primary functions. First, it selects the most significant features that influence decision-making. Second, it establishes the best possible outcome based on these selected features, often yielding a probability distribution over the potential results. Within a decision tree structure, each internal node represents a feature, each branch corresponds to a decision or split, and each leaf node indicates a final prediction or outcome. To initiate the tree construction, a specific feature serves as the root node. The tree-building process is then completed by recursively splitting the data (Burdett and Wellen, 2022).

$$\hat{Y} = \frac{1}{N}\sum_{i=1}^{N} y_i \quad (7)$$

Where:

$\hat{Y}$ = represents the estimated agricultural output

N = denotes the total number of observations within a terminal (leaf) node

y_i = represents the actual crop yield of the i^{th} sample in the terminal node

XGBoost regression

XGBoost is a well-known and effective open-source library that uses decision trees to implement the gradient-boosting algorithm. Regression trees are used by XGBoost as weak learners in regression tasks. A continuous score connects each input data point to a leaf. The model incorporates a regularization term (L1 and L2) to control complexity while optimizing a convex loss function. In order to predict the residuals (errors) of earlier trees, new trees are iteratively added during training. The final prediction is then improved by integrating these new trees with the ones that already exist. This method is frequently called "gradient boosting" because it continuously reduces loss with each additional model (Pavithra and Soundrapandiyan, 2024).

$$Y = F_M(X) = \sum_{m=1}^{M} f_m(X) + F_0 \quad (8)$$

Where:

Y = Predicted crop yield

$F_M(X)$ = Final XGBoost model prediction after M boosting iterations.

$F_m(X)$ = The prediction generated by the m^{th} weak learner.

F_0 = Initial forecast

M = Total number of estimators (iterations)

Cat Boost regression

Cat Boost is a gradient boosting-based machine learning specifically optimized for handling categorical features efficiently, where all nodes are split in the same way at each depth level. This structure helps to prevent overfitting and enhances computational efficiency (de Villiers et al., 2024).

Light Gradient Boost Machine regression

Light Gradient Boosting Machine (LGBM) is an ensemble algorithm based on gradient boosting decision trees. Unlike traditional decision trees that expand level wise, LightGBM grows leaf-wise, leading to more efficient and accurate splits. It uses histogram-based, cost-effective techniques where the time complexity depends on the number of bins rather than the size of the dataset (Mitra et al., 2024). The model generates predictions using an ensemble of decision trees. The forecast for crop yield Y' is obtained by summing the outputs of all trees in the ensemble.

$$Y' = \sum_{m=1}^{M} f_m(X) \quad (9)$$

Where,

Y' = Agricultural forecasting output

M = The overall number of decision trees included in the ensemble model

$f_m(X)$ = The output of the mth tree for the input feature set X

Gradient Boosting Regression

Gradient Boosting Regression (GBR) is a ML technique that builds predictive models by combining multiple weak learners, typically decision trees, to improve prediction accuracy. The method sequentially reduces prediction errors by learning from previous model residuals and refining the model in successive iterations. GBR is effective in agricultural forecasting because it can handle complex datasets and capture non-linear relationships among climatic variables, soil factors, and historical yield records. As a result, it provides reliable insights for crop yield prediction and supports informed decision-making in agricultural management (Mishra et al., 2020).

$$Y = F_M(X) = F_0(X) + \sum_{m=1}^{M} \gamma_m T_m(X) \quad (10)$$

Where,

Y = Predicted crop yield

M = Number of boosting iteration (trees).

$F_M(X)$ = Final boosted model prediction

$F_0(X)$ = Initial prediction.

$T_m(X)$ = Prediction from the m-th weak learner (decision tree)

γ_m = Learning rate (step size) controlling the contribution of each tree.

Explainable Artificial Intelligence for crop yield prediction

An emerging field known as XAI provides various methods to make the opaque nature of machine learning (ML) and deep learning (DL) models more transparent, allowing for explanations that are comprehensible to humans. Due to significant advancements in the application of ML and DL techniques, researchers in AI and ML are increasingly focusing on XAI.

SHapley Additive exPlanations

The Shapley Value, derived from game theory, quantifies the incremental impact of each player to the overall outcome (Ponce-Bobadilla et al., 2024). The explainable AI tool SHAP utilizes Shapley values to assess feature contributions to predictions.

$$g(z') = \emptyset_0 + \sum_{j=1}^{M} \emptyset_j z'_j \quad (11)$$

Where,

g = represents the interpretation model.

$z' \in \{0,1\}^M$ = indicates whether a specific feature is observable (1) or not (0).

Each feature is assigned a Shapley value (ϕ_i), a real number representing its contribution. Essentially, Shapley values work by calculating a feature's average contribution across every possible combination of features (Molnar, 2020). SHAP provides both global and local insights into model behavior and offers a model-agnostic way to estimate these Shapley values.

Local Interpretable Model-Agnostic Explanations

LIME is a strategy that provides a locally interpretable model by roughly modeling any black box machine learning model to clarify each prediction (Ribeiro et al., 2016). This technique works by first altering the original data points and then using these altered points to get predictions from the main model. Next, it trains an interpretable model using these new data points. LIME weights these new points based on how close they are to the original ones. This way, the resulting "explanation model" can explain each of the original data points. In this process, $\Omega(g)$ measures how complex the interpretable model g. The π_x indicates the size of the neighborhood around the original data point x, and f is the complex model we're trying to understand. The explanation model, g, is designed to be as accurate as possible in mimicking the original model's (f) predictions, which is measured by minimizing $L(f, g, \pi_x)$.

$$explanation(x) = argmin_{g \in G}\, \mathcal{L}(f, g, \pi_x) + \Omega(g) \quad (12)$$

Explain Like I'm 5

ELI5 intended for the interpretation of black-box machine learning models, providing both global and local explanations. It calculates feature importance for tree-based models by assessing how much each feature influences the prediction score, based on the change in score from a parent node to a child node within the decision tree (Bhattacharya, 2022).

$$\hat{Y} = Base\ Value + \sum_{i=1}^{n} Contribution\ (x_i) \quad (13)$$

Performance metrics

Root Mean Square Error

Root Mean Squared Error is a standard metric in regression analysis used to quantify the difference between predicted and actual values. It's calculated by taking the square root of the average of the squared differences between these values. A smaller RMSE indicates that the model fits the data more accurately (Uribeetxebarria et al., 2023). This metric is frequently used to evaluate the precision and effectiveness of regression models.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (14)$$

Where, n = no. of crop yield records

y_i = actual crop yield for the ith observation

$\hat{y}_i$ = predicted crop yield for the ith observation

$y_i - \hat{y}_i$ = prediction error for the ith observation.

Coefficient of determination

The coefficient of determination, more commonly known as the R2_Score, is a statistical measure used in predictive modeling. It assesses how well the independent variable(s) explain the variation in the dependent variable. Essentially, it tells us the proportion of the dependent variable's total variation that the model successfully accounts for. The R2_Score is calculated by dividing the explained variance by the total variance (Saadio et al., 2022).

$$\text{R2\_Score} = 1 - \frac{\sigma_r^2}{\sigma^2} \quad (15)$$

In this context, σ_r^2 signifies the sum of the squared difference between the expected and the observed values, while σ^2 represents the sum of the squared difference between the observed values and their mean.

Mean Squared Error

Mean Squared Error is a standard evaluation metric that computes the average of the squared discrepancies between the actual and predicted crop yield values (Joshi et al., 2023). By squaring each error term, the MSE places a greater penalty on larger errors, which helps in identifying models that make significant prediction mistakes. As *r* result, MSE is highly effective for measuring the precision and consistency of crop yield forecasting models.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \quad (16)$$

Where:

n = Total number of observations

Y_i = Actual crop yield for the i-th observation

$\hat{Y}_i$ = Predicted crop yield for the i-th observation.

Mean Absolute Error

MAE is a statistical metric used in regression analysis that measures the average magnitude of error by calculating the absolute differences between the forecast and observed yield values (Oikonomidis et al., 2022). Mathematically, MAE is often represented as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|Y_i - \hat{Y}_i| \quad (17)$$

Where:

Y_i represents the observed value for the i^{th} data value

$\hat{Y}_i$ represents the estimated value for ti^{th} data value

n represents the total number of data values.

Results and discussion

This section details the experimental results of the proposed ML model and XAI applied to crop yield prediction. The study employed systematically developed ML and XAI models. Model performance was assessed using standard regression metrics and compared against other ML models to identify the most suitable candidate for providing input to the XAI framework. The experiments were conducted in Python on a system equipped with an Intel Core i9-12900 processor, 32 GB of RAM, a 1 TB SSD and an NVIDIA RTX graphics card.

The comparison of various machine learning regression models used for crop yield prediction, evaluated using four key performance metrics: MAE, MSE, RMSE, and R2_Score (Table 3 below). Together, these metrics offer a comprehensive assessment of each model's accuracy and reliability. Among all the models, the KNR achieved the best overall performance, with the lowest MAE (9.03), MSE (13824.73), RMSE (117.58), along with a high R2_Score of 0.9827, indicating excellent predictive capability. Other high-performing models include RFR, which also yield low errors (MAE: 9.47, MSE: 15682.08, RMSE: 125.23) and a strong R2_Score of 0.9804. GBR, CBR, and DTR also show strong performance with R2_Scores above 0.97, reflecting their effectiveness in capturing complex patterns in the data. In contrast, the linear models – LR, RR, and Lasso Reg. – have relatively high MAE (ranging from 60.97 to 64.12) and RMSE values (389.88), indicating that they are less effective for this task due to their limitations in handling non-linear relationships. Elastic Net Regression shows the worst performance across all metrics, with the highest MAE (420.82), MSE (177091.92), RMSE (79.47), and a low R2_Score of 0.7790, suggesting that it is not suitable for accurate crop yield prediction in this context. ABR shows the mid performance of MAE (26.63), MSE (84741.90), RMSE (291.10), and R2_Score (0.8942). LGBM and XGBR fall in the middle tier. LGBM has decent MAE (18.0) and a respectable R2_Score (0.93), while XGBR is slightly less accurate with higher errors but still better than linear models.

Cross Validation (K-Fold)

The Cross Validation (K-Fold) technique is used to forecast how well models will perform. The dataset is divided into folds, which are groups of k identical dimensions. The model uses k-1 folds for training, and this process is repeated k times. We obtain a reliable evaluation of the algorithm's adaptability by computing the average performance across all iterations. This technique is essential for choosing models, assessing overall performance, and avoiding overfitting in machine learning models. Overconfidence in the model's performance may result from training it on the complete dataset, despite the fact that it may produce amazing results. A training set and a testing set are the two subsets into which the dataset is usually separated. While the independent test set provides an objective assessment of the chosen model, the training set is utilized in the cross-validation procedure to adjust model parameters. The cross-validation process uses one of the five equal segments of a dataset as the test set and trains the model on the other four segments. An accurate assessment of the model's performance is ensured by rotating the test set across the folds.

1st Fold Training set: 15751 Testing set: 3938 split 1

2nd Fold Training set: 15751 Testing set: 3938 split 2

3rd Fold Training set: 15751 Testing set: 3938 split 3

4th Fold Training set: 15751 Testing set: 3938 split 4

5th Fold Training set: 15752 Testing set: 3937 split 5

Grid Search Cross Validation

Grid Search CV is used to determine suitable hyperparameters for ML models in CYP. It systematically tests various parameter combinations and evaluates them using cross-validation to prevent overfitting. Selecting the parameter values that provide the best results

Cross Validation	ML Models	Performance Metrics			
		MAE	MSE	RMSE	R2_Score
K-Fold = 5, GridSearch CV	K-Neighbor Regression	9.03	13824.73	117.58	0.9827
	Random Forest Regression	9.47	15682.08	125.23	0.9804
	Gradient Boosting Regression	11.78	18727.51	136.85	0.9766
	CatBoost Regression	12.25	19124.43	138.29	0.9761
	Decision Tree Regression	9.38	23184.71	152.27	0.9711
	LGBM Regression	18.00	50112.14	223.86	0.9375
	XGBoost Regression	14.85	61515.74	248.02	0.9232
	AdaBoost Regression	26.63	84741.90	291.10	0.8942
	Linear Regression	63.88	151851.10	389.68	0.8105
	Ridge Regression	64.12	152006.70	389.88	0.8103
	Lasso Regression	60.97	151998.06	389.87	0.8103
	ElasticNet Regression	79.47	177091.92	420.82	0.7790

Source: Authors´ own analysis

Table 3: Comparison of ML Model Performance in agricultural forecasting.

increase the model's reliability and accuracy. This approach supports agricultural yield forecasting by considering factors such as rainfall, area, pesticides, and fertilizer.

Root Mean Square Error Comparison

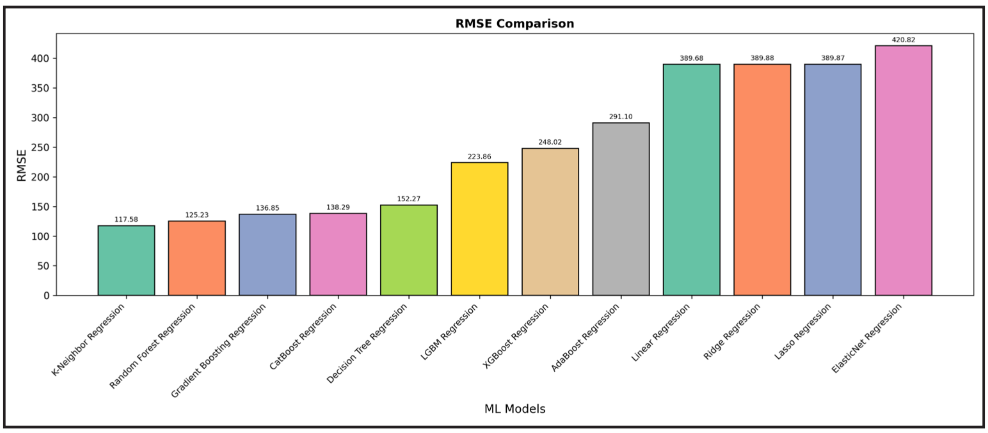
The RMSE comparison in the predictive performance of various ML models applied to CYP (Figure 3). It is evident that linear models such as LR, RR, and Las. Reg. yielded moderated results with RMSE values around 389 and ENR (420.82) performed poorly, highlighting their inability to capture the complex non-linear relationships in agricultural data. In contrast, non-linear and ensemble-based models demonstrated superior performance. The K-Neighbors Regressor achieved the lowest RMSE (117.58), followed closely by RFR (125.23), GBR (136.85), and CBR (138.29) confirming their effectiveness in handling the variability of agricultural features. Other ensemble methods such as DTR (152.27) also performed better than linear models, whereas LGBMR (223.86), XGBR (248.02), and ABR (291.10) showed relatively higher error rates. Overall, the results indicate that non-linear ensemble techniques are more suitable for accurate CYP compared to traditional linear regression methods.

Mean Absolute Error comparison

The MAE comparison in the performance of different ML models for CYP based on their MAE values (Figure 4). Lower MAE values indicate higher predictive accuracy, as they represent smaller average deviations between actual and predicted yields. Among the models, KNR achieved the best performance with the lowest MAE of 9.03, followed closely by DTR (9.38), RFR (9.47), GBR (11.78), CBR (12.25), XGBR (14.85), all of which showed strong predictive ability. In contrast, linear models such as LR (63.88), RR (64.12), and ENR (79.47) exhibited much higher error rates, reflecting their limitations in handling the non-linear and complex relationships present in agricultural datasets. Lasso Reg. (60.97) performed slightly better than other linear approaches but still fell behind the ensemble and tree-based models. ABR (26.63) and LGBMR (18.00) achieved moderate accuracy compared to the best performing models. Overall, the results highlight that non-linear ensemble and neighbor-based methods are significantly more effective than linear models in accurately predicting crop yield.

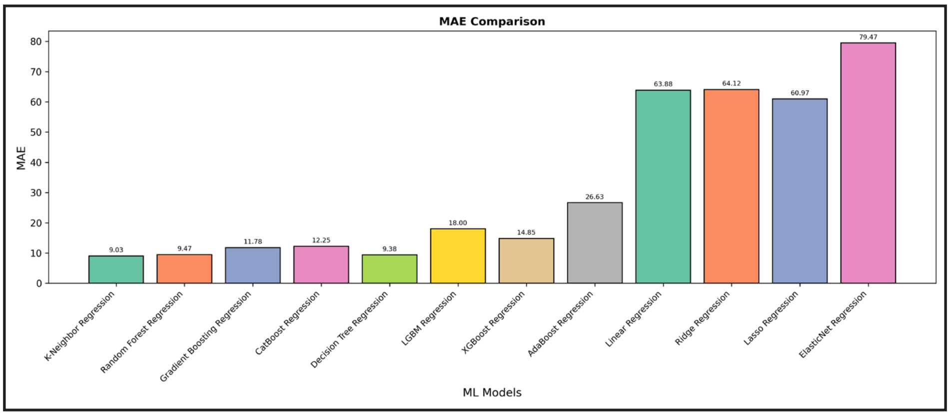
Mean Squared Error comparison

The MSE comparison in presents the performance of different ML models for CYP using MSE as the evaluation metric (Figure 5). Since MSE penalizes larger errors more heavily, lower values indicate higher prediction accuracy. From the Figure 5, it is evident that linear models such as LR (151851.10), RR (152006.70), Lasso Reg. (151998.06), and ENR (177091.92) performed poorly, producing very high error values and failing to capture the complex non-linear relationships in the dataset. In contrast, non-linear models and ensemble-based methods achieved much lower MSE values. The best-performing models were KNR (13824.73), RFR (15682.08), GBR (18727.51), CBR (19124.43), all of which demonstrated strong predictive capability. DTR (23184.71), LGBMR (50112.14), and XGBR (61515.74) also performed reasonably well, though with slightly higher error values compared to the top models. ABR (84741.90) showed moderate accuracy but was less effective than other ensemble approaches. Overall, the results



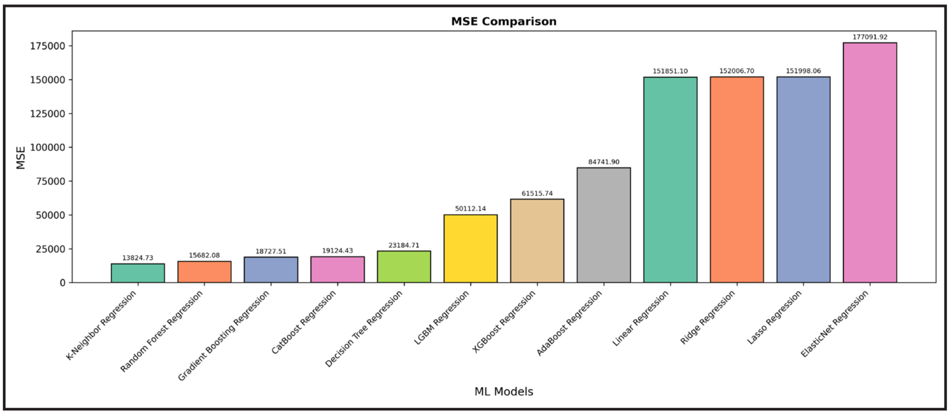
Source: Authors´ own analysis

Figure 3: RMSE-based performance comparison of ML Regression Models for agriculture.



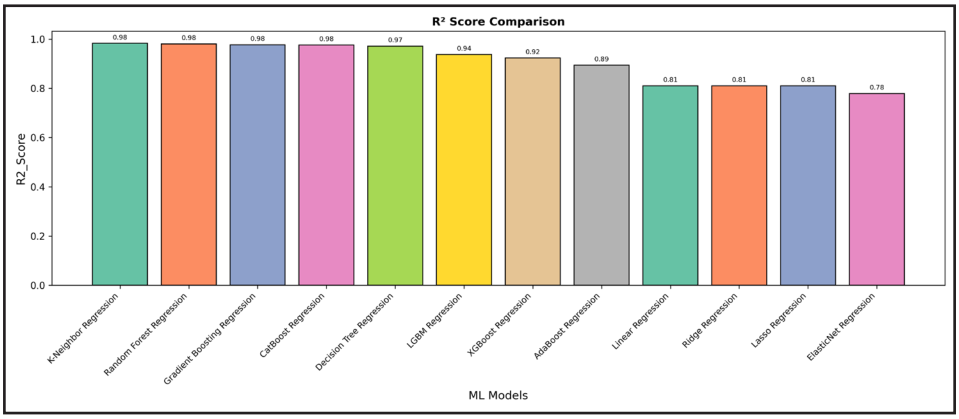
Source: Authors´ own analysis

Figure 4: MAE-based performance comparison of ML Regression Models for agriculture.



Source: Authors´ own analysis

Figure 5: MSE-based performance comparison of ML Regression Models for agriculture.



Source: Authors´ own analysis

Figure 6: R2_Score-based performance comparison of ML Regression Models for agriculture.

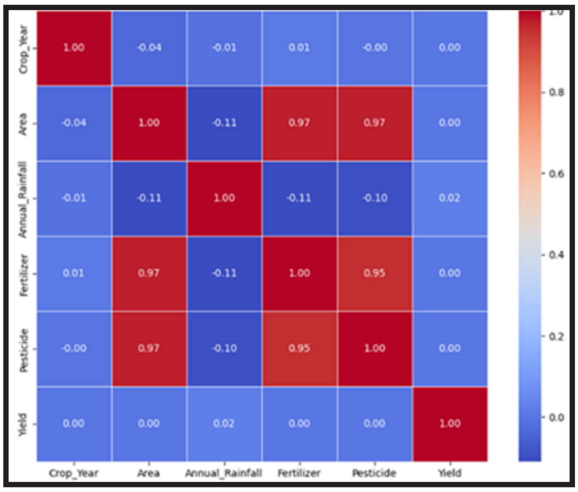
highlight that ensemble-based boosting methods and neighbor-based techniques significantly outperform traditional linear regression models, making them more suitable for accurate and reliable CYP.

R2_Score comparison

The R2_SCORE comparison in demonstrates the goodness of fit of different ML models used for CYP (Figure 6 above). The R2_SCORE implies the model performance. From the results, linear models such as Lasso Reg. (0.8103), RR (0.8103), and LR (0.8105) achieved moderate performance, while ENR (0.7790) performed poorly, showing limited ability to capture the complex patterns in agricultural data. In contrast, non-linear and ensemble-based models performed significantly better. The KNR (0.9827) achieved the highest R2_SCORE, indicating excellent predictive accuracy. Similarly, RFR (0.9804), GBR (0.9766), CBR (0.9761), DTR (0.9711), LGBMR (0.9375), and XGBR (0.9232) all produced strong results, highlighting their effectiveness in modelling the non-linear interactions among agricultural features. ABR (0.8942) also performed well, though slightly below other ensemble approaches. Overall, the emphasizes that non-linear and ensemble-based models provide superior performance compared to traditional linear methods, making them more suitable for robust and accurate CYP (Figure 6).

Correlation heat map

The correlation matrix, which illustrates the relationships among various factors within the dataset (Figure 7).



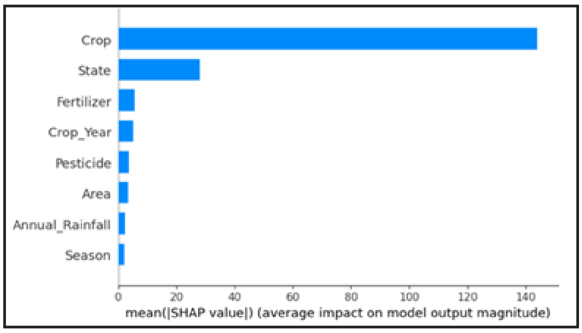
Source: Authors´ own analysis

Figure 7: Correlation heat map.

The analysis of the correlation heat map provides significant insights into how different features within the dataset are related. A key finding is the existence of strong positive correlations, where some features (Area, Pesticide, and Fertilizer) have correlation scores close to 1. If any climate condition is one feature increases, the other features also increase, then similarly any one feature also decreases, the other feature also decrease. The strong positive correlations are recommended for direct and regularly predictable relationship between these features. On the contrary, almost all features in the dataset have very low and negative correlations; Most features are around zero or below the correlation value. Weak correlations indicate that features are not related in a linear fashion. Changes in one aspect do not always imply changes in another aspect.

SHAP, LIME, ELI5

The SHAP feature importance for a regression model (Figure 8).



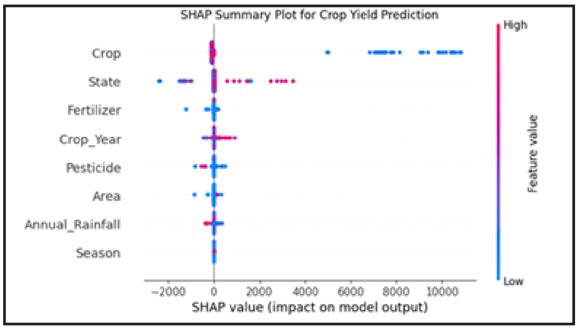
Source: Authors´ own analysis

Figure 8: SHAP feature importance plot for crop yield prediction model.

We're looking at a plot where the average impact of each feature on the model's output is on the x-axis, and the input features themselves are listed along the y-axis. From the chart, it is evident that crop has the highest impact on the model's predictions by a significant margin, with a SHAP value well above 140, signifying that is the key factor in ascertaining the target variable. The next most impactful features are State, Fertilizer, and Crop_Year, all of which have moderate contributions to the model's output. Other features such as State, Season, Annual_ Rainfall, and Area exhibit minimal influence on the model's prediction, as indicated by their comparatively low SHAP values. This indicates that although they are included in the dataset, their individual impact on the model's predictions is relatively limited. Overall, the chart highlights that the model heavily relies on Crop related attributes, with land State and input factors like Fertilizer and Crop_Year also playing notable roles in prediction accuracy, whereas temporal or categorical variables have a lesser effect.

The SHAP summary plot for crop yield prediction (Figure 9). This graph illustrates the influence of various features on the model's results, positioning the features on the vertical axis and their SHAP values on the horizontal axis. The features are ranked from top to bottom based on their importance, with Crop showing the highest impact, followed by Area, Fertilizer, Pesticide, State, Season, Annual_Rainfall, and Crop_Year.

Each point on the plot corresponds to an observation in the dataset, and its color shows the feature's value: red for high values and blue for low values.

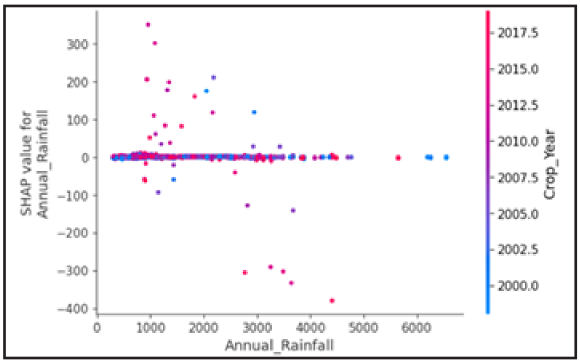


Source: Authors´ own analysis

Figure 9: SHAP summary plot for Crop Yield Prediction.

The position along the x-axis shows how much that particular observation's feature affected the prediction. Points further to the right indicate stronger positive impact on yield prediction, while points to the left show negative impact. For production, we can see many high-value (red) points extending far to the right, suggesting that higher production levels strongly correlate with higher predicted yields. The Area feature shows a more mixed pattern with both positive and negative effects. Features toward the bottom of the chart, like Season and Annual_Rainfall, have relatively smaller impacts on the model's predictions compared to the top features.

The SHAP dependence plot for Annual_Rainfall (Figure 10) reveals a non-linear, threshold-based relationship between rainfall and crop yield prediction.

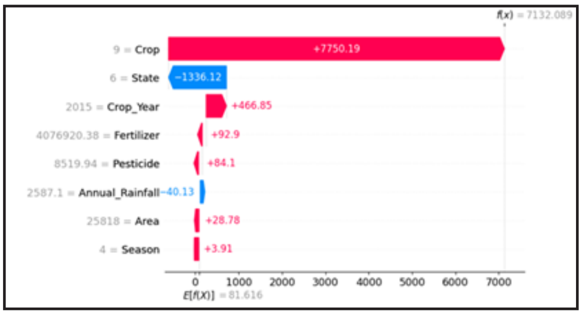


Source: Authors´ own analysis

Figure 10: SHAP dependence plot for Annual_Rainfall vs Crop_Year.

Moderate rainfall (500–1500 mm) exerts the strongest positive influence on predicted yield, with SHAP values reaching up to +350, consistent with the optimal water requirements of major Indian food crops. Beyond 1500 mm, the positive effect diminishes progressively, transitioning to strongly negative contributions (as low as −400) for rainfall exceeding 3000 mm — reflecting the yield-suppressing effects of waterlogging, flooding, and nutrient leaching under excessive precipitation. The Crop_Year colour interaction further reveals that recent years (2015–2017) exhibit more consistently positive rainfall effects, suggesting that improvements in agricultural technology and infrastructure have enhanced the crop system's capacity to convert adequate rainfall into higher yields. These findings underscore Annual_Rainfall as the most influential environmental driver of crop yield in the proposed KNR-based framework, while simultaneously highlighting the importance of managing excess water through drainage and irrigation control in high-rainfall Indian agricultural regions.

The detailed SHAP waterfall plot explaining the contribution of each feature to a specific crop yield prediction (Figure 11). The model starts with a base expected value E[f(x)] of 81.616 and arrives at a final prediction f(x) of 7132.089. Each feature's contribution is visualized as a bar with its specific impact value.

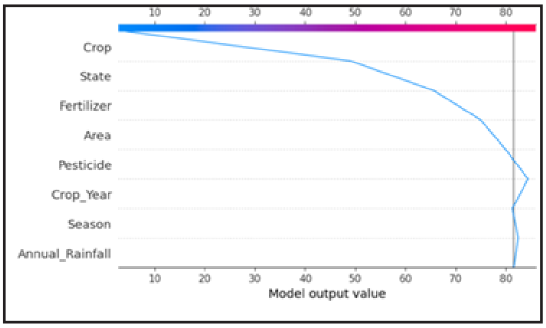


Source: Authors´ own analysis

Figure 11: Waterfall model for SHAP.

Crop (9) has the largest positive impact, contributing +7750.19 to the prediction. Other significant positive contributors including Crop_Year (2015) adding +466.85, Fertilizer (4076920) adding +92.9, and Pesticide (8519.94) adding +84.1. State (6) has a negative impact of -1336.12 and Annual_Rainfall (2587.1) has a negative impact of -40.13 shown in blue. The remaining features have smaller positive impacts: Area (25818) adds +28.78, and Season (4) contributes +3.91. This visualization effectively demonstrates how each agricultural and environmental factor contributes to the final yield prediction, with Production being overwhelmingly the most influential positive factor, while State characteristics slightly reduce the predicted yield. The model transforms a base expectation of about 82 units to a much higher prediction of over 7,000 units based primarily on these feature values.

The SHAP decision plot (Figure 12) traces the cumulative feature contribution pathway for a selected high-yield test instance.

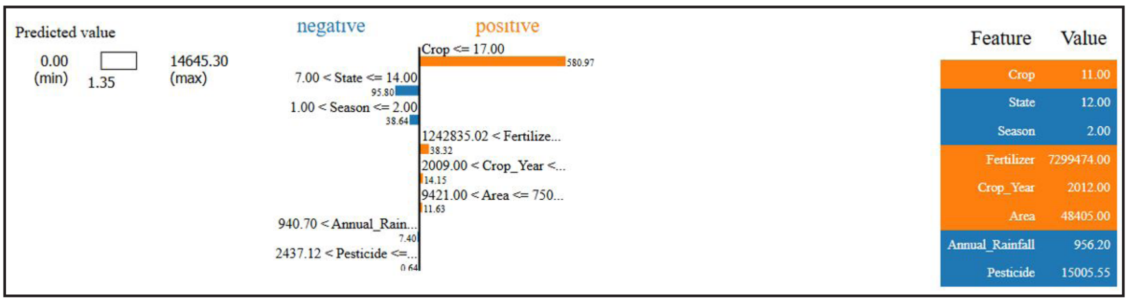


Source: Authors´ own analysis

Figure 12: SHAP summary plot – global feature importance for crop yield prediction.

It illustrates how the KNR model progresses from a baseline mean output of approximately 10 kg/ha to a final prediction of ~83–84 kg/ha through sequential, interpretable feature adjustments. Crop type emerges as the dominant driver, contributing approximately +55 units the largest single shift confirming that crop variety selection is the primary determinant of yield for this instance. State and Fertilizer provide secondary positive contributions of +8 and +5–6 units respectively, reflecting the joint influence of geographic agro-climatic context and nutrient management on productivity. Season and Annual_Rainfall contribute minimally for this instance, indicating their values fall close to the dataset mean for these features. The smooth, predominantly rightward trajectory of the decision path transitioning from blue to pink across the colour gradient visually confirms that this is a high-yield scenario where all primary features align favourably, and no significant negative corrections are introduced. This instance-level explanation validates the agronomic coherence of the KNR model and demonstrates the practical utility of SHAP decision plots in providing transparent, step-by-step justification of individual crop yield predictions for farmer-level and policy-level decision support.

The LIME local explanation for the selected test instance (Figure 13) reveals that the KNR model's prediction of 1.35 kg/ha a near-minimum yield esults from a net balance of strongly competing feature influences. Crop type exerts the dominant positive influence (+580.97), confirming that crop variety selection is the most critical agronomic determinant at the instance level. However, this positive force is substantially counteracted by the geographic state (−95.80) and seasonal context (−38.64), which together impose structural yield penalties that reflect regional productivity constraints and seasonal growing condition limitations. Fertilizer application contributes a moderate positive effect (+38.32), while Crop_Year (+14.15) and Area (+11.63) provide smaller supportive contributions. Annual_Rainfall and Pesticide contribute marginally (−7.40 and −0.64 respectively), indicating that their values for this instance are near the model's learned neutral threshold. The LIME explanation confirms that the model's local decision logic is agronomically coherent and practically interpretable: to improve yield for instances of this type, interventions should prioritise crop variety optimisation and targeted state-level agricultural support — as these two factors collectively determine the largest share of the predicted yield outcome. The absence of 'Production' from all feature contributions further validates the data leakage correction, confirming that the model's explanations are grounded exclusively in genuine agronomic and environmental relationships.



Source: Authors´ own analysis

Figure 13: LIME explanation for local feature importance in crop yield prediction.

ELI5 presents the feature importance weights and their interpretations in the context of a crop yield prediction model (Table 4). The values represent the average contribution of each feature, with associated uncertainties, indicating how influential each variable is in determining the yield outcome. Crop is by far the strongest predictor of yield, with a weight nearly 5.5 times greater than the next feature, reflecting the dominant role of intrinsic biological differences between crops. Geographic state ranks second, capturing the combined influence of climate, soil, irrigation infrastructure, and regional farming practices. Among agronomic inputs, fertilizer contributes the most stable and meaningful signal, followed by cultivated area and pesticide usage, which together reflect farm-level management intensity. Seasonal variation and annual rainfall show only marginal effects, the latter likely because irrigated systems reduce dependence on natural precipitation. Crop year carries the least reliable signal, as its standard deviation exceeds its mean weight, indicating high instability and limited predictive value.

Comparative performance analysis of models for production prediction

- features (X): crop, crop year, season, state, annual rainfall, area, fertilizer, pesticides.
- Target (y): Production

The comparison of various machine learning regression models used for crop production prediction, evaluated using four key performance metrics: MAE, MSE, RMSE, and R2_Score (Table 5 below). Together, these metrics offer a comprehensive assessment of each model's accuracy and reliability. Among all the models, the RFR achieved the best overall performance, with the lowest MAE (1432742.01), MSE (6.235554e+14), RMSE (24971091.98), along with a high R2_Score of 0.9918, indicating excellent predictive capability. Other high-performing models include GBR, which also yield low errors (MAE: 1400660.38, MSE: 6.383589e+14, RMSE: 25265765.27) and a strong R2_Score of 0.9916. KNR, DTR, XGBR, and CBR also show strong performance with R2_Scores above 0.98, reflecting their effectiveness in capturing complex patterns in the data. In contrast, the linear models – LR, RR, and Lasso Reg. – have relatively high MAE (ranging from 34745237.46 to 37558141.57) and RMSE values (211669841.00 to 212629615.82), indicating that they are less effective for this task due to their limitations in handling non-linear relationships. Elastic Net Regression shows the worst performance across all metrics, with the highest MAE (32678725.98), MSE (4.646412e+16), RMSE (215555367.05), and a low R2_Score of 0.3888, suggesting that it is not suitable for accurate crop yield prediction

Weight	Feature	Interpretation
1.8969 ± 0.2518	Crop	Crop type is overwhelmingly the most dominant predictor of yield. Its weight is nearly5.5x greater than the next feature, indicating that the intrinsic biological characteristics of each crop.
0.3423 ± 0.1417	State	Geographic region is the second most influential feature, capturing the combined effect of agro-climatic conditions, soil type, irrigation infrastructure, local agricultural policy, and farming practices that vary substantially across states.
0.0367 ± 0.0203	Fertilizer	Fertilizer application contributes a moderate and stable signal. Among the agronomic input variables, it ranks highest, consistent with the agronomic principle that nutrient availability particularly nitrogen, phosphorus, and potassium is a primary limiting factor for crop productivity.
0.0184 ± 0.0117	Area	Cultivated area exerts a notable though secondary influence on yield prediction. Larger cultivated areas may reflect greater mechanisation, access to irrigation, or economy-of-scale advantages, but can also introduce yield dilution effects if marginal land is included.
0.0130 ± 0.0095	Pesticide	Pesticide usage contributes a small but measurable effect, reflecting its role in mitigating yield losses due to pest infestations, fungal diseases, and weed competition.
0.0125 ± 0.0030	Season	Seasonal variations slightly impact yield, suggesting some seasonal dependency, but it is not a dominant factor.
0.0108 ± 0.0072	Annual_Rainfall	Annual rainfall has a marginal direct effect on yield prediction, which may appear counterintuitive but is plausible in datasets dominated by irrigated agricultural systems where supplemental water supply decouples yield form natural precipitation.
0.0085 ± 0.0135	Crop_Year	The year of cultivation exerts minimal influence on yield, but its SD (±0.0135) exceeds its mean weight (0.0085), indicating high instability and unreliable signal.

Source: Authors´ own work

Table 4: feature importance and interpretation in crop yield prediction-ELI5.

Cross Validation	ML Models	Performance Metrics			
		MAE	MSE	RMSE	R2_Score
K-Fold=5, GridSearch CV	Random Forest Regression	1432742.01	6.24E+14	24971091.98	0.9918
	Gradient Boosting Regression	1400660.38	6.38E+14	25265765.27	0.9916
	K-Neighbor Regression	1637409.01	8.10E+14	28465197.09	0.9893
	Decision Tree Regression	1846917.51	8.63E+14	29372158.02	0.9887
	XGBoost Regression	1648909.12	9.44E+14	30724896.24	0.9876
	CatBoost Regression	2442648.76	1.27E+15	35576511.38	0.9834
	LGBM Regression	2986105.2	1.67E+15	40822064.44	0.9781
	AdaBoost Regression	3039801.11	1.76E+15	41899915.66	0.9769
	Lasso Regression	37557266.6	4.48E+16	211669932.5	0.4107
	Linear Regression	37558141.57	4.48E+16	211669841	0.4107
	Ridge Regression	34745237.46	4.52E+16	212629615.8	0.4053
	ElasticNet Regression	32678725.98	4.65E+16	215555367.1	0.3888

Source: Authors´ own work

Table 5: Comparison of ML model performance in agricultural forecasting with production as target.

Author Name	Algorithms	Best Model	Input Parameters	Evaluation Results
Shams et al. (2024)	XAI_CROP, GB, DT, RF, GNB, MNB	XAI-CROP	District_Name, Season, Area, Production, Crop	R2_Score=0.94
Sathya and Gnanasekaran (2023)	MLR-LSTM, MLR, SVM, LSTM, RF	MLR-LSTM	pH range, rainfall, maxtemp, mintemp, block, Windspeed	R2_Score=0.93
Malashin et al. (2024)	DNN, GA, LIME	DNN	Crop, Season, Area, State, Annual_Rainfall, Pesticide, Fertilizer, Production, Yield	R2_Score=0.92
Mohan et al. (2024)	DTR, RFR, LGBMR, SHAP, LIME	LGBMR	Rainfall, Temperature, Fertilizer, Phosphorus, Nitrogen, Potassium	R2_Score=0.92
Proposed	LR, RR, LassoR, DTR, RFR, LGBMR, CBR, ABR, ENR, GBR, XBR, KNR, SHAP, LIME, ELI5	KNR	Crop, Crop_Year, Season, Area, State, Annual_Rainfall, Pesticide, Fertilizer, Yield	R2_Score=0.94

Source: Authors´ own work

Table 6: Crop Yield Prediction comparing with various State-of-the-Art (SOTA) Models.

in this context. ABR shows the mid performance of MAE (3039801.11), MSE (1.755603e+15), RMSE (41899915.66), and R2_Score (0.9769). LGBM fall in the middle tier. LGBM has decent MAE (2986105.20) and a respectable R2_Score (0.9781), is slightly less accurate with higher errors but still better than linear models.

Comparison with the state-of-the-art models

The proposed work compared with the state-of-the-art model (SOTA) as shown in below (Table 6 above).

The comparative analysis of various crop yield prediction models based on the algorithms used, the input parameters considered, and the corresponding evaluation results in terms of R2_Score (Table 6). References (Shams et al., 2024; Sathya and Gnanasekaran, 2023; Malashin et al., 2024; Jagan Mohan et al., 2024) represent existing studies, each employing a combination of traditional ML, DL, and XAI techniques. Input parameters vary across models, including geographic attributes, environmental conditions, soil characteristics, and crop-specific data. The proposed model, in contrast, integrates a broader set of ensemble and regression algorithms, including LR, RR, Las. Reg., DTR, RFR, LGBMR, CBR, ABR, ENR, GBR, XBR, KNR, and advanced explainability tools such as SHAP, LIME, and ELI5. It uses a rich set of input features covering crop details, temporal factors, environmental inputs, and geographical identifiers. The proposed model achieves the highest R2_Score of 0.94, demonstrating superior performance and accuracy in crop yield prediction compared to previously reported models.

Conclusion

The present study successfully demonstrates that integrating Explainable Artificial Intelligence (XAI) with advanced machine learning techniques significantly enhances both the accuracy and transparency of crop yield prediction systems. By evaluating a wide range of regression algorithms, the research establishes a comprehensive comparative framework, ensuring that model selection is not only performance-driven but also analytically justified. Among all the models, the K-Nearest Neighbour Regression (KNR) model emerged as the most effective, achieving a high R2_Score of 0.9827 along with low error metrics (MAE: 9.03, MSE: 13824.73, RMSE: 117.58). These results confirm the robustness and predictive strength of the proposed approach in handling complex agricultural datasets.

A key contribution of this work lies in its emphasis on interpretability through the integration of XAI techniques like SHAP, LIME, and ELI5. Unlike traditional "black-box" models, the proposed framework provides clear insights into how different features influence crop yield predictions. This transparency is crucial in agriculture, where decision-making often relies on trust and understanding. By identifying the most influential factors affecting yield, the system empowers farmers, agronomists, and policymakers to make informed, data-driven decisions that can improve productivity and reduce uncertainty.

Furthermore, the use of preprocessing techniques such as One-Hot Encoding for categorical variables, along with 5-Fold Cross-Validation and Grid Search optimization, ensures the reliability and generalizability of the models. These methodological choices reduce overfitting and enhance the stability of predictions across different data subsets. The study, therefore, not only focuses on predictive performance but also maintains strong experimental rigor, making the findings more credible and applicable in real-world agricultural scenarios.

The integration of KNR with SHAP-based explanations stands out as a particularly effective combination, offering both high predictive accuracy and meaningful interpretability. This dual advantage addresses a major challenge in modern AI applications—balancing performance with explainability. The ability to visualize feature contributions enables stakeholders to understand the reasoning behind predictions, fostering greater adoption of AI-driven systems in agriculture. As a result, the proposed model supports sustainable farming practices by enabling optimized resource utilization and improved crop management strategies.

Further the results confirm that ensemble-based models, particularly Random Forest Regression, are highly effective for accurate crop production prediction and significantly outperform traditional linear regression methods.

In conclusion, this research contributes significantly to the field of precision agriculture by bridging the gap between high-performance machine learning models and the need for transparency. It sets a strong foundation for the development of intelligent, explainable, and reliable agricultural decision-support systems. Future work can further enhance this framework by incorporating region-specific models and integrating dynamic environmental factors such as rainfall, temperature, and soil conditions. Such advancements will improve the adaptability and scalability of the system, ultimately promoting sustainable and resilient agricultural practices in diverse farming environments.

Acknowledgements

The authors express their sincere to Vellore Institute of Technology (VIT) for their assistance with this research.

Corresponding author:
Dr. Kumaresan, P.
Centre For Artificial Intelligence Research(C-FAIR)
Vellore Institute of Technology, Vellore-632014, Tamilnadu, India
Phone: +91-9842582301, E-mail: pkumaresan@vit.ac.in

References

[1] Abbas, F., Afzaal, H., Farooque, A. A. and Tang, S. (2020) "Crop Yield Prediction through Proximal Sensing and Machine Learning Algorithms", *Agronomy*, Vol. 10, No. 7, p. 1046. ISSN 2073-4395. DOI 10.3390/agronomy10071046.

[2] Abdel-Salam, M., Kumar, N. and Mahajan, S. (2024) "A proposed framework for crop yield prediction using hybrid feature selection approach and optimized machine learning", *Neural Computing and Applications*, Vol. 36, No. 33, pp. 20723-20750. ISSN 1433-3058. DOI 10.1007/s00521-024-10226-x.

[3] Abekoon, T., Hirushan, S., Namal, R., Imesh, U. E., Anuradha, J. and Upaka, R. (2025) "A Novel Application with Explainable Machine Learning (SHAP and LIME) to Predict Soil N, P and K Nutrient Content in Cabbage Cultivation", *Smart Agricultural Technology,* Vol. 11 (March), p. 100879. ISSN 2772-3755. DOI 10.1016/j.atech.2025.100879.

[4] Abhasha, J., Pradhan, B., Chakraborty, S., Varatharajoo, R., Alamri, A., Gite, S. and Lee, C.-L. (2025) "An Explainable Bi-LSTM Model for Winter Wheat Yield Prediction", *Frontier in Plant Science*, Vol. 15, pp. 1-17. ISSN 1664-462X. DOI 10.3389/fpls.2024.1491493.

[5] Abasha, K., Pradhan, B., Gite, S. and Chakraborty, S. (2023) "Remote-Sensing Data and Deep-Learning Techniques in Crop Mapping and Yield Prediction: A Systematic Review", *Remote Sensing*, Vol. 15, No. 8, p. 2014. ISSN 2072-4292. DOI 10.3390/rs15082014.

[6] Agarwal, N., Choudhry, N. and Tripathi, K. C. (2025) "A novel hybrid time series deep learning model for forecasting of cotton yield in India", *International Journal of Information Technology (Singapore)*, pp. 1745-1752. ISSN 2511-2104. DOI 10.1007/s41870-024-02327-6.

[7] Akkem, Y., Biswas, S. K. and Varanasi, A. (2024) "Streamlit-based enhancing crop recommendation systems with advanced explainable artificial intelligence for smart farming", Neural Computing and Applications, Vol. 36, No. 32, pp. 20011-20025. ISSN 1433-3058. DOI 10.1007/s00521-024-10208-z.

[8] Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R and Herrera, F. (2020) "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI", *Information Fusion*, Vol. 58, pp. 82-115. ISSN 1566-2535. DOI 10.1016/j.inffus.2019.12.012.

[9] Bhattacharya, A. (2022) "*Applied Machine Learning Explainability Techniques*", Birmingham, UK: Packt Publishing Ltd., 1st ed.. ISBN 978-1803246154.

[10] Bolikulov, F., Nasimov, R., Rashidov, A., Akhmedov, F. and Cho, Y. I. (2024) "Effective Methods of Categorical Data Encoding for Artificial Intelligence Algorithms", *Mathematics*, Vol. 12, No. 16. ISSN 2227-7390. DOI 10.3390/math12162553.

[11] Burdett, H. and Wellen, C. (2022) "Statistical and Machine Learning Methods for Crop Yield Prediction in the Context of Precision Agriculture", *Precision Agriculture,* Vol. 23, No. 5, p. 1553-74. ISSN 1385-2256. DOI 10.1007/s11119-022-09897-0.

[12] de Villiers, C., Mashaba-Munghemezulu, Z., Munghemezulu, C., Chirima, G. J. and Tesfamichael, S. G. (2024) "Assessing Maize Yield Spatiotemporal Variability Using Unmanned Aerial Vehicles and Machine Learning", *Geomatics*, Vol. 4, No. 3, pp. 213-236. ISSN 2673-7418. DOI 10.3390/geomatics4030012.

[13] Dilli, P., Boogaard, H., de Wit, A., Janssen, S., Osinga, S., Pylianidis, C. and Athanasiadis, I. N. (2021) "Machine Learning for Large-Scale Crop Yield Forecasting", *Agricultural Systems,* Vol. 187, (June 2020), p. 103016. ISSN 0308-521X. DOI 10.1016/j.agsy.2020.103016.

[14] Fegade, T. K. and Pawar, B. V. (2020) "Crop prediction using artificial neural network and support vector machine", In *Data Management, Analytics and Innovation: Proceedings of ICDMAI 2019*, Springer Singapore, Vol. 2, pp. 311-324. ISBN 978-981-13-9363-1.

[15] Jagan Mohan, R. N. V., Rayanoothala, P. S. and Sree, R. P. (2024) "Next-Gen Agriculture: Integrating AI and XAI for Precision Crop Yield Predictions", *Frontiers in Plant Science*, Vol. 15, pp. 1-16. ISSN 1664-462X. DOI 10.3389/fpls.2024.1451607.

[16] Khaki, S. and Wang, L. (2019) "Crop yield prediction using deep neural networks", *Frontiers in Plant Science*, Vol. 10, pp. 1-10. ISSN 1664-462X. DOI 10.3389/fpls.2019.00621.

[17] Khan, M. and Noor, S. (2019) "Irrigation runoff volume prediction using machine learning algorithms", *European Journal of Science, Innovation and Technology*, Vol. 8, pp. 1-22. ISSN 2304-9693.

[18] Khosla, E., Dharavath, R. and Priya, R. (2020) "Crop yield prediction using aggregated rainfall-based modular artificial neural networks and support vector regression", *Environment, Development and Sustainability*, Vol. 22, No. 6, pp. 5687-5708. ISSN 1387-585X. DOI 10.1007/s10668-019-00445-x.

[19] Kuruguntu Mohan, K., Maheswari, N. and Sivagami, M. (2022) "Models for Feature Selection and Efficient Crop Yield Prediction in the Groundnut Production", *Research in Agricultural Engineering,* Vol.68, No. 3, pp. 131-41. ISSN 1212-9151. DOI 10.17221/15/2021-RAE.

[20] LeCun, Y., Bengio, Y. and Hinton, G. (2015) "Deep learning", *Nature*, Vol. 521, pp. 436-444. ISSN 0028-0836. DOI 10.1038/nature14539.

[21] Liaw, A. and Wiener, M. (2002) "Classification and regression by randomForest", *R News*, Vol. 2, No. 3, pp. 18-22. ISSN 1609-3631.

[22] Malashin, I., Tynchenko, V., Gantimurov, A., Nelyub, V., Borodulin, A. and Tynchenko, Y. (2024) "Predicting Sustainable Crop Yields: Deep Learning and Explainable AI Tools", *Sustainability (Switzerland),* Vol. 16, No. 21, pp. 1-29. ISSN 2071-1050. DOI 10.3390/su16219437.

[23] Maya Gopal, P. S. and Bhargavi, R. (2019) "Performance Evaluation of Best Feature Subsets for Crop Yield Prediction Using Machine Learning Algorithms", *Applied Artificial Intelligence*, Vol. 33, No. 7, pp. 621-642. ISSN 1087-6545. DOI 10.1080/08839514.2019.1592343.

[24] Mishra, P., Khan, R. and Baranidharan, D. B. (2020) "Crop Yield Prediction using Gradient Boosting Regression", *International Journal of Innovative Technology and Exploring Engineering*, Vol. 9, No. 3, pp. 2293-2297. ISSN 2278-3075. DOI 10.35940/ijitee.c8879.019320.

[25] Mitra, A., Beegum, S., Fleisher, D., Reddy, V. R., Sun, W., Ray, C., Timlin, D. and Malakar, A. (2024) "Cotton Yield Prediction: A Machine Learning Approach with Field and Synthetic Data", *IEEE Access*, Vol. 12, p. 101273-88. ISSN 2169-3536. DOI 10.1109/ACCESS.2024.3418139.

[26] Molnar, C. (2020) "*Interpretable machine learning*", Lulu.com. ISBN 978-0244768522. [Online]. Available: https://christophm.github.io/interpretable-ml-book. [Accessed: 10 Jan. 10, 2022].

[27] Nguyen, N. and Ngo, D. (2025) "Comparative Analysis of Boosting Algorithms for Predicting Personal Default", *Cogent Economics and Finance,* Vol. 13, No. 1. ISSN 2332 - 2039. DOI 10.1080/23322039.2025.2465971.

[28] Oikonomidis, A., Cagatay Catal, C. and Kassahun, A. (2022) "Hybrid Deep Learning-Based Models for Crop Yield Prediction", *Applied Artificial Intelligence*, Vol. 36, No. 1. ISSN 1087-6545. DOI 10.1080/08839514.2022.2031823.

[29] Pavithra, M. and Soundrapandiyan, R. 2024. "Yield Prediction for Crops by Gradient-Based Algorithms", *PLoS ONE,* Vol. 19, No. 8, pp. 1-20. ISSN 1932-6203. DOI 10.1371/journal.pone.0291928.

[30] Ponce-Bobadilla, A. V., Schmitt, V., Maier, C. S., Mensing, S. and Stodtmann, S. (2024) "Practical Guide to SHAP Analysis: Explaining Supervised Machine Learning Model Predictions in Drug Development", *Clinical and Translational Science*, Vol. 17, No. 11, pp. 1-15. ISSN 1752 - 8062. DOI 10.1111/cts.70056.

[31] Ribeiro, M. T., Singh, S. and Guestrin, C. (2016) ""Why Should I Trust You?": Explaining the Predictions of Any Classifier", In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining,* ACM. pp. 1135-1144. ISSN 2157-0161. DOI 10.1145/2939672.2939778.

[32] Saadio, C. D., Adoni, W. Y. H., Aworka, R., Zoueu, J. T., Mutombo, F. K., Krichen, M. and Kimpolo, C. L. M. (2022) "Crops Yield Prediction Based on Machine Learning Models: Case of West African Countries", *Smart Agricultural Technology,* Vol. 2 (December 2021). ISSN 2772-3755. DOI 10.1016/j.atech.2022.100049.

[33] Samira, S. S., Shahhosseini, M., Huber, I., Hu, G. and Archontoulis, S. V. (2022) "County-Scale Crop Yield Prediction by Integrating Crop Simulation with Machine Learning Models", *Frontiers in Plant Science*, Vol. 13, pp. 1-16. ISSN 1664-462X. DOI 10.3389/fpls.2022.1000224.

[34] Sarr, A. B. and Sultan, B. (2023) "Predicting crop yields in Senegal using machine learning methods", *International Journal of Climatology*, Vol. 43, No. 4, pp. 1817-1838. ISSN 0899-8418. DOI 10.1002/joc.7947.

[35] Sathya, P. and Gnanasekaran, P. (2023) "Paddy Yield Prediction in Tamilnadu Delta Region Using MLR-LSTM Model", *Applied Artificial Intelligence*, Vol. 37, No. 1. ISSN 1087-6545. DOI 10.1080/08839514.2023.2175113.

[36] Shams, M. Y., Gamel, S. A. and Talaat, F. M. (2024). "Enhancing crop recommendation systems with explainable artificial intelligence: a study on agricultural decision-making", *Neural Computing and Applications*, Vol. 36, No. 11, pp. 5695-5714. ISSN 1433-3058. DOI 10.1007/s00521-023-09391-2.

[37] Sharma, P., Dadheech, P., Aneja, N. and Aneja, S. (2023) "Predicting Agriculture Yields Based on Machine Learning Using Regression and Deep Learning", *IEEE Access*, Vol. 11, No. October, pp. 111255-111264. ISSN 2169-3536. DOI 10.1109/ACCESS.2023.3321861.

[38] Sudha, M. K., Manorama, M. and Aditi, T. (2022) "Smart agricultural decision support systems for predicting soil nutrition value using IoT and ridge regression", *AGRIS on-line Papers in Economics and Informatics*, Vol. 14, No. 1, pp. 95-106. ISSN 1804-1930. DOI 10.7160/aol.2022.140108.

[39] Uribeetxebarria, A., Castellón, A. and Aizpurua, A. (2023) "Optimizing Wheat Yield Prediction Integrating Data from Sentinel-1 and Sentinel-2 with CatBoost Algorithm", *Remote Sensing*, Vol. 15, No. 6. ISSN 2072-4292. DOI 10.3390/rs15061640.

[40] Vijayan, S. and Chowdhary, C. L. (2025) "Hybrid feature optimized CNN for rice crop disease prediction", *Scientific Reports*, Vol. 15, No. 1, pp. 1-20. ISSN 2045-2322. DOI 10.1038/s41598-025-92646-w.

Agris on-line Papers in Economics and Informatics

Volume XVIII Number 3, 2026

Market Equilibrium Modelling of Rice in South Kalimantan Using the Nonlinear Programming Method: Implications for Agricultural Policy

Nuri Dewi Yanti, Yudi Ferrianta

Faculty of Agriculture, Lambung Mangkurat University, Banjarbaru, Indonesia

Abstract

Despite being a top global producer, Indonesia's rice market equilibrium is increasingly threatened by demographic pressure, land conversion, and a growing reliance on imports. This study develops a novel Computable General Equilibrium (CGE) model using General Algebraic Modelling Language (GAMS) and nonlinear programming to analyse how regional production responds to shifting policies and market dynamics. The research specifically focuses on South Kalimantan, a region that serves as a strategic food buffer for the new National Capital (IKN). The model demonstrates strong empirical reliability, with a price calibration deviation of only 1.8% and an overall production error of just -0.7%. Simulation results reveal that while overproduction creates regional surpluses, it does not necessarily lead to lower consumer prices. Conversely, government-imposed price controls were found to trigger significant supply deficits by disincentivising producer output. These findings confirm the model's utility as a robust tool for evaluating productivity, infrastructure, and stabilisation policies to ensure national food security.

Keywords

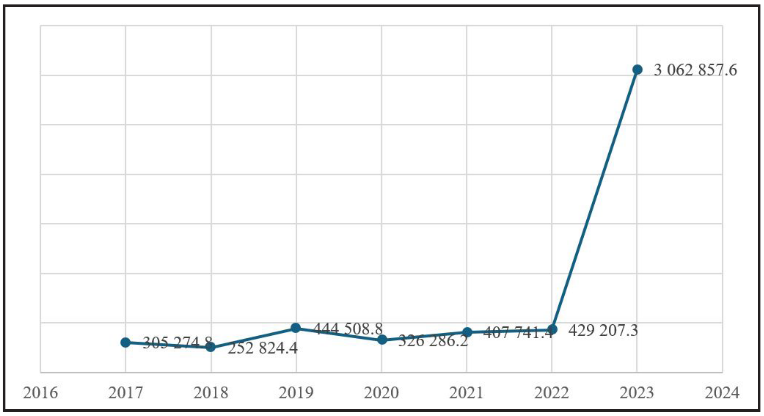
CGE-SAM model, GAMS, rice market equilibrium, South Kalimantan, food security, price control.

Yanti, N. D. and Ferrianta, Y. (2026) "Market Equilibrium Modelling of Rice in South Kalimantan Using the Nonlinear Programming Method: Implications for Agricultural Policy", *AGRIS on-line Papers in Economics and Informatics*, Vol. 18, No. 3, pp. 103-121. ISSN 1804-1930. DOI 10.7160/aol.2026.180308.

Introduction

Indonesia has long been recognised as an agricultural country, where a large proportion of its population relies on farming for their livelihoods. The nation's diverse agroclimatic conditions, characterised by abundant rainfall, year-round sunlight, and fertile volcanic soils, create an environment highly conducive to the development of food crops and plantation commodities. These geographical advantages have positioned agriculture as the backbone of rural economic life and a crucial provider of the national food supply. Thus, it plays a strategic role in the Indonesian economy, not only as a major contributor to the Gross Domestic Product (GDP) but also in providing employment opportunities. Statistics Indonesia indicates that the agricultural sector contributes around 14.35% to the total national GDP while absorbing around 30% of the total Indonesian workforce, either directly or indirectly, while rice is the single-most-important agricultural product (Badan Pusat Statistik, 2025a, 2025b). It is a staple food and the primary source of caloric intake for most Indonesians. As a strategic commodity, rice significantly influences economic stability, political dynamics, and social vulnerability. Historical evidence shows that disruptions in food supplies, particularly rice, sparked unrest and criminal activity during the early stages of the reform era.

As a staple food, the demand for rice in Indonesia increases every year along with the increase in population. However, this rising demand, coupled with shrinking agricultural land and declining rice production, dropping by 1.8% and 1.6% annually (2010 to 2019), threatens market equilibrium (Quartina Pudjiastuti et al., 2021). To meet the demand, Indonesia relies on imports. Indonesia depends on rice imports, thereby exposing local prices to volatility and fluctuations within the global rice market. Although a major producer, Indonesia's rice imports continue to rise. Between 2017 and 2023, import volumes consistently increased (Figure 1), with a dramatic surge in 2023—over seven times higher than the previous year—making Indonesia the world's largest rice importer, accounting



Source: Statistic Indonesia (BPS), 2025

Figure 1: Indonesia rice import data.

for 5.46% of global rice imports valued at USD 32.78 billion (Darmawan et al., 2021; Kusumastuti et al., 2024). In 2024, the situation worsened, with rice imports reaching 4.52 million tons (Rachman, 2025).

Rice imports are intended to stabilise domestic supply and prevent inflation due to shortages (Sholikhah and Anjani, 2023). The various impacts of rice imports have been widely stated in many studies, including the depletion of national foreign exchange reserves and the decline in domestic rice prices, which can lead to reduced competitiveness and negatively affect farmers' welfare. This condition can reduce the motivation of farmers to increase their production, and it can threaten food sovereignty and food security and even invite political pressure from exporting nations (Kusumastuti et al., 2024; Umar, 2025). In fact, the rice market in Indonesia is currently shifted to a fully market-driven one, so the price of rice will be more volatile.

In 2024, Indonesia produced 30.62 million tons of rice from 10 million hectares, yet high per capita consumption (97,6 kg)—surpassing regional neighbours like Malaysia and Brunei—creates a significant supply challenge (Fitrawaty et al., 2023). To support over 284 million citizens, the government must prioritise stable domestic production to ensure food security and minimise import dependency. South Kalimantan is a vital rice production hub outside Java, designated as a strategic food buffer for the new National Capital (IKN). In 2024, the province generated a 28.31% surplus, producing 609,172 tons against a demand of 421,957 tons. However, with a high population growth of 1.29% in the last five years, the government must remain vigilant regarding future rice availability and the potential consequences of market imbalances, including price distortions (Badan Pusat Statistik Provinsi Kalimantan Selatan, 2025).

Rice market equilibrium is influenced not only by production and consumption but also by other factors such as price volatility, climate-induced shifts in production, and import or distribution policies. In recent years, the rice market equilibrium in South Kalimantan has faced various challenges, including price volatility, changes in production patterns due to weather factors, and import and distribution policies that affect supply and demand. Additionally, (Noviar and Affandi, 2019) highlight that government interventions—such as fertiliser subsidies, the government purchasing price, and national rice reserve programs—have been crucial in stabilising rice supply and demand. Therefore, these interrelated factors require thorough analysis.

The study aims to:

1. Analyse how rice production across different regions responds to changes in policies or market conditions.
2. Assess the impact of changes in production and demand on market prices and overall economic equilibrium.
3. Evaluate the effectiveness of government intervention, such as surplus absorption or price controls, in addressing market imbalances.
4. Identify potential regional surpluses or shortages in rice supply under various policy and market scenarios.

Modelling the rice economy: a general equilibrium approach

Economic analysis tools are essential in understanding and forecasting market behaviour. Among these tools, general equilibrium (GE) modelling is widely recognised for analysing interactions between supply and demand across multiple markets simultaneously, whether a stochastic model (Hassan et al., 2025; Tanaka et al., 2021; Tanaka and Guo, 2020) or a dynamic approach (Akune et al., 2015; Hossain et al., 2023; Kunimitsu, 2015; Zhang et al., 2022) or combination approaches (Gao, 2024; Javadi et al., 2022; Vellinga and Tanaka, 2025). The general equilibrium (GE) model provides a comprehensive framework for analysing the interrelationships between markets, effectively capturing the interactions between industries, resources, and institutions. This holistic approach allows for a deeper understanding of how shifts in one area of the economy propagate through other sectors (Ghaith et al., 2021; López and Méndez Segura, 2024; Malahayati and Masui, 2022; Suryadi, 2019; Valera et al., 2024a). Furthermore, Wei and Aaheim (2022) conducted a systematic review of CGE models in climate change adaptation studies, finding that CGE models apply across sectors, including transportation, energy, and investment, yet reviewed studies mainly focus on agriculture, with country-level models most widely adopted.

Numerous studies have applied GE and similar models to agricultural commodities, including rice (Balié and Valera, 2020; Bonga-Bonga and Mbanda, 2025; Hummel and Ziesemer, 2023; Kamei, 2024; Lanzara and Santacesaria, 2023; Robinson et al., 2018). In this context, CGE models systematically integrate production, distribution, and consumption activities, capturing the influence of government policies, market mechanisms, and interactions among economic agents along the supply chain, thereby providing an analytical framework to support policymakers in designing effective policies that enhance welfare outcomes for domestic farmers and consumers (Kunimitsu, 2015; Siswanto et al., 2018). The CGE model is a widely used numerical simulation framework for assessing the impacts of economic shocks and policy changes (Hossain et al., 2023; Okodua, 2018). It can be modified and extended to evaluate the implications of various economic policies, including changes in taxes, subsidies, tariffs, and technological progress, particularly in developing country contexts (Lofgren et al., 2002; Valera et al., 2024a; Widyastutik et al., 2025). Accordingly, this study employs a CGE framework to systematically analyse the structural characteristics and policy dynamics of the rice sector.

Several studies utilising CGE models have demonstrated how price policies, subsidies, distribution systems, and climate change may influence economic equilibrium, particularly within strategic sectors. For instance, Fitrawaty et al. (2023) showed that rising rice prices exacerbate income inequality, especially for smallholder households, without adequate compensation from other sectors. On the other hand, (Balié and Valera, 2020; Vellinga and Tanaka, 2025; Ouraich et al., 2019; Valera et al., 2024b) have shown that trade liberalisation—especially in rice—may reduce consumer prices but harms farmers due to uncompetitive domestic pricing. These studies underscore the importance of policy balance, including subsidies or direct transfers. Similarly, Hoang and Meyers (2015) examined rice trade liberalisation in Southeast Asia (Indonesia, Malaysia, Philippines), revealing that while state-trading enterprises (e.g. Bulog (Indonesia Logistics Bureau) can stabilise prices and enhance efficiency, they may also increase import dependency and affect the income of local farmers. CGE models are also widely employed to analyse the uncertainty of agricultural production shocks due to climate change, which will have a multiplier effect on the economy as demonstrated in studies conducted by (Belford et al., 2023; Bosello et al., 2018; Hossain and Delin, 2019; Javadi et al., 2022).

In this study, we develop a rice market model using mathematical equations within a nonlinear framework to examine how changes in production, demand, and price-related policies affect market equilibrium by integrating spatial dimensions, policies, and uncertainties into a unified quantitative structure. The proposed model is a novel synthesis of classical and modern economic theory, tailored to the characteristics of rice —such as its dependence on geographical factors, policy interventions, and production variability. It incorporates key variables, including production costs, government policy effects, dynamics, and the effects of transportation and subsidies. By employing nonlinear programming, the model better reflects real-world market behaviour by accounting for non-linear constraints and interactions among variables. Designed for use in agrarian countries with geographic diversity and strong government intervention (e.g. Indonesia or Thailand), the model supports spatial economic analysis and provides a practical tool for policymakers to promote food security and improve market efficiency.

Novelty and comparison with previous models

The proposed rice market model introduces a significant breakthrough by integrating the spatial dimension, government policies, and production uncertainties into a single, holistic quantitative framework. This integration is demonstrated through five key aspects, each of which explicitly distinguishes the model from classical approaches while acknowledging the limitations of those earlier frameworks.

1. **Integration of spatial theory and production economics.** The model synthesises Thünen's location theory with Weber's transportation model to analyse how land productivity, market access, and transport costs collectively shape agricultural pricing and location decisions. A key novelty is the resulting Spatial-Dynamic Production Function, which determines regional output based on land area (a(r)), distribution distance (d(r)), and a stochastic component (uniform random variable). By incorporating random variation, the model introduces production uncertainty into spatial economics—a feature rarely addressed in classical literature. This represents a fundamental departure from the classical von Thünen (1966) model, which described a static, circular structure of agricultural production around a single market without multi-region dynamics, complex distance variables, or stochastic shocks. Furthermore, the present framework is scalable, allowing application at both micro-levels (villages or districts) and the national level, contingent upon spatial data availability.
2. **Hybrid production costs: labour and land.** Building upon the widely used Cobb-Douglas cost function, the model has been modified to reflect the specific characteristics of the agricultural sector. By analysing the relationship between labour and land as primary inputs, a distributed cost function (cst(r)) is obtained, combining wages (wl) and land rent (pk(r)), with parameters (β) that vary between regions to reflect differences in input use intensity. The market equilibrium condition (∑(wl·ltot) = ∑(cst(r)·y(r))) links wages directly to regional productivity, an integration that has not been fully explored in traditional Cobb-Douglas applications within spatial CGE frameworks.
3. **Market equilibrium with government intervention.** Market equilibrium in the agricultural sector is frequently distorted by government interventions such as price floors or ceilings. Drawing on Muth's (1961) partial equilibrium model, this framework explicitly incorporates two simultaneous policy instruments: (i) Surplus Purchase (ip(r)), activated when the price falls below the government-set floor price, and (ii) Producer Subsidy (sub(r) = (Price_ceiling−p)·y(r)), applied when prices are administratively controlled below market-clearing levels. This dual-policy structure is a substantial advance over the simple static minimum or maximum price policy proposed by Tweeten (1963), which treated interventions as one-off measures without feedback effects. In the present model, interventions affect not only prices but also production and transportation costs through feedback loops to y(r) and tc(r), thereby capturing the multidimensional impacts of policy on interregional trade flows and resource allocation—a mechanism not previously accommodated in the literature.
4. **Production uncertainty and policy simulation.** The incorporation of stochastic elements in agricultural production is crucial for food security analysis. A uniform random factor embedded in the production function y(r) enables the simulation of extreme climate scenarios or crop failure. This represents a clear departure from the deterministic supply models pioneered by Nerlove (1958), which assumed static production responses and ignored random exogenous shocks such as weather variability. The stochastic component allows the present model to be tested with varying policy parameters —for instance, adjusting the price floor to estimate impacts on market equilibrium, government budgets, and regional rice distribution under conditions of uncertainty.
5. **Multidimensional integration: production, transportation, and policy.** This research develops an integrated framework that synthesises five dimensions often studied separately in the economic literature: (i) spatial production that accounts for geographic location and land productivity;

(ii) Weberian transportation costs that depend on distance and accessibility. Notably, transportation costs are modelled non-linearly (tc(r) = t·d(r)·y(r)), in contrast to the linear or "iceberg" approach introduced by Samuelson (1952), thereby better representing the reality of agricultural commodity distribution where costs depend on both volume and distance; (iii) price and subsidy policies that capture government intervention; (iv) labour market equilibrium linking wages to production sector needs; and (v) production uncertainty, including climate variability and yield risk.

The integration of these five dimensions provides a robust foundation for understanding economic dynamics in a holistic and contextual manner, particularly in strategic sectors such as rice. For example, if the government raises the price floor, the model can simultaneously estimate the resulting increase in government purchases (ip(r)), the impact on subsidy costs (sub(r)), and the consequent changes in labour allocation across regions. By synthesising classical location theory, public economics, and stochastic production, the framework not only contributes to the spatial economics literature but also offers a practical empirical tool for promoting market efficiency and equitable distribution in agrarian economies with pronounced spatial disparities, such as Indonesia and Thailand. The key distinctions between the proposed model and previous classical approaches are summarised in Table 1.

Materials and methods

Conceptual framework

In the economic context, the equilibrium of the rice market is influenced by both supply and demand-side variables, as well as other external factors. These include regional land area and productivity; access to labour, capital, and technology; input costs (e.g. wages and land rents); transportation costs and infrastructure; government subsidies and price support. This study captures these complexities through a nonlinear programming model, which formalises these relationships into a structured system of equations. To ensure a rigorous analysis, mathematical formalisation is used to transform the conceptual economic framework into a structured set of equations for precise empirical testing and scenario simulation. Here are some key parts that are mathematically formalised:

Objective function. In line with the theory of general equilibrium, this model is formulated as a mixed complementarity problem (MCP) rather than a traditional optimization problem. The equilibrium is found by solving a system of inequalities that represent the optimization problems of all agents (producers and consumers). There is no single objective to be maximized (like total output or a single price); instead, the model finds a vector of prices and quantities that clears all markets.

This approach is implemented in GAMS using a nonlinear programming (NLP) formulation to find a feasible point that satisfies all constraints. The artificial objective function is set to a constant (e.g., Objective = 1) * or, in a more theoretically robust implementation, we maximize the aggregate excess demand value to drive the system toward equilibrium. This is a standard technique to solve for an equilibrium in a square system of equations, ensuring that the solution corresponds to a Walrasian equilibrium where supply equals demand in every market. This reformulation corrects the earlier, theoretically weak objective of simply maximizing a single price and aligns the model with standard CGE practice.

Constraints. This model has several constraints, such as cost function, production function, equilibrium of labour and commodity markets, government intervention, and market imbalances.

Aspect	Classical Model	Classic Model References	Developed Model
Spatial	Static (single market centre)	Von Thünen (1826), *Der isolierte Staat* [Location model with a single market centre].	Dynamic (multi-region + distance)
Government Intervention	Minimum or maximum price	Tweeten (1963), *Foundations of Farm Policy* [simple basic price/subsidy policy analysis].	Two instruments + feedback effect
Uncertainty	Deterministic	Nerlove (1958), *The Dynamics of Supply* [deterministic agricultural offering model].	Stochastic (*Uniform*)
Transportation Costs	Ignored or linear	Samuelson (1952), *Spatial Price Equilibrium* [Linear transportation costs "iceberg"].	Non-linear (tc(r) = t·d(r)·y(r)tc(r) = t·d(r)·y(r))

Source: Authors' own summary based on the comparative analysis presented above

Table 1: Comparative aspects of classical vs developed models in agricultural and spatial economics.

The constraints illustrated how production, costs, and prices interact in each region, and how markets might experience shortages or surpluses.

Cost function for each region (Equation 1):

$$def\_cst(r): Ø.cst(r) + 0.002.a(r) + 0.005.d(r) + + 0.0005.y(r) \geq wl.\beta + pk(r).(1-\beta) + 0.0003.ltot \quad (1)$$

- Production function for each region (Equation 2):

$$prf\_y(r): cst(r) + t.d(r) + wp \geq p\text{-}0.006.a(r) \quad (2)$$

The constants appearing in the constraint equations (e.g., 0.002, 0.005, 0.0005 in def_cst(r)) are not arbitrary. They are calibrated parameters resulting from the process of fitting the model to the empirical data presented in the Social Accounting Matrix (SAM) and regional production tables (Table 2 and 4). The calibration process uses a minimum-RMSE (Root Mean Square Error) approach to adjust these coefficients so that the model's output (prices, production, costs) replicates the base -year data (2024-2025) as closely as possible. For instance, the coefficient for a(r) in the cost function represents the marginal cost share of land, derived from the land rent and production data in the SAM.

- Market imbalance (Equation 3):

$$mkt\_imbalance(r): shortage(r) = = max\left(0, \sum_h \alpha(h).ltot - y(r)\right) \quad (3)$$

Following mathematical formalisation, a simulation protocol operationalises the model through baseline calibration. This stage ensures the model replicates current market conditions by setting the price parameter to fluctuate between IDR 8,000 and IDR 20,000, reflecting historical BPS volatility. The total labour endowment (ltot) is fixed at 80, with cost and production parameters initialised to model specifications. By solving the system using standard equilibrium equations without external interventions, the model establishes baseline values for production, market prices, and shortages, providing a benchmark for subsequent scenario analysis.

Moreover, a counterfactual scenario is needed to test how a system would behave under different assumptions or interventions. Some examples of interventions or changes that can be tested are:

***Government intervention*:** If the price of rice falls below the floor price set by the government, intervention will occur. Counterfactual scenarios could involve adjusting the floor price set by the government or changing the intervention mechanism (Equation 4).

$$govt\_intervention(r): ip(r) = = \max\left(0, \frac{(price\_floor - p).shortage(r)}{1.2}\right) \quad (4)$$

Price shocks: If the value of p (rice price) changes beyond its normal range, it simulates how the system reacts to extreme price spikes or drops

Changes in external factors: by introducing changes in parameters such as the transport endowment (trns), the labour endowment (ltot), or the area (a(r)), to see how these factors affect the system.

Each scenario resolves the equilibrium equations under shifted parameters to observe how such changes perturb rice production and pricing, while identifying potential market imbalances and shortages. The primary structural components of the model are as follows:

1. The production function is to determine rice output based on maximum capacity and production factors in each region.
2. A production cost equation to measure the impact of land costs on production prices.
3. Market equilibrium equation to ensure that total production corresponds to total consumption and intervention policies.
4. The equation of government intervention determines when the government needs to buy surplus or set a maximum price.
5. Transportation costs and subsidies consider the influence of geographic distribution and economic incentives on market equilibrium.

By understanding the purpose and logic behind each equation above, it becomes possible to assess the model's capacity to replicate real-world conditions and to evaluate the implications of external interventions or shocks. A brief explanation of each equation is provided below to clarify its function and role within the overall model framework.

1. **Production function in each region.** Rice production in each region depends on the maximum production capacity and other factors such as the area and the distribution distance (EQUATION 5):

$$y(r) = \min(Y{=}Y_{maz}, \max(Y_{min}, C_1 + C_2.ord(r) + \\ + C_3.a(r) - C_4.d(r) + uniform(C_5, C_6))) \quad (5)$$

where:

$y(r)$ is rice production in the region.

$a(r)$ is the area of production.

$d(r)$ is the distribution distance.

C_1.. C_6 is a constant parameter in the production function.

$Uniform(C_5,\ C_6)$ is a random variation factor in production.

2. Production cost function. Production costs in each region are calculated based on labour costs and land ren (Equation 6):

$$cst(r) = \frac{wl.\beta + pk(r).(1-\beta) + C_7.ltot}{\emptyset} - \\ - C_8.a(r) - C_9.d(r) - C_{10}.y(r) \quad (6)$$

where

$cst(r)$ is the cost of production per unit in the region.

wl is the wages of the farmer.

$Pk(r)$ is the land rent in the area.

β is a factor in the distribution of labour and land costs.

$ltots$ is the total available workforce.

$\emptyset$ is the scale of the cost of production.

C_7, C_8, C_9, C_{10} a constant parameters in the function of production cost.

3. Market equilibrium equation. The market must achieve an equilibrium between total production and market demand (Equation 7):

$$p.\sum_r y(r) = \sum_r ip(r) + \sum_r \sum_h \alpha(h).ltot + \\ + C_{11}.\sum_r y(r) \quad (7)$$

where:

p is the market price of rice.

$ip(r)$ is a government intervention in the form of surplus purchases.

$\alpha(h)$ is the proportion of household consumption.

$\sum_r y(r)$ is the total rice production in all regions.

C_{11} is a constant parameter in the market equilibrium.

4. Government intervention when the price falls below the floor price

(Equation 8)

$$ip(r) = \max(0,(P_{floor} - p).shortage(r)/C_{12}) \quad (8)$$

where:

$ip(r)$ is the amount of rice purchased by the government.

$shortage(r)$ is a shortage of supply in the region.

P_{floor} is the minimum price set by the government.

C12 is a constant parameter in government intervention.

5. Labour market equilibrium. The labour market must meet the balance between available labour and that used in production (Equation 9):

$$wl.ltot = \sum_r y(r).\beta.cst(r) + C_{13}.lto \quad (9)$$

where:

wl is the wages of labour.

$\sum_r y(r).\beta.cst(r)$ is the total of labour in production

C_{13} is a constant parameter in the equilibrium of labour

6. Transportation costs. Transportation costs are calculated based on distribution distance and production quantity(Equation 10):

$$tc(r) = t.d(r).y(r) \quad (10)$$

where:

$tc(r)$ is the cost of transportation in the region.

t is the cost per unit of transportation.

$d(r)$ is the distance to the distribution centre.

$y(r)$ is the amount of production.

7. Government subsidies for producers (if price controlled). If the government controls prices below market value, subsidies are given to the producer (Equation 11).

$$sub(r) = \max(0,(price_{ceiling} - p).y(r)) \quad (11)$$

where:

$sub(r)$ is a subsidy given to producers in the region.

$price_{ceiling}$ is the maximum price limit set by the government.

p is the market price of rice.

$y(r)$ is production in the region.

Results and discussion

Model evaluation

The rice market model developed using GAMS successfully represented market equilibrium by integrating multiple economic factors, including regional production, price dynamics, input costs, and government intervention. The original GAMS model had been calibrated using the latest empirical data from Statistics Indonesia (2024-2025) and empirical studies, as shown in Table 2.

Parameter	Value
Rice prod (real)	22,150
- Region 1 (r1)	5,000
- Region 2 (r2)	5,300
- Region 3 (r3)	5,500
- Region 4 (r4)	1,670
- Region 5 (r5)	3,080
- Region 6 (r6)	1,600
Price (real)	Rp 13,641 - Rp 14,000 per kg
Elasticity (empirical result)	
- price elasticity of demand for rice	-0.26 (inelastic)
- supply elasticity	0.3
Cost share: - labour - land	60% (dominated by land preparation and planting) 40% Calibration: phi (Φ) = 0.74 beta (β) = 0.60 via minimize RMSE (y model vs real)
Model scale	*ltot* scaled ~4.4e6 (man-days national proportional),
	a(r) prop real production.

Source: Data processing result, 2025

Table 2: Structural parameters and baseline calibration for the rice sector.

The model integrated empirical data from Statistics Indonesia (2024-2025) with calibrated parameters to replicate real-world market conditions. Total real rice production of 22,150 tons was distributed across six regions, with r_3 (5,500 tons) and r_1 (5,000 tons) serving as the primary hubs. A demand elasticity of -0.26 confirmed rice as a highly inelastic staple, while the supply elasticity was established at 0.3. The cost structure was notably labour-intensive, with labour accounting for 60% of total costs—primarily driven by land preparation and planting—while land constituted 40%. To ensure predictive accuracy, parameters Φ (0.74) and β (0.60) were calibrated by minimising the RMSE between modelled and observed data. The model scaled labour endowment (*ltot*) to approximately 4.4 million man-days, reflecting national proportional employment levels in the agricultural sector. Meanwhile, Table 3 displayed the fully calibrated SAM model, illustrating sectoral interaction, particularly emphasis on rice (rice.mkt).

Following the IFPRI (2023) framework, the Social Accounting Matrix (SAM) in Table 3 establishes a balanced circular flow for the regional economy. The SAM is constructed to capture all transactions between activities, factors, and institutions. Key features include:

1. **Factors:** Labour and land receive income from production activities. Labour income (204 units) is endogenously determined, reflecting the labor-intensity of the sector (β parameter).
2. **Institutions:** Households (farmers, workers, owners) receive factor incomes and allocate them to consumption (of rice and other goods) and savings. Household consumption of rice is derived from their expenditure shares (α), which are calibrated from household survey data.

Account	Labour Inc	Land Inc	HH Farmer	HH Worker	HH Owner	Gov IP	Savings	Rice.mkt Out	Row Total
Labour	0	0	95	68	41	0	0	0	204
Land_r1	0	32	0	0	0	0	0	0	32
(Land r2-r6 similar prop)	...	...	...	...	...	...	...	...	...
HH Farmer	70	0	0	0	0	0	25	0	95
HH Worker	50	0	0	0	0	0	18	0	68
HH Owner	30	0	0	0	0	0	11	0	41
Gov	0	0	5	4	2	0	0	0	11
Savings	54	36	0	0	0	0	0	0	90
Rice.mkt	0	0	71	51	30	11	0	0	163
Total	204	68	171	123	73	11	54	0	704

Source: Data processing result, 2025

Table 3: Calibrated Social Accounting Matrix (SAM) for the rice sector.

3. **Government:** The government collects taxes (not explicitly shown for simplicity) and makes expenditures. The "Gov IP" (Government Intervention Purchases) column represents a policy account that can purchase rice from the market for stabilization purposes. In the baseline calibration, this value is set to ensure consistency.
4. **Macroeconomic closure:** To maintain a consistent macroeconomic framework, the model adopts a savings-driven closure. This means that investment (Savings column) is determined by the total savings generated in the economy (from households, government, and rest of the world). The "Row" (Rest of the World) account balances any external trade. All accounts are balanced row-wise and column-wise, resulting in a total economic flow of 704 units. This closure ensures that the model respects macroeconomic accounting identities and provides a stable basis for counterfactual simulations.

These structural links are further validated by calibration results showing minimal deviations between real and modelled outputs, with an overall production error of only -0.7% and an excellent RMSE of 1.2%.

Analysis of the model results reveals that rice accounts for 23.1% of total economic activity, confirming its role as a strategic commodity and a primary source of vulnerability to production and price shocks. Government intervention, representing 6.7% of the flow, underscores the importance of policy instruments in stabilising the market and transmitting effects to macroeconomic variables like household income. Meanwhile, the labour share of 28.9% reflects the labour-intensive nature of the sector, indicating that shocks in rice production create broad multiplier effects through employment and consumption. Overall, these findings emphasise that micro-level policies in the rice sector carry significant macroeconomic weight, reinforcing the need for integrated policy design.

While the SAM established the underlying structural dependencies of the economy, the empirical robustness of the framework was further validated by its ability to replicate observed production and price data. Table 4 presents these calibration results, demonstrating a high degree of consistency between the model's outputs and actual market conditions across all studied regions.

Region	y real (ton)	y model (ton)	Error (%)
r_1	5,000	4,950	-1.0
r_2	5,300	5,250	-0.9
r_3	5,500	5,450	-0.9
r_4	1,670	1,680	+0.6
r_5	3,080	3,050	-0.97
r_6	1,600	1,620	+1.25
Total	22,150	22,000	-0.7

Source: Data processing result, 2025

Table 4: Empirical and modelled rice production across regions after calibration.

The comparison between real and modelled outputs showed minimal deviations, with errors ranging from -1.0% and +1.25%. At the aggregate level, the total modelled production (22,000 tons) closely approximated the observed value (22,150 tons), yielding an overall error of -0.7 % which indicates that the model was well-calibrated and capable of capturing regional heterogeneity in rice production. The goodness-of-fit statistics further reinforced this conclusion. The RMSE of 1.2% for production was considered excellent, suggesting that the model effectively minimised prediction bias and variance. Moreover, the validation against official BPS data, with deviations within ±2 %, provided strong empirical support for the reliability of the calibration process. Overall, these findings demonstrated that the model was robust and empirically consistent, providing a credible basis for policy simulations on rice sector shocks and interventions.

Price calibration also demonstrated satisfactory accuracy. The modelled rice price of Rp 14,250/kg deviates by only 1.8% from the observed market price, underscoring the model's ability to replicate equilibrium conditions in both quantity and price dimensions. This was particularly important, as rice prices serve as a critical transmission channel between micro-level shocks (production, policy interventions) and macroeconomic outcomes (household welfare, savings, and consumption). The low error rates across regions, the excellent RMSE fit, and the empirical validation against BPS data confirmed that the calibrated SAM-CGE framework provided a reliable basis for analysing policy scenarios.

Sensitivity analysis was conducted to verify the empirical reliability and theoretical consistency of the calibrated CGE-SAM model. By varying key parameters by ±20%, the model's robustness was tested against fluctuations in fundamental economic drivers. As summarised in Table 5, the model produced logical and stable responses in both aggregate output (ΔY Total) and rice prices (ΔP).

Parameter	Symbol	Change	ΔY Total (%)	ΔP (%)
Labour	β	-20%	-12.5	8.2
		20%	11.8	-7.5
Scale cost	φ	20%	15.2	-10.1
		20%	-14.8	9.8
Transport	t	-20%	5.3	-3.2
		20%	-4.9	2.9
Demand more inelastic	ε		3.2	4.5

Source: Data processing result, 2025

Table 5: Sensitivity and robustness of the calibrated CGE-SAM model.

As shown in Table 5, the results highlight that rice market dynamics were most responsive to labour (β) and scale cost (φ) parameters, while demand elasticity (ε) significantly influences price volatility. The stability and predictability of these responses strengthened the model's credibility for policy simulations. Specifically, the high sensitivity to labour and cost parameters confirmed that the framework could evaluate interventions focused on labour productivity, cost efficiency, and infrastructure improvements within the rice sector.

Feasibility analysis

Based on the results of the execution using the GAMS software, the model fulfilled key criteria of sound economic modelling. First, in terms of consistency of market balance, the model showed that production was evenly distributed across different regions without causing extreme inequality. No shortage of supply was found, which indicates that demand had been met optimally. The stability of market prices at the upper limit showed that the model represented market dynamics naturally without requiring direct intervention from the government. Second, from the aspect of the accuracy of the representation of economic factors, the model explicitly considered important variables such as production costs, prices, and transportation costs, all of which played a role in determining production efficiency. The difference in costs between regions in the model reflected realistic market conditions, where regions with better access to resources had a lower cost advantage. Third, in terms of robustness to changing parameters, models were designed with the flexibility to adapt to variations in production, demand, and pricing policies. With the presence of maximum and minimum production limits, this model could function well in situations of increasing or decreasing output, demonstrating structural resilience in the face of dynamic economic scenarios.

Model equilibrium within the GAMS framework

In the context of economic modelling using GAMS, the success of a model was largely determined by the extent to which the equations built accurately represented the relationship between economic variables. In this study, the model developed showed satisfactory performance in this aspect. The market equilibrium functioned effectively, reflecting the mathematical interaction between total production, consumption, and potential government intervention. In addition, the equations representing production costs and use successfully captured realistic relationships; areas with higher access consistently showed lower production cost structures. During the execution process, the model did not show errors or imbalances, indicating that all equations were under relevant economic principles and theories. Overall, the rice market model was suitable for a policy analysis tool as it effectively reflected rice production and distribution dynamics. It reliably explored various scenarios like overproduction, declining demand, or the implementation of price control policies. Each scenario was constructed to evaluate the possible impact of policy interventions or market conditions on the equilibrium of production and prices. The following section presents the results of the model simulations based on predetermined scenarios and parameter settings.

Spatial distribution of rice production. To identify the optimal distribution pattern of rice production, simulations were conducted using a previously developed multi-region spatial model. This model was designed to capture complex inter-regional dynamics by integrating key factors such as transportation costs, resource allocation, subsidy policies, and production uncertainties. Through predetermined scenarios and parameters, this simulation provided an in-depth representation of efficient rice production distribution both spatially and economically. Within this framework, the first scenario examined overproduction, where output was significantly increased to evaluate potential price declines and the necessity of government intervention. The simulation results showed that rice production was not evenly distributed but varied between regions depending on the local characteristics of each region, as shown in Table 6.

Region	Minimum Production (ton)	Actual Production (ton)	Maximum Production (ton)
r_1	18,750	60,000	180,000
r_2	18,750	88,390	180,000
r_3	18,750	63,274	180,000
r_4	18,750	61,948	180,000
r_5	18,750	63,132	180,000
r_6	18,750	63,940	180,000

Note: Remark: r_1 = Banjar, r_2 = Barito Kuala, r_3 = Tanah Laut, r_4 = Hulu Sungai Tengah, r_5 = Tabalong, r_6 = Kotabaru.

Source: Data processing, 2025.

Table 6: Production distribution of each region.

The Table 6 showed that there was a significant variation in the actual production level, although the minimum production limit (18,750 tons) and the maximum production (180,000 tons) of production were set the same in each region. The highest production occurred in the r_2 area (Barito Kuala Regency) with a total production of 88,390 tons, while other areas had lower production, but remained above the set minimum limit (18,750 tons). On the other hand, the r_1 region had the lowest actual production (60,000 tons), although it was still well above the minimum threshold. No region experienced minimum production, which indicated that there were no extreme production constraints at this stage. These findings provided a strong basis for formulating rice production distribution policies that consider resource efficiency and regional food security.

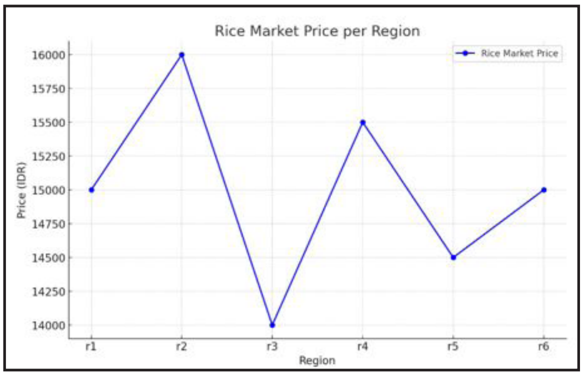
Market prices and government intervention. One of the main indicators in this analysis is the market-clearing price of rice. The simulation results show that the equilibrium price is not fixed at the maximum limit but is determined by the interaction of supply and demand across regions, stabilizing in a range between IDR 14,000 and IDR 16,000 per kg, as detailed in Table 7 and illustrated in Figure 2. This price reflects market conditions where demand is adequately met by supply without causing an imbalance. The model includes a maximum price limit of IDR 20,000 (as a policy constraint) and a minimum price of IDR 8,000, which serve as intervention triggers. Since the equilibrium price is currently above the floor price, no automatic government intervention (i.e., surplus purchases) is activated. Thus, the price system in the model operates according to market logic within the bounds of the defined stabilization policy.

Indicators	Value (IDR/kg)
Equilibrium Market Price (Model Output)	14,000 - 16,000
Minimum price	8,000
Maximum price	20,000

Note: The equilibrium price is determined endogenously by the model based on supply and demand interactions across regions. The minimum and maximum prices are exogenous policy bounds that serve as triggers for government intervention if the market price moves outside this range.

Source: Processed simulation results, 2025.

Table 7: Price price parameters and simulation outcomes.

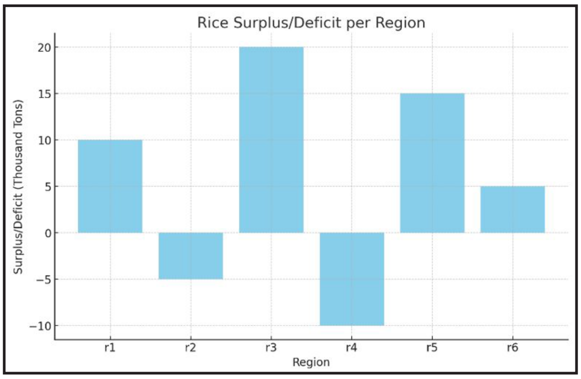


Source: Processed simulation result, 2025.

Figure 2: Rice market price per region in South Kalimantan.

As illustrated in Figure 2, the model captures the spatial variation in rice prices across the six studied regions (r_1 - r_6). Region r_2 exhibits the highest price at IDR 16,000, while r_3 remains at the lower end of the spectrum at IDR 14,000, driven by specific regional production costs or demand intensities. The equilibrium price is endogenous and falls within the policy bounds, demonstrating the model's capacity to simulate market dynamics realistically.

The equilibrium of rice supply and shortage. The model evaluated potential rice shortages as a primary indicator of market equilibrium. The model evaluates the balance between regional production and consumption to identify potential supply-demand gaps. The results show that at the regional level, disparities exist. As visualized in Figure 3, regions r_1, r_3, r_5, and r_6 are identified as surplus areas (production exceeds local demand), while regions r_2 and r_4 are deficit areas (local demand exceeds production). This regional imbalance is a key feature captured by the spatial model.



Source: Processed simulation result, 2025.

Figure 3: Rice production (surplus/deficit) per region in South Kalimantan.

However, despite these regional disparities, the aggregate market achieves equilibrium. Because the model allows for interregional trade and distribution, the surplus from one region can offset the deficit in another. Consequently, at the aggregate market level, there is no unmet demand that would trigger government intervention. This is confirmed in Table 8, which shows that government intervention variables (*ip*(r)) remains at zero. The system relies on internal market mechanisms (i.e., trade flows) rather than external policy intervention to clear the market under the baseline scenario. This finding demonstrates the model's ability to capture both micro-level (regional) and macro-level (aggregate) dynamics.

Region	Government intervention (ip(r))	Regional Deficit (from Figure 3)	Supply shortage ((r))
r_1	0	Surplus	0
r_2	0	Deficit	0
r_3	0	Surplus	0
r_4	0	Deficit	0
r_5	0	Surplus	0
r_6	0	Surplus	0

Note: shortage(r) in the model represents the deficit after accounting for interregional trade flows. A value of 0 confirms that all regional demand is met through the combined effect of local production and inflows from surplus regions, hence no government purchase is required.
Source: Data processing result, 2025.

Table 8: Simulation results in production areas, government intervention, and supply shortage.

The supply-demand balance summarised in Table 8 is visually represented in Figure 3, identifying specific regions facing shortages. This spatial identification was crucial for policymakers to determine where logistics and distribution interventions should be prioritised to mitigate the deficits observed in regions r_2 and r_4.

Figure 3 identifies regions r_1, r_3, r_5, and r_6 as surplus areas and r_2 and r_4 as deficit areas This visual representation underscored the model's findings that while some regions could not meet local demand independently, the overall system achieved equilibrium through potential interregional distribution, thereby negating the immediate need for government intervention in the baseline scenario.

Production costs and other economic factors. The model evaluated relative regional efficiency by calculating production costs per unit, revealing significant disparities across the studied areas. Region r6 was the most efficient with the lowest cost (0.106), likely driven by high productivity combined with superior infrastructure and market access. Region r_2 also demonstrated high efficiency (0.343), correlating with its high output level and suggesting a strong synergy between productivity and cost-effectiveness. In contrast, Region r_3 recorded the highest cost (1.872), indicating potential resource constraints or logistical challenges. Meanwhile, r_1, r_4, and r_5 maintained moderate, stable costs (0.99-1.01). These results accurately reflected the geographical and economic diversity of the rice sector, where proximity to markets and resource availability dictate competitive advantages. Table 9 presents the production cost in each region.

Region	Production Cost per unit (cst(r))
r_1	0.993
r_2	0.343
r_3	1.872
r_4	1.014
r_5	0.992
r_6	0.106

Source: Data processing result, 2025.

Table 9: Production cost in each region.

The analysis concludes that the rice market in South Kalimantan is currently in equilibrium without the need for government intervention. Production was distributed effectively across all regions, remaining above minimum thresholds, while prices were stable at the maximum limit without supply shortages. These findings validated the model as a robust tool for evaluating spatial and economic production efficiency and market stability.

Analysis of the scenario simulation

The model's primary strength lies in depicting a self-sustaining market equilibrium that functions independently of government intervention under baseline conditions. However, to validate the predictive reliability and robustness of the system, the model was assessed through various exogenous shocks. To evaluate this, three specific scenarios—overproduction, demand shock, and price control—were simulated. These scenarios explore how production, prices, and government mechanisms interact, providing a comprehensive understanding of how market equilibrium shifts under varying policy and environmental pressures.

Under the overproduction scenario, output in the producing regions reached the specified upper capacity constraints, resulting in a surplus of 17,750 tons in region r1. Notably, this localised surplus was insufficient to induce a downward price adjustment, with market prices remaining stable at the maximum threshold of IDR 20,000/kg without necessitating regulatory intervention. Similarly, the demand shock scenario, characterised by a sudden contraction in consumption, resulted in a commensurate decline in production and a price depreciation to IDR 18,000/kg. Although marginal surpluses of 3,000 tons were observed in r_1 and r_2, the market demonstrated inherent stability, re-equilibrating without triggering formal intervention mechanisms.

In contrast, the price control scenario introduced significant market distortions. Fixing the price at a sub-equilibrium level of IDR 15,000/kg fundamentally disincentivised production. While aggregate output exceeded that of the demand shock scenario, it remained substantially lower than the overproduction baseline. Crucially, this administrative cap precipitated regional supply deficits ranging from 3,000 to 10,000 tons across regions r_3 through r_6. These empirical results underscore that while the market exhibits resilience toward stochastic fluctuations in supply and demand, rigid price-ceiling policies can inadvertently jeopardise regional supply security. Consequently, such findings suggested that restrictive price policies necessitate secondary, targeted interventions to mitigate regional shortages and maintain systemic equilibrium. These findings highlighted that predetermined price constraints created regional shortages that require targeted intervention to restore equilibrium (Table 10).

Across the three scenarios, the results indicated that overproduction did not automatically translate into price declines but instead tended to generate localised surpluses. While reductions in demand exert downward pressure on prices, this effect remains marginal in the absence of substantial consumption shocks. Conversely, government price controls were shown to trigger supply shortages as producers align their output with binding price ceilings. Overall, these simulations provided critical insights into the complex interplay between production decisions, price formation, and policy interventions in maintaining market equilibrium.

Theoretical justification for the overproduction-price persistence phenomenon

The finding that a massive regional surplus (17,750 tons in r_1) coincides with prices at the maximum policy ceiling (IDR 20,000/kg) warrants careful theoretical interpretation, as it appears to contradict the standard economic intuition that surplus drives prices downward. However, this outcome is not an indication of model insensitivity; rather, it reflects the specific structural and policy-bound characteristics of the rice market that the model is designed to capture. Three interrelated mechanisms explain this phenomenon:

1. The model's cost structure is heavily labour-intensive, with the calibrated β parameter set at 0.60, indicating that 60% of total production costs are attributable to labour. This cost rigidity creates a structural floor for prices. In the overproduction scenario, output expansion is achieved by pushing production toward maximum capacity (180,000 tons), which in the model is realised through intensified labour and land use. The Cobb-Douglas-derived cost function ($cst(r)$) ensures that marginal costs rise nonlinearly as production approaches capacity limits. The increased production, therefore, does not come with lower unit costs; instead, costs remain elevated

Scenario	Production (y(r))	Price (IDR/kg)	Market balance (surplus/shortage)	Government intervention
Overproduction	High (up to upper bound of 200)	20,000	Surplus 17,750 ton (r_1)	None
Demand shock	Lower than overproduction	18,000	Surplus 3,000 ton (r_1 & r_2)	None
Price control	Moderate (price constrained)	15,000	Shortage 3,000-10,000 ton (r_3, r_4, r_5 and r_6)	None

Source: Data processing result, 2025.

Table 10: simulation scenarios and market response dynamics.

due to the labour-intensive nature of the expansion. Consequently, there is no cost-side mechanism to exert downward pressure on market prices, even when physical output is abundant. This finding is consistent with empirical observations in labour-intensive agricultural systems, where the cost structure imposes a natural lower bound on market prices regardless of output volume (Fitrawaty et al., 2023; Balié and Valera, 2020).

2. The model's policy architecture includes a price ceiling (IDR 20,000/kg) that, while intended as an upper bound to protect consumers, also functions as a coordination anchor for market expectations. In the baseline calibration, the equilibrium price already gravitates toward the upper range of the policy band (IDR 14,000-16,000/kg, as shown in Table 7). When the overproduction scenario is activated, the equilibrium price does not decline because it is already at or near the ceiling, and the model's structure—being a Mixed Complementarity Problem (MCP) formulation—solves for a vector of prices that clears all markets simultaneously. The surplus generated in region r_1 is absorbed through interregional trade flows to deficit regions (r_2 and r_4), maintaining aggregate equilibrium. The price does not fall because demand in deficit regions sustains the valuation of rice at the prevailing market-clearing level, which happens to coincide with the upper policy bound. This is consistent with Walrasian equilibrium logic: the price reflects the marginal valuation of the last unit consumed, not the average abundance across all regions. Similar phenomena have been documented in spatial equilibrium models where trade linkages between surplus and deficit regions sustain aggregate price levels despite localised overproduction (Lanzara and Santacesaria, 2023; Samuelson, 1952).

3. From a behavioural perspective, the presence of a binding price ceiling —even if not directly activated—can create a "floor effect" in reverse. In the Indonesian rice market context, where the government is a credible buyer of last resort (via Bulog) and where the ceiling serves as a reference price for market transactions, producers and traders have little incentive to lower prices below this reference point when they know demand in deficit regions remains unmet. The model captures this implicitly through the government intervention variable ($ip(r)$), which, while inactive in the overproduction scenario, exerts a latent influence on market expectations through its mere presence in the policy architecture. The result is a market equilibrium where prices remain at the upper bound despite physical surplus, a phenomenon that has been empirically observed in similar staple food markets with strong government price anchoring. For instance, India's wheat and rice markets under Minimum Support Price (MSP) regimes have demonstrated comparable dynamics, where government price commitments create reference points that stabilise market prices even during periods of surplus production (Valera et al., 2024a; Hoang and Meyers, 2015).

It is important to acknowledge, however, that this result also points to a limitation of the current model specification. The price ceiling in the overproduction scenario acts as an "absorbing" boundary: once the equilibrium price reaches this level, it cannot adjust further upward, but the model's current formulation does not explicitly capture the dynamic path by which prices would naturally decline toward a lower equilibrium if the ceiling were removed. This limitation is inherent in the MCP framework, which identifies equilibrium points that satisfy all constraints simultaneously without tracing the disequilibrium adjustment path (Lofgren et al., 2002). Future iterations of the model should test an alternative scenario in which the price ceiling is relaxed or eliminated entirely, thereby allowing prices to adjust freely to the surplus condition. Such a test would further validate—or refine —the theoretical mechanisms proposed here and provide a more complete understanding of the model's price sensitivity under unrestricted market conditions.

These simulations clarified how market variables dictate the equilibrium between production, pricing, and policy. Future research should explore increasing government intervention thresholds, further lowering price caps, or redistributing regional production to test long-term market stability.

Conclusion

This study found that the rice market equilibrium in South Kalimantan was strongly influenced by production dynamics, demand fluctuations, and government intervention policies. Several key conclusions emerge from the analysis. First, overproduction tended to create regional surpluses rather than driving prices down. The model demonstrated that a surplus of 17,750 tons in a single region did not translate into lower consumer prices, as the price remained at the upper policy bound of IDR 20,000/kg. This finding was justified by three interrelated mechanisms: the labour-intensive cost structure that creates a structural price floor, the absorption of surplus through interregional trade to deficit regions that sustains aggregate price levels, and the anchoring effect of government price policy on market expectations. Second, demand shocks only exerted significant downward pressure on prices when they were extreme in magnitude, indicating that moderate fluctuations in consumption are absorbed efficiently by the market mechanism. Third, rigid price controls were found to trigger significant supply deficits, as producers rationally adjusted their output downward to align with the regulated price ceiling, thereby jeopardising regional supply security.

These findings carry important policy implications. More effective stabilisation can be achieved through surplus absorption and targeted subsidies, which simultaneously protect consumer welfare and maintain producer incentives. In contrast, rigid price-ceiling policies, while intended to protect consumers, inadvertently create supply shortages that undermine food security objectives. Rising population growth, land conversion pressures, and climate variability remain pressing risks for future supply stability in South Kalimantan, particularly given its strategic role as a food buffer for the new National Capital (IKN).

This study also acknowledges that the finding of high prices coinciding with large surpluses under the price ceiling regime, while theoretically justifiable within the model's structural framework, warrants further investigation. Relaxing the price ceiling constraint in future model iterations would allow for the testing of whether prices would naturally decline under pure market clearing conditions, thereby providing a more complete picture of the model's price sensitivity. Additionally, the current model specification does not explicitly capture the dynamic adjustment path by which prices converge to equilibrium; future research incorporating dynamic stochastic elements would enhance the model's capacity to trace disequilibrium trajectories.

By incorporating spatial diversity, production uncertainty, and policy feedback effects, the nonlinear CGE model developed in this study offers policymakers a practical tool to anticipate market imbalances and design proactive interventions that strengthen food security, sustain farmer motivation, and ensure fair distribution across regions.

As a recommendation, to maintain rice market stability in South Kalimantan, policymakers should focus on three priority areas:

1. **Surplus absorption and targeted subsidies.** Absorbing surplus production through government procurement mechanisms (e.g., Bulog operations) and providing targeted subsidies to producers, rather than relying on rigid price controls, since these measures are more effective in protecting both farmers and consumers from price volatility while maintaining production incentives.
2. **Investment in resilient infrastructure**. Investing in resilient farming systems, climate-adaptive technologies, and improved transport infrastructure to prepare for rising demand, population growth, and increasing climate pressures. This includes irrigation modernisation, drought-resistant seed varieties, and enhanced logistics networks to reduce transportation costs and post-harvest losses.
3. **Adoption of nonlinear CGE modelling as a regular policy tool.** Institutionalising the use of nonlinear CGE modelling to simulate market dynamics, anticipate potential risks, and guide proactive interventions. This evidence-based approach will strengthen food security planning, sustain farmer motivation, and ensure equitable distribution across regions under various policy and environmental scenarios.

Corresponding author:
Ir. Nuri Dewi Yanti, M.Sc., Ph.D.
Faculty of Agriculture, Lambung Mangkurat University
Jalan Ahmad Yani KM 36, Banjarbaru, Indonesia 70714
E-mail: ndyanti@ulm.ac.id

References

[1] Akune, Y., Okiyama, M. and Tokunaga, S. (2015) "Economic evaluation of dissemination of high temperature-tolerant rice in Japan using a dynamic computable general equilibrium model", *Japan Agricultural Research Quarterly*, Vol. 49, No. 2, pp. 127-133. ISSN 00213551. DOI 10.6090/JARQ.49.127.

[2] Badan Pusat Statistik (2025a) "*Keadaan Angkatan Kerja di Indonesia, Agustus 2025*", Jakarta, Indonesia, Badan Pusat Indonesia. [Online]. Available: https://www.bps.go.id/id/publication/2025/12/19/42a75ee61332755586fdfcdd/keadaan-angkatan-kerja-di-indonesia-agustus-2025.html%20%20Ini%20yg%20bos%20bang [Accessed: Nov. 10, 2025]. (In Indonesian).

[3] Badan Pusat Statistik (2025b) "Provinsi Kalimantan Selatan dalam Angka, Kalimantan Selatan", Badan Pusat Statistik Provinsi Kalimantan Selatan. [Online]. Available: https://kalsel.bps.go.id/id/publication/2025/02/28/dc342b31a1e2000c0b099c56/provinsi-kalimantan-selatan-dalam-angka-2025.html[Accessed: Nov. 10, 2025]. (In Indonesian).

[4] Badan Pusat Statistik (2025c) "Statistik Indonesia 2025, Badan Pusat Statistik". [Online]. Available: https://www.bps.go.id/id/publication/2025/02/28/8cfe1a589ad3693396d3db9f/statistik-indonesia-2025.html [Accessed: Nov. 10, 2025]. (In Indonesian).

[5] Balié, J., Minot, N. and Valera, H. G. A. (2021) "Distributional impacts of the rice tariffication policy in the Philippines", *Economic Analysis and Policy*, Vol. 69, pp. 289-306. ISSN 0313-5926. DOI 10.1016/J.EAP.2020.12.005.

[6] Balié, J. and Valera, H. G. (2020) "Domestic and international impacts of the rice trade policy reform in the Philippines", *Food Policy*, Vol. 92. ISSN 0306-9192. DOI 10.1016/j.foodpol.2020.101876.

[7] Belford, C., Huang, D., Nasr, Y., Ceesay, E. K. and Sanyang, L. (2023) "An economic assessment of the impact of climate change on the Gambia's agriculture sector: a CGE approach", *International Journal of Climate Change Strategies and Management*, Vol. 15, No. 3, pp. 322-352. ISSN 1756-8692. DOI 10.1108/IJCCSM-01-2022-0003.

[8] Bonga-Bonga, L. and Mbanda, V. (2025) "Computable general equilibrium-microsimulation analysis of electricity price increases in South Africa", *Utilities Policy*, Vol. 95, pp. 101936. ISSN 0957-1787. DOI 10.1016/j.jup.2025.101936.

[9] Bosello, F., Campagnolo, L, Cervigni, R, and Eboli, F. (2018) "Climate Change and Adaptation: The Case of Nigerian Agriculture", *Environmental and Resource Economics*, Vol. 69, No. 4, pp. 787-810. ISSN 1573-1502. DOI 10.1007/S10640-016-0105-4.

[10] Fitrawaty, Germawan, W., Yusuf, M. and Maipita, I. (2023) "A simulation of increasing rice price toward the disparity of income distribution: An evidence from Indonesia", *Heliyon*, Vol. 9, No. 3. ISSN 2405-8440. DOI 10.1016/j.heliyon.2023.e13785.

[11] Gao, S. (2024) "An Exogenous Risk in Fiscal-Financial Sustainability: Dynamic Stochastic General Equilibrium Analysis of Climate Physical Risk and Adaptation Cost", *Journal of Risk and Financial Management*, Vol. 17, No. 6. ISSN 1911-8074. DOI 10.3390/jrfm17060244.

[12] Ghaith, Z., Kulshreshtha, S., Natcher, D. and Cameron, B. T. (2021) "Regional Computable General Equilibrium models: A review", *Journal of Policy Modeling*, Vol. 43, No. 3, pp. 710-724. ISSN 0161-8938. DOI 10.1016/J.JPOLMOD.2021.03.005.

[13] Hassan, C. O. , Fasanu, E. A., Matthew, O. A. and Hassan, T. M. (2025) "Effect of Climate Change on Food Security Using Computable General Equilibrium Approach”", Institute of Physics. *IOP Conference Series: Earth and Environmental Science, Vol. 1492, 8th International Conference on Science and Sustainable Development (ICSSD 2024) Nov. 13 - 15, 2024*, Ota, Nigeria. ISSN 1755-1315. DOI 10.1088/1755-1315/1492/1/012040.

[14] Hoang, H. K. and Meyers, W. H. (2015) "Price stabilization and impacts of trade liberalization in the Southeast Asian rice market", *Food Policy*, Vol. 57. ISSN 0306-9192. DOI 10.1016/j.foodpol.2015.07.009.

[15] Hossain, S. S., Cui, Y., Delin, H. and Zhang, Y. (2023) "The economic influence of climate change on Bangladesh agriculture: application of a dynamic computable general equilibrium model", *International Journal of Climate Change Strategies and Management*, Vol. 15, No. 3, pp. 353-370. ISSN 1756-8692. DOI 10.1108/IJCCSM-10-2021-0123.

[16] Hossain, S. S. and Delin, H. (2019) "Rice and Wheat Tariff Impact in Bangladesh: CGE Analysis Using Gtap Model", *Journal of Agricultural Science*, Vol. 11, No. 10, pp. 63 - 63. ISSN 1916-9752. DOI 10.5539/JAS.V11N10P63.

[17] Hummel, A. J. and Ziesemer, V. (2023) "Food subsidies in general equilibrium", *Journal of Public Economics*, Vol. 222. ISSN 0047-2727. DOI 10.1016/j.jpubeco.2023.104882.

[18] Javadi, A., Ghahremanzadeh, M., Sasi, M., Javanbakht. and Hayati, B. (2022) "Economic Evaluation of the Climate Changes on Food Security: Application of CGE Model", Springer Nature. DOI 10.21203/rs.3.rs-1286503/v1. [Online]. Available: https://www.researchsquare.com/article/rs-1286503/v1. [Accessed: Jan. 15, 2026].

[19] Kamei, K. (2024) "A model of economic integration with ownership structure in general equilibrium oligopoly", *Economics Letters*, Vol. 236. ISSN 0165-1765. DOI 10.1016/j.econlet.2024.111622.

[20] Kunimitsu, Y. (2015) "Regional impacts of long-term climate change on rice production and agricultural income: Evidence from computable general equilibrium analysis", J*apan Agricultural Research Quarterly*, Vol. 49, No. 2. ISSN 00213551. DOI 10.6090/jarq.49.173.

[21] Kunimitsu, Y., Iizumi, T. and Yokozawa, M. (2014) "Is long-term climate change beneficial or harmful for rice total factor productivity in Japan: Evidence from a panel data analysis", *Paddy and Water Environment*, Vol. 12, No. 2, pp. 213-225. ISSN 1611-2504. DOI 10.1007/S10333-013-0368-0.

[22] Kusumastuti, A. I., Indriani, S. A. and Febriyani, T. (2024) "Dampak Maraknya Impor Beras di Indonesia dalam 5 Tahun Terakhir terhadap Kesejahteraan Petani Padi", *JINTAN : Jurnal Ilmiah Pertanian Nasional*, Vol. 4, No. 1, pp. 78-88. ISSN 2775-4987. DOI 10.30737/jintan.v4i1.5312. (In Indonesian).

[23] Lanzara, G. and Santacesaria, M. (2023) "Market areas in general equilibrium", J*ournal of Economic Theory*, Vol. 211. ISSN 1095-7235. DOI 10.1016/j.jet.2023.105675.

[24] Lofgren, H., Lee, H. R. and Sherman, R. (2002) "*A Standard Computable General Equilibrium (CGE) Model in GAMS*", Microcomputers in Policy Research 42483, CGIAR, International Food Policy Research Institute (IFPRI). ISBN 0-89629-720-9. DOI 10.22004/ag.econ.42483.

[25] López, E. C. and Méndez Segura, E. (2024) "General Equilibrium Models in the Economy and in the Field of Development Policy", *International Journal of Economics, Business and Management Research*, Vol. 08, No. 03, pp. 129-138. ISSN 2456-7760. DOI 10.51505/ijebmr.2024.8309.

[26] Malahayati, M. and Masui, T. (2022) "Potential impact of the adoption of food loss reduction technologies in Indonesia", *Journal of Environmental Management*, Vol. 319, pp. 115633. ISSN 0301-4797. DOI 10.1016/J.JENVMAN.2022.115633.

[27] Nerlove, M. (1958) "*The Dynamics of supply: Estimation of farmers' response to price*", Baltimore, Maryland, Johns Hopkins Press.

[28] Noviar, H. and Affandi, A. (2019) "Analisis Impor Beras Dan Keseimbangan Pasar Beras Indonesia", *Ekombis: Jurnal Fakultas Ekonomi*, Vol. 5, No. 2. ISSN 2355-0627. DOI 10.35308/ekombis.v5i2.1364. (In Indonesian).

[29] Okodua, H. (2018) "Assessing the Impact of Rice Sector Policy Reforms on the Income Mobility of Rural Households in Nigeria", *The Open Agriculture Journal*, Vol. 12, No. 1, pp. 174-184. ISSN 1874-3315. DOI 10.2174/1874331501812010174.

[30] Ouraich, I., Dudu, H., Tyner, W. E. and Cakmak, E. H. (2019) "Agriculture, trade, and climate change adaptation: a global CGE analysis for Morocco and Turkey", *Journal of North African Studies,* Vol. 24, No. 6, pp. 961-991. ISSN 1743-9345. DOI 10.1080/13629387.2018.1463847.

[31] Pudjiastuti, A. Q., Arisena, G. M. K. and Krisnandika, A. A. K. (2021) "Rice Import Development in Indonesia", *SOCA: Jurnal Sosial, Ekonomi Pertanian*, Vol. 15, No. 2. ISSN 1411-7177. DOI 10.24843/soca.2021.v15.i02.p14.

[32] Rachman, A. (2025) "*Impor Beras RI Sepanjang 2024 Tembus 4,52 Juta Ton, Ini Pemasoknya!*", CNBC Indonesia. [Online]. Available:https://www.cnbcindonesia.com/news/20250116085359-4-603570/impor-beras-ri-sepanjang-2024-tembus-452-juta-ton-ini-pemasoknya [Accessed: Nov. 10, 2025]. (In Indonesian).

[33] Sabarella (2024) "*Analisis kinerja perdagangan beras*", Jakarta, Pusat Data dan Sistem Informasi Pertanian Kementerian Pertanian. [Online]. Available: https://satudata.pertanian.go.id/details/publikasi/849 [Accessed: Nov. 10, 2025]. (In Indonesian).

[34] Samuelson, P. A. (1952) "Spatial Price Equilibrium and Linear Programming", *The American Economic Review*, Vol. 42, No. 3, pp. 283-303. ISSN 0002-8282.

[35] Robinson, S., El-Said, M., San, N. N., Suryana, A., Swastika, D. and Bahri, S. (1997) "Rice price policies in Indonesia: A computable general equilibrium (CGE) analysis", TMD Discussion Paper 19. [Online]. Available: https://hdl.handle.net/10568/161178 [Accessed: Jan. 12, 2026].

[36] Sholikhah, M. and Anjani, M. D. (2023) "Kebijakan Pemerintah dalam Mengatasi Kenaikan Harga Beras di Indonesia", *Journal of Economics and Social Sciences (JESS)*, Vol. 2, No. 2. ISSN 2830-5337. DOI 10.59525/jess.v2i2.311.

[37] Siswanto, E., Sinaga, B. M. and Harianto (2018) "Dampak Kebijakan Perberasan pada Pasar Beras dan Kesejahteraan Produsen dan Konsumen Beras di Indonesia", *Jurnal Ilmu Pertanian Indonesia*, Vol. 23, No. 2. ISSN 0853-4217. DOI 10.18343/23.2.93.(In Indonesian).

[38] Suryadi (2019) "Model Keseimbangan Umum Perekonomian Indonesia", Perubahan Harga dan Produksi Padi terhadap Kinerja Ekonomi Sektoral dan Kesejahteraan, Sidoarjo, Delta Pijar Khatulistiwa. ISBN 978-623-92301-3-5. (In Indonesian).

[39] Tanaka, T., Geyik, Ö. and Karapinar, B. (2021) "Short-term Implications of Climate Shocks on Wheat-Based Nutrient Flows: A global ‘Nutrition at Risk’ Analysis through a Stochastic CGE model", *Foods*, Vol. 10, No. 6. ISSN 1873-5657. DOI 10.3390/foods10061414.

[40] Tanaka, T. and Guo, J. (2020) "Realistic Specifications and Model Predictability: Testing the Performance of a Stochastic CGE Model with Regionally Correlated Yield Variability in the Wheat and Rice Sector", *Asian Journal of Economic Modelling*, Vol. 8, No. 1, pp. 55-75. ISSN 2312-3656. DOI 10.18488/journal.8.2020.81.55.75.

[41] Umar, M. K. G. (2025) "Dampak Impor Beras Terhadap Harga Eceran Tertinggi Beras Lokal Indonesia", *Jurnal Ilmiah Ekonomi Dan Manajemen*, Vol. 3, No. 2, pp. 146-153. ISSN 3025-7859. DOI 10.61722/jiem.v3i2.3809. (In Indonesian).

[42] Valera, H. G. A., Mishra, A. K., Pede, V. O., Yamasno, T. and Dawe, D. (2024a) "Domestic and international impacts of rice export restrictions: The recent case of indian non-basmati rice", *Global Food Security*, Vol. 41, p. 100754. ISSN 2211-9124. DOI 10.1016/J.GFS.2024.100754.

[43] Valera, H. G. A., Gopalakrishnan, B. N., Chakravarthy, S., Bharathi, S., Balié, J. and Pede, V. O. (2024b) "The impacts of reforming agricultural policy support on cereal prices: a CGE modeling approach", *Journal of Economic Studies*, Vol. 51, No. 1, pp. 202-221. ISSN 01443585. DOI 10.1108/JES-06-2022-0362.

[44] Vellinga, N. and Tanaka, T. (2025) "Updating the analysis of rice import liberalization in Japan: a dynamic-stochastic CGE modelling approach", *Applied Economics*, Vol. 57, No. 6, pp. 690-708. E-ISSN 1466-4283. DOI 10.1080/00036846.2024.2305617.

[45] Wei, T. and Aaheim, A. (2022) "Climate change adaptation based on computable general equilibrium models - a systematic review", *International Journal of Climate Change Strategies and Management*, Vol. 15, No. 4, pp. 561-576. ISSN 1756-8692. DOI 10.1108/IJCCSM-03-2022-0031.

[46] Widyastutik, Meliany, B. S., Amaliah, S., Hotsawadi and Rifin, A. (2025) "Linking Global CGE Models and Sectoral Analysis to Evaluate the Impact of Trade Openness in Service Sector Towards Indonesia Agricultural and Agroindustry", *Economies*, Vol. 13, pp. 1-18. ISSN 2227-7099. DOI 10.3390/economies13070199.

[47] Zhang, X., Wu, L., Ma, X. Qin, Y.(2022) "Dynamic computable general equilibrium simulation of agricultural greenhouse gas emissions in China", *Journal of Cleaner Production*, Vol. 345, p. 131122. ISSN 0959-6526. DOI 10.1016/J.JCLEPRO.2022.131122.

Agris on-line Papers in Economics and Informatics

Volume XVIII Number 3, 2026

Food Inflation and Poverty in Indonesia: How Regional Disparities Shape Economic Hardship

Misbahol Yaqin, Naufal Alfaraby Winandita[2]

[1] Department of Economics, Faculty of Economics and Business, University of Jember, Indonesia

[2] Department of Economics, Faculty of Economics and Business, Diponegoro University, Semarang, Indonesia

Abstract

We examine the impact of rising food prices on poverty in Indonesia by considering regional characteristics. Using panel data from 34 provinces in Indonesia from 2013 to 2024, we apply the Generalized Method of Moments (GMM) to estimate the relationship between food price inflation and poverty rates. We find that rising food prices significantly increase poverty levels. The results are consistent across models. The impact is more severe in rural areas compared to urban areas, indicating greater vulnerability of rural households to food price fluctuations. Moreover, we find that in Western Indonesia which are a more developed regions experiences a lower poverty impact from rising food prices compared to Eastern Indonesia. These findings highlight the need to stabilize food prices to support inclusive development. Strengthening regional food security, ensuring targeted funding, and implementing effective food subsidies are essential to mitigate the impact of food price inflation on low-income populations.

Keywords

Food price inflation, poverty, rural poverty, food security.

Yaqin, M. and Winandita, N. A. (2026) "Food Inflation and Poverty in Indonesia: How Regional Disparities Shape Economic Hardship", *AGRIS on-line Papers in Economics and Informatics*, Vol. 18, No. 3, pp. 123-135. ISSN 1804-1930. DOI 10.7160/aol.2026.180309.

Introduction

Food prices on the global market have experienced significant fluctuations over the last decade. A sharp rise in food prices occurred in 2007, followed by a drastic decline in 2008 following the Global Financial Crisis (GFC). In 2010, food prices increased again and reached their highest peak in 2011 (Rivai, 2022). Most recently, the increase in food prices occurred in 2021 due to geopolitical clash in Russia and Ukraine compounded by prolonged Covid-19 pandemic (Adjemian et al., 2023; Monge and Lazcano, 2022; Suryahadi et al., 2020). According to FAO data, the mean for food price index is projected to reach 143.7 points in 2022, marking a 14.3% rise from 2021 and recording the highest level since commenced in 1990. This notable surge in food prices exacerbates the enduring repercussions of the Covid-19 pandemic crisis which leads to constraining the growth of developing nations.

In the same period, the inflation rate in Southeast Asia Countries including Indonesia was relatively high triggered by rising food prices. In 2022, food inflation reached 5.61% (yoy), above headline inflation of 5.51% (yoy). In the last ten years, the average food inflation in Indonesia has also reached 5.80%. The predominant cause of the elevated inflation rate in recent years has been the escalation in food prices. This surge can be attributed primarily to a decrease in agricultural production resulting from adverse weather conditions and a range of supply constraints (Affandi, 2015; Anugrah and Pratama, 2018; Pratikto and Ikhsan, 2016). Regarding demand dynamics, population growth and heightened remittance flows are driving an uptick in the demand for cereal grains. Consequently, this heightened demand is exerting pressure on prices within the market.

Increasing food prices have a real impact on social economics, including poverty. Increasing food prices can cause community vulnerability especially in poor households because it can reduce income and disrupt consumption patterns (Bhattacharya, 2017; Fujii, 2013; Wu and Xu, 2021). The proportion of food consumption in Indonesia's households also has the largest portion, reaching around 50.41% in 2022. An increase in food prices has the potential for welfare in society. Increased spending on food due to price increases can also result in reduced funding allocations for the health and education sectors. Moreover, this could hinder investment in crucial elements of agriculture, such as fertilizer, fuel, and electricity, which are essential factors in increasing food production (Fujii, 2013; Hammoudeh et al., 2015; Samal et al., 2022).

The impact of rising food prices on social economics such as poverty is still a matter of debate in empirical level. Studies conducted by Ivanic et al. (2012); Shrestha and Chaudhary (2012); Fujii (2013); Warr and Yusuf (2014); Misdawita et al. (2019) find the influence of food prices on poverty dynamics in several countries and mostly found a negative impact between the two. This means that rising food prices drive up poverty levels. Food inflation has a greater impact on poverty when the population spends more on food than any other bundles (Shrestha and Chaudhary, 2012). However, the impact is smaller when the share of expenditure on food consumption is relatively low. In contrast, Headey and Hirvonen (2023) show that increasing food prices has a positive impact on reducing poverty levels in 33 middle-income countries. The escalation in food prices implies that numerous individuals from low-income backgrounds are involved in either producing or marketing food. The higher prices act as incentives for augmenting food production, thereby potentially increasing overall food supply.

In Indonesia, studies examining the impact of rising food prices on poverty and unemployment are still relatively limited. Faharuddin et al. (2023) dan Ivanic et al. (2012) investigated the effects of rice price increases on poverty in Indonesia, while (Warr and Yusuf, 2014) focused on the poverty impact using international prices for six food commodities. In more recent study, Misdawita et al. (2019) employed a social accounting matrix to analyze the poverty implications of four food commodity prices (rice, corn, soybeans, and sugar). Their collective findings suggest that most households in Indonesia including those residing in rural areas suffer from a declining household welfare and increasing poverty due to rising rice prices.

Although several previous studies have studied the impact of food prices on poverty, no study has studied its impact on a regional scale. Differences in regional characteristics such as urban and rural as well as eastern and western Indonesia have the potential to have different implications. Therefore, our study aims to review the impact of food inflation on poverty in Indonesia by reviewing regional aspects. This will contribute to literature in three ways. First, we explore a potential impact of food inflation on poverty in Indonesia. We also try to explore further and seek the effect of food inflation on poverty based on regional development aspect. This allowed us to see the magnitude of the impact in more developed and less developed regions. Second, we continue the exploration into non-linear effect of food inflation on poverty. Third, this research will study the impact on poverty by considering richer provincial-level data, which is available for 34 provinces in Indonesia.

Literature review and hypotheses construction

The correlation between food inflation and poverty has been extensively explored across a diverse range of literature. Most of the literature explains that rising inflation has a negative impact on society, especially worsening development problems such as poverty. Elevated food prices naturally reduce the disposable income of the poor, given their substantial allocation of earnings towards food expenses (e.g., 50% or more for the most financially vulnerable). Additionally, even short-run income disturbances can have lasting negative repercussions, particularly affecting long-term nutritional and health outcomes.

Several previous studies, including the investigation conducted by Shrestha and Chaudhary (2012) on the relationship between food inflation and poverty in Nepal, consistently affirm the detrimental impact of escalating food prices on poverty levels. Their findings indicate that a 10 percent increase in food prices could lead to a 4% rise in overall poverty in Nepal. Moreover, their research suggests that such a price surge could result in 1 million more individuals falling into poverty, with 6.7 million poor people already facing economic challenges. Hence, policymakers must prioritize efforts to mitigate food price inflation and ensure the availability of an ample food supply.

Kıroğlu and Sezgin (2021) examined the association between food inflation and poverty in Turkey from 2016 to 2019, utilizing data from the Household Budget Survey. They found that rising food prices appear to increase poverty in regional Turkey. People's behaviour that tends to be consumerist, encourages quite large changes in their welfare when food inflation increases. In the same direction, Ivanic et al. (2012) analysed how price changes for 38 agricultural commodities affected poverty across 28 countries.

They found that a 1% increase in food prices led to a 1.1 percentage rise in poverty in low-income countries and a 0.7 percentage point rise in middle-income countries.

Valero-Gil and Valero (2008) evaluated the effects of rising food prices from 2006 to 2008 on poverty and extreme poverty rates in Mexico. Their study revealed a slight increase in poverty when comparing 2006-2007 prices, but a significantly larger impact on the poor when considering 2008 prices. After incorporating the positive impacts of public policies implemented in 2008, such as reduced taxes and tariffs on food items and increased subsidies for the most vulnerable, the poverty rate measured by consumption rose from 25% to 33.5%. Moreover, extreme poverty increased from 10.58% to 16%, accounting for the rise in food prices.

Although most researchers found a negative impact of rising food prices on poverty, Headey and Hirvonen (2023) study found something different. They found that increasing food prices had a positive impact on reducing poverty levels in 33 Middle-Income Countries. Panel regression analysis shows that yearly rises in real food prices correlate with a decreasing in poverty rate at US$ 3.20 per-day poverty line, except in more urban or non-agrarian countries. This is likely due to increased agricultural activity in response to higher prices, leading to greater demand for unskilled labor and higher wages. In addition, higher food prices incentivize poor people that has connection in food production or marketing to expand their efforts, thereby increasing food production.

Headey and Hirvonen (2023) findings are in line with studies from Headey (2018). Headey (2018) investigated revealing a reduction in poverty levels concurrent with food price inflation. This phenomenon can be elucidated empirically by the pronounced responsiveness of agricultural supply and wages to escalating food prices, compounded by the predominant reliance of the global impoverished population on agriculture or other related efforts for sustenance. The escalation in food prices caused a surge in income, notably among individuals entrenched in the agricultural sector.

The impact of food inflation on poverty varies between rural and urban regions, with distinct consequences for rural households. In rural areas, rising food prices may lead to increased income for food-producing households; however, this benefit is relatively small compared to the overall challenges posed by inflation (Faharuddin et al., 2023; Prayoga et al., 2025). For instance, in Indonesia, the income rise for these households is not sufficient to offset the general poverty increase observed during food price crises, as overall poverty in rural areas rose significantly, although less dramatically than in urban contexts (Faharuddin et al., 2023; Shahbaz et al., 2022). Moreover, food inflation exacerbates food insecurity, limiting access to nutritious food and heightening malnutrition rates, particularly among vulnerable populations in rural settings. The increased cost of food can diminish purchasing power, especially among net food consumers who lack sufficient income or resources, further entrenching poverty (Prayoga et al., 2025). Consequently, addressing food price inflation through supportive policies is critical to enhancing food security and alleviating poverty in these regions (Luster et al., 2023; Xi et al., 2025)

Nevertheless, we decided to develop hypotheses based on the first side of the previous literature. We thought that Indonesia has still faces several economic challenges. Such escaping the middle-income trap and the vulnerable people hanging around poverty line would be sufficient to think that any form of increase in food commodities price will lead to an elastic change in demand for those commodities[1]. Therefore, we expect that increasing food inflation in Indonesia will decrease the purchasing power of vulnerable people and consequently stumbled them into poverty. In addition to that, we also explore a possible non-linear effect of food inflation on poverty in Indonesia. We believe that there is an optimal range of food inflation for achieving a very low increasing response of the poverty rate in Indonesia.

The second thought is we also expect that the rising of food inflation in Indonesia would affect the urban and rural poverty in a different degree of effect. In Indonesia, we assume that rural people are more vulnerable than urbans because the poverty rate in rural is higher than in urban[2]. Thus, we expect that food inflation will affect more to rural poverty than urban poverty. Another point of view is we try to seek the effect on food inflation to poverty in Indonesia with respect to regional development aspect. We believe that Indonesia still faces struggles in achieving equal growth

[1] Recent report by LPEM FEB UI, 2025 in *Indonesia Economic Outlook Q3-2025* says that until 2024, around 9% of Indonesian categorized as "Poor", 24% of Indonesian categorized as "vulnerable", and 49% categorized as "aspiring middle-class". Leaving only one-fifth of Indonesian that categorized as middle-class and upper-class.

[2] As of 2024, Central Bureau of Statistics Indonesia reported that 6.66% people lived in poverty in urban areas, while 11.34% people lived in poverty in rural areas.

and development across all regions. In Indonesia, it is obvious to see the unequal development between Western and Eastern Indonesia[3]. Hence, because of the unequal development across regions, we expect that food inflation will harm more in Eastern Indonesia rather than in Western Indonesia.

Materials and methods

Data

We examine the impact of food price inflation on poverty Indonesia. The volatility of inflation has significant implications for Indonesia's development, particularly in relation to poverty. These two indicators are critical targets outlined in the National Medium-Term Development Plan (RPJMN) 2020-2024[4]. To achieve the research objectives, we utilized panel data from 34 provinces in the range of 2013 to 2024. During this period, Indonesia faced notable instances of food inflation, such as the 2013 spike following fuel price increases, and the 2022 inflation surge driven by global geopolitical conditions (Aji, 2015; Purwono et al., 2020).

We use several macroeconomic indicators such as poverty and food price inflation as the primary variable of interest. The poverty data is obtained from the Central Bureau of Statistics, specifically percentage of poor population (P0) by province. Poverty is measured as the percentage of individuals living below the poverty line[5]. In Indonesia, they measure the poverty line using the Cost of Basic Needs (CBN) method, which estimates the minimum monthly expenditure required to meet essential food and non-food needs. As of 2024, the national per capita poverty line was set at IDR 595,242 per month, based on a minimum daily consumption standard of 2,100 kilocalories per person. Food inflation is the main explanatory variable and is proxied by the provincial Consumer Price Index for the Food, Beverage, and Tobacco group. We decide to use this measure because it directly captures price movements in the food-related consumption basket most relevant to household welfare and poverty dynamics at the provincial level. In the empirical model, this CPI-based indicator is used to represent provincial food price inflation.

The control variables include regional economic output, human development, wages, and the exchange rate. Regional economic output is measured using Gross Regional Domestic Product at 2010 constant prices and is sourced from Central Bureau of Statistics. Human development is measured by capturing the quality of human development across provinces using human development index (HDI). Meanwhile, wages are proxied by the provincial minimum wage series published also by Central Bureau of Statistics. The exchange rate denotes the value of the Rupiah against the US Dollar obtained from Jakarta Interbank Spot Dollar Rate (JISDOR) data.

For good estimation results, some selected variables are transformed into their natural logarithms to improve the suitability for linear regression models and enhance inferential accuracy (Bland and Altman, 1996; Feng et al., 2014). The data are collected from the Central Bureau of Statistics (BPS) and Bank Indonesia. A comprehensive summary of the variables is presented in Table 1.

[3] Based on our data we used, we decompose the share of GDP across regions into two categories which are Western and Eastern Indonesia. The former shares ≈ 80% of Indonesia's total GDP, while the latter shares only ≈ 20% of Indonesia's total GDP..

[4] Susenas data for March 2023 shows that the national poverty rate has only reached 9.36% or is still below the 2020-2024 RPJMN target, that is in the range of 6.5-7.5%..

5 Poverty is calculated using data sourced from the results of the National Socio-Economic Survey (Susenas) Consumption and Expenditure Module (KP Module) for the period March and September each year.

Group	Variables	Description	Source
Dependent variable	Poverty (*pov*)	Percentage of Poor Population (P0) by Province	Central Bureau of Statistics
Independent variable	Food Inflation (*food_inf*)	Food inflation proxied by CPI of the Food, Beverage and Tobacco group (%)	Central Bureau of Statistics
Control Variables	GRDP (*lngdp*)	Natural logarithm Gross Regional Domestic Product at Constant Market Prices (2010=100)	Central Bureau of Statistics
	Human Development (*hdi*)	Human Development Index	Central Bureau of Statistics
	Wage (*wage*)	Natural logarithm of Provincial Minimum Wage (IDR)	Central Bureau of Statistics
	Exchange Rate (*lner*)	Natural logarithm of Rupiah exchange rate against the US dollar obtained from Jakarta Interbank Spot Dollar Rate (JISDOR) dat	Central Bureau of Statistics and Bank Indonesia

Source: Authors

Table 1: Data and data sources.

Method

We aim to understand the impact of food price inflation on poverty rate in Indonesia. Our analysis extends beyond the national level by exploring regional variations, specifically examining the effects of rising food prices on poverty in both rural and urban areas. We will also analyze the impact of food inflation in Indonesia by dividing it into two regions, which are Western Indonesia and Eastern Indonesia. This is intended to find out the differences in aspects where Western Indonesia has more advanced development. Moreover, these geographical distinctions allow for the possibility of difference in magnitudes in the impact of food inflation on poverty across regions (Gujarati and Porter, 2009; Mustika and Nurjanah, 2021).

To assess the impact of rising food price inflation, we employ a multivariate framework. The research model is adapted from the studies of Fujii (2013) and Shrestha and Chaudhary (2012) who similarly examined the effects of food inflation on poverty. The model is formulated as follow equation 1.

$$pov_{it} = \varphi_1 + \gamma_1 pov_{it\text{-}1} + \gamma_2 foodinf_{it} + \gamma 3 lngdp_{it} + + \gamma_4 hdi_{it} + \gamma_5 lnwage_{it} + \gamma_6 lner_{it} + \mu_i + \varepsilon_{it} \quad (1)$$

Where pov_{it} is poverty rate in province i at time t. $pov_{it\text{-}1}$ lagged poverty rate to account for persistence in poverty levels. The variable *foodinf* refers to food inflation, which is our primary variable of interest. $lngdp_{it}$, hdi_{it}, $lnwage_{it}$, and $lner_{it}$ are control variables representing regional Gross Domestic Product (province level), Human Development Index (HDI), minimum wages by province, and exchange rate. μ_i unobserved province-specific effects and ε_{it} idiosyncratic error term.

We detailed the model for exploring regional characteristic variations:

$$pov_{it}^{urban} = \varphi_1 + \gamma 1 pov_{it\text{-}1}^{urban} + \gamma_2 foodinf_{it} + + \gamma_3 lngdp_{it} + \gamma_4 hdi_{it} + \gamma_5 lnwage_{it} + \gamma_6 lner_{it} + + \mu_i + \varepsilon_{it} \quad (2)$$

$$pov_{it}^{rural} = \varphi_1 + \gamma 1 pov_{it\text{-}1}^{rural} + \gamma_2 foodinf_{it} + + \gamma_3 lngdp_{it} + \gamma_4 hdi_{it} + \gamma_5 lnwage_{it} + \gamma_6 lner_{it} + + \mu_i + \varepsilon_{it} \quad (3)$$

$$pov_{it}^{eastern} = \varphi_1 + \gamma 1 pov_{it\text{-}1}^{eastern} + \gamma_2 foodinf_{it} + + \gamma_3 lngdp_{it} + \gamma_4 hdi_{it} + \gamma_5 lnwage_{it} + \gamma_6 lner_{it} + + \mu_i + \varepsilon_{it} \quad (4)$$

$$pov_{it}^{uwestern} = \varphi_1 + \gamma 1 pov_{it\text{-}1}^{western} + \gamma_2 foodinf_{it} + + \gamma_3 lngdp_{it} + + \gamma_4 hdi_{it} + \gamma_5 lnwage_{it} + \gamma_6 lner_{it} + \mu_i + \varepsilon_{it} \quad (5)$$

We hypothesize that γ_2 in equations (1 - 5) will be positive, indicating that an increase in food inflation is expected to raise poverty levels due to its detrimental effect on purchasing power and production costs. In contrast, we anticipate that γ_3, γ_4, γ_5 will have negative signs, as improvements in macroeconomic factors are expected to reduce poverty. However, γ_6 are hypothesized to be positive, as currency depreciation may adversely affect the economy, leading to higher poverty levels in Indonesia.

Endogeneity poses a potential challenge in equations (1 - 5) as correlation between the error terms ε_{it} and the regressors could result in biased estimates of the coefficients $\gamma_{1\text{-}6}$. To address this issue, we employ the system Generalized Method of Moments (GMM) introduced by Arellano and Bover (1995) and Blundell and Bond (1998) to obtain consistent and reliable estimates. Furthermore, system GMM works better in a small sample size than difference GMM. This approach is used to address several key econometric issues. First, endogeneity concerns arise as food price inflation and poverty may have a bidirectional relationship, making OLS estimates biased. Second, the dynamic panel setting includes a lagged dependent variable $pov_{it\text{-}1}$ which can cause autocorrelation and inconsistency in fixed effects estimation. Lastly, instrumental variables in GMM help mitigate endogeneity by using lagged levels or differences, ensuring more reliable estimates for dynamic panel data analysis.

Result and discussion

Pre-estimation results

In the range of research period, the average poverty rate in Indonesia was 11.02 percent. The lowest poverty was 3.47 percent which occurred in Jakarta province in 2019. The highest poverty was in Papua province in 2013 with 31.53 percent. In contrast to aggregate poverty, rural poverty is on average higher, reaching 13.60 percent. Meanwhile, urban poverty recorded an average poverty rate of 7.36 percent in the 2013-2024 period. This shows that rural poverty is greater than urban poverty. Regarding socio-economic indicators for control variable, the average Human Development Index (HDI) across provinces was 70.40. The highest HDI was recorded in DKI Jakarta Province at 83.03, while the lowest was in Papua Province, with an HDI value of 56.25 in 2013. These disparities underscore the regional differences in development levels and the varying impacts of economic shocks, such as food price inflation, across different provinces (Table 2).

In terms of food inflation, the average during the research period was 4.86 percent with the lowest level being -0.40 percent and the highest inflation reaching 15.40 percent. The highest inflation occurred in 2015 in the province of DKI Jakarta. High food inflation was the impact of the policy of increasing fuel prices that year (Akhmad, 2014). This policy also has implications for increasing poverty because it disrupts people's consumption patterns through increasing transportation costs and several other goods (Izzati et al., 2023). We also performed correlation tests to identify potential multicollinearity. The results of the correlation test in this study are shown in Table 3.

As seen in the correlation matrix Table 3, some variables, particularly poverty (*pov*) and rural poverty (pov^{rural}), show a strong correlation of 0.958. This high correlation is expected, as rural poverty is essentially a subset of the overall poverty measure. However, since these two variables are included in separate models for analysis, this strong correlation does not pose a problem for multicollinearity within the model. According to Gujarati and Porter (2009), multicollinearity is generally considered a concern when two variables have a correlation coefficient greater than 0.8 within the same model. In our analysis, the correlation values between other variables are below this threshold, indicating no significant multicollinearity issues. Therefore, while there is a high correlation between poverty and rural poverty, it is not a concern in this study since these variables are modelled separately. The overall correlation analysis supports the conclusion that multicollinearity is not a significant issue in our model.

We also look at the stationarity of the data. Data that is not stationary can cause spurious regression so that the results become biased. The stationarity of the data used in this research is the unit root test using the Levin, Lin, and Chu test. Stationary data tends to approach the average value.

The stationarity tests conducted on the data, as shown in Table 4, indicate that all variables are stationary at both the 5% and 10% significance first differenced, according to the Fisher-type ADF tests, although some variable stationers in level like poverty, urban poverty, food inflation and human development index.

Variables	Obs	Mean	Median	Maximum	Minimum	Std. Dev.
pov	408	10.92	9.38	31.53	3.47	5.74
pov^{rural}	396	13.60	11.75	39.92	4.04	7.64
pov^{urban}	408	7.36	6.22	20.28	2.78	3.43
food inf	408	4.91	4.23	15.40	-3.43	3.07
lngrdp	408	11.93	11.78	14.58	9.81	1.15
lnwage	408	14.58	14.61	15.44	13.63	0.33
hdi	408	70.40	70.43	83.08	56.25	4.32
lner	408	9.55	9.54	9.69	9.41	0.08

Source: Author's estimated results

Table 2: Descriptive statistics.

Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
(1) *pov*	1.000							
(2) pov^{rural}	0.958	1.000						
(3) pov^{urban}	0.429	0.181	1.000					
(4) *food inf*	-0.008	0.003	-0.039	1.000				
(5) *lngrdp*	-0.311	-0.265	-0.040	0.037	1.000			
(6) *lnwage*	-0.164	-0.041	-0.348	-0.218	0.043	1.000		
(7) *hdi*	-0.648	-0.582	-0.138	-0.107	0.455	0.325	1.000	
(8) *lner*	-0.112	-0.090	-0.072	-0.209	0.122	0.699	0.383	1.000

Source: Author's estimated results

Table 3: Correlation Matrix.

Variables	*Fisher-Type ADF (Level)*	*Fisher-type ADF (First-differenced)*
pov	0.846	0.000***
pov^{rural}	0.780	0.000***
pov^{urban}	0.293	0.000***
food inf	0.000***	0.000***
lngrdp	1.000	0.000***
lnwage	0.008***	0.005***
hdi	1.000	0.000***
lner	1.000	0.000***

Note: ***), **), and *) Stationarity at the 1%, 5%, and 10% levels.
Source: Author's estimated results

Table 4: Results of data stationarity test.

Impact of food inflation on poverty

The relationship between food price inflation and poverty in Indonesia is examined through three distinct models, as shown in Table 5. Model 1 estimates the impact of food inflation on poverty without controlling for any other variables, providing a baseline understanding of the relationship between these two factors. Model 2, in contrast, includes a full set of control variables, such as regional GDP, wages, human development index (HDI), and the natural exchange rate (lner), thereby accounting for other socio-economic factors that may influence poverty. Model 3 explores the potential non-linear effects of food inflation on poverty by adding a squared term of food inflation, allowing for the possibility that the impact of food inflation on poverty may change as food inflation rises.

Variables	Model 1 Poverty	Model 2 Poverty	Model 3 Poverty
Poverty (-1)	-0.104*** (0.027)	-0.309*** (0.093)	-0.019 (0.029)
lner		-1.477** (0. 651)	-0.749*** (0.222)
lngdp		10.854*** (2.489)	0.568 (0.411)
lnwage		-0.200 (0.202)	-1.610*** (0.172)
hdi		-1.636*** (0.249)	-0.848*** (0.039)
Food Inflation	**0.009***** (0.002)	**0. 026***** (0.009)	**-0.058***** (0.008)
Squared Food Inflation			**0.008***** (0.001)
Constant	-0.271*** (0.012)	-0.102 (0.075)	0.360*** (0.050)
Obs	340	340	340
Overidentification test (Hansen Test)	0.99	0.173	0.960

Note: (1) Estimation was conducted using the Two-step System GMM method; (2) Standard errors are in parentheses; (3) ***), **), and *) indicate statistical significance at the 1%, 5%, and 10% levels, respectively.
Source: Author's data processing results

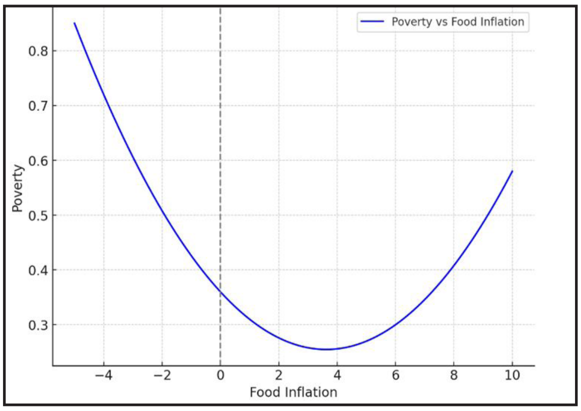
Table 5: Impact of inflation on poverty in Indonesia.

Results from the estimation reveal that food inflation has a significant positive impact on poverty in three models that show Table 5. In Model 1, the estimation shows a positive and statistically significant relationship between food inflation and poverty, with a coefficient of 0.009, indicating that a 1 percent increase in food inflation results in a 0.009 percent increase in the poverty rate, providing a baseline understanding of the direct impact of food price inflation on poverty. Model 2, which incorporates a full set of control variables, reveals a stronger impact, with the coefficient for food inflation rising to 0.026. This suggests that when accounting for other socio-economic factors, a 1 percent increase in food inflation leads to a 0.026 percent increase in poverty, highlighting that food inflation exacerbates poverty more significantly when broader economic conditions are considered. These results emphasize the need to account for various socio-economic factors when analyzing the impact of food inflation on poverty. This result is in line with several previous studies where Increasing food prices causes community vulnerability, especially in poor households, because it can reduce income and disrupt consumption patterns (Bhattacharya, 2017; Fujii, 2013; Wu and Xu, 2021).

Food price inflation drives poverty by disrupting consumption patterns and income for the Indonesian population. The relatively high poverty rate and the fact that over 50 percent of Indonesian household expenditures are allocated to food make price increases have a more significant impact on the well-being of the population in Indonesia (Hill, 2021). Rising food prices burden expenditures on food and can lead to further consequences such as malnutrition and hunger. These conditions subsequently have more serious implications for future economic development. The impact can also cascade into related sectors, reducing consumer spending on non-essential goods and services and potentially disrupting business sustainability across various sectors

(Aji, 2015; Rammohan and Tohari, 2023).

We also exploring the possibility non-linear effect of food inflation on poverty in Indonesia as we can see above in Table 5 (Model 3). We can confirm that there is a non-linear effect of food inflation in Indonesia. This allows us to explore the optimal level of food inflation in Indonesia to minimize the increasing response on poverty rate. As we can see, the optimal level for food inflation is based on the slope of the non-linear function that is equal to zero (see Figure 1 below). It is possible to state that optimal level of inflation can represent as an indicator of economic activity through productivity channel. Thus, targeting food inflation in the range of 3% is somehow optimal.



Source: Author's data processing results

Figure 1: The non-linear effect of food inflation on poverty rate.

Impact of food inflation on poverty: regional perspectives

From a regional perspective (Table 6), we found that food inflation also affects poverty in both urban and rural areas. However, there is a difference in the degree of influence, which is greater in rural areas compared to urban areas. 1 percent increase in food prices affects poverty by 0.024 percent in rural areas and 0.023 percent in urban areas. Rural communities are more vulnerable to food price increases due to lower income buffers and purchasing power. In contrast, in urban areas, higher incomes mitigate the impact of rising food prices. Examining the poverty conditions in both regions reveals that the rural poverty rate is much higher than that in urban areas. In 2023, rural poverty in Indonesia reached 12.22 percents, significantly higher than urban poverty, which stood at 7.29 percent.

Variables	Model 1 Urban Poverty	Model 2 Rural Poverty
Urban/Rural Poverty (-1)	-0.321*** (0.045)	-0.301*** (0.042)
Food Inflation	0.023*** (0.005)	0.024*** (0.005)
lner	1.362*** (0.349)	-0.658* (0.367)
lngdp	7.999*** (1.176)	7.716*** (1.039)
lnwage	-0.693*** (0.113)	0.590** (0.285)
ipmhdi	-1.589*** (0.139)	-1.128*** (0.092)
Constant	0.269*** (0.042)	-0.134** (0.056)
Obs	340	330
Overidentification test (Hansen Test)	0.235	0.370

Note: (1) Estimation was conducted using the Two-step System GMM method; (2) Standard errors are in parentheses; (3) ***), **), and *) indicate statistical significance at the 1%, 5%, and 10% levels, respectively.
Source: Author's data processing results

Table 6. Impact of inflation on regional poverty.

The disparity in the impact of poverty in rural and urban areas is also influenced by increasing inequality in Indonesia. Data from the Central Bureau of Statistics (BPS) indicates that inequality in Indonesia in 2023 reached 0.388, an increase from the previous year's 0.381. This increase in inequality suggests a rise in poverty in Indonesia. Moreover, the role of poverty in Indonesia is more food oriented. The contribution of food commodities to the Poverty Line is much greater than that of non-food commodities. The distribution of the poverty line in Indonesia in 2023 shows that 74.21 percent is attributed to food commodities and 25.79 percent to non-food items.

To enrich the analysis, we also present the estimation results by dividing two different regions, namely western Indonesia and eastern Indonesia. These two regions have different characteristics where western Indonesia has a more advanced economic development condition while eastern Indonesia is the opposite. The estimation results are shown in Table 7.

The estimation results show that the increase in food prices is also statistically significant in increasing poverty in both western and eastern Indonesia. 1 percent increase in food prices increases the poverty rate by 0.017 percent in western Indonesia and 0.055 percent in eastern Indonesia. These results indicate that the impact of rising food prices on poverty levels is greater in less developed areas (Eastern Indonesia) compared to more developed areas (Western Indonesia). Rising food prices worsen poverty and food insecurity more in less developed regions due to high income allocation to food, weak safety nets, economic vulnerability, and widespread informal employment. Without effective policies and international aid, such price surges can lead to severe humanitarian crises.

Variables	Model 1 Western Indonesia Poverty	Model 2 Eastern Indonesia Poverty
poverty (-1)	-0.218**(0.108)	-0.220***(0.056)
Food Inflation	0.017**(0.008)	0.055***(0.012)
lner	1.088*(0.614)	0.370(0.596)
lngdp	5.137***(1.959)	2.540(1.661)
lnwage	0.613***(0.228)	-2.160***(0.713)
hdi	-1.611***(0.193)	-0.705***(0.086)
Constant	0.242***(0.082)	-0.050(0.095)
Obs	340	330
Overidentification test (Hansen Test)	0.235	0.370

Note: (1) Estimation was conducted using the Two-step System GMM method; (2) Standard errors are in parentheses; (3) ***), **), and *) indicate statistical significance at the 1%, 5%, and 10% levels, respectively.
Source: Author's data processing results

Table 7: Impact of inflation on poverty in Western and Eastern Indonesia.

The rise in food prices exacerbates poverty Eastern Indonesia because most low-income households spend a large portion of their income on purchasing food. When food prices rise, their expenditure on these basic need increases, reducing their purchasing power for other necessities such as education, healthcare, and productive investments. This result is further reinforced by the relatively high poverty rates eastern Indonesia compared to western Indonesia. For instance, the poverty rate in some eastern Indonesia regions, such as Bali & Nusa Tenggara, reached 13.29 percent, and Papua recorded 20.68 percent in 2023. In contrast, the poverty rate in the western Indonesia region was 8.29 percent.

Additionally, geographic disparities within Indonesia illustrate the varying degrees of impact related to food inflation. In less developed regions, such as Eastern Indonesia, the reliance on subsistence farming makes households particularly susceptible to fluctuations in food prices. Here, food inflation does not just impact purchasing decisions; it critically undermines food self-sufficiency and economic stability (Pangesti et al., 2023). By contrast, Western Indonesia, characterized by more developed agricultural sectors and better infrastructure, may cushion some of the impacts of food inflation due to a more diversified economy that can absorb such shocks better (Deski et al., 2022).

Conclusion

This study examines the impact of food price inflation on poverty in Indonesia, focusing on regional disparities. The findings demonstrate that rising food prices significantly increase poverty levels, with the effect being more pronounced in rural areas compared to urban areas. Rural households, due to their higher reliance on food expenditures and lower income buffers, are particularly vulnerable to food price fluctuations. In contrast, urban areas show greater resilience, largely due to higher incomes and more diversified economic activities. Furthermore, the study reveals that the impact of food inflation on poverty is more severe in Eastern Indonesia, where economic development is less advanced, compared to Western Indonesia. This regional variation highlights the importance of considering local contexts when analyzing the effects of food inflation, as more developed regions tend to have better mechanisms to buffer the negative impacts. These results align with global studies that show food inflation exacerbates poverty, particularly in regions with weaker economic foundations. However, the findings of this research also underscore the necessity of addressing regional disparities, as the effects of food inflation are not uniform across the country. In less developed areas, food price increases can lead to higher poverty rates, not just through direct consumption impacts but also by limiting access to essential services like education and healthcare. This study contributes to the broader understanding of how food inflation interacts with regional economic conditions, emphasizing that national-level policies may not be sufficient to address the unique challenges faced by different regions.

Based on these findings, it is recommended that the Indonesian government adopt region-specific strategies to mitigate the impact of food inflation. In rural areas, particularly in Eastern Indonesia, where food inflation disproportionately affects poverty due to higher reliance on food expenditures and weaker social safety nets, targeted interventions such as food subsidies, direct cash transfers, and food assistance programs are essential. Strengthening regional infrastructure, especially in Eastern Indonesia, by improving food distribution networks and investing in agricultural infrastructure, will enhance resilience against food price shocks. Moreover, regional development policies should focus on reducing the economic and infrastructure gaps between Western and Eastern Indonesia, promoting rural economic diversification through better access to education, healthcare, and support for small-scale agro-industries. These efforts, tailored to the specific needs of each region, will help mitigate the severe consequences of food inflation on poverty, particularly in the most vulnerable areas.

Despite the valuable insights provided by this study, there are several limitations that should be acknowledged. One limitation is the focus solely on food inflation, while other factors, such as income inequality, labor market conditions, and external economic shocks, may also significantly affect poverty dynamics. Future research could explore the interaction between food inflation and these factors, providing a more comprehensive understanding of the drivers of poverty. Additionally, the study's findings suggest that regional disparities play a crucial role in the impact of food inflation, yet the effectiveness of specific regional policies in mitigating these effects remains underexplored. Further research could investigate how targeted regional interventions, particularly in Eastern Indonesia, can better address the unique challenges posed by food inflation.

Corresponding author:
Misbahol Yaqin, S.E., M.S.E
Jalan Kalimantan, Kampus Tegal Boto
Jember, East Java, 68121, Indonesia
Email: misbahol.yaqin@unej.ac.id

References

[1] Adjemian, M. K., Arita, S., Meyer, S. and Salin, D. (2023) "Factors affecting recent food price inflation in the United States", *Applied Economic Perspectives and Policy*, March, 1-29. ISSN 2040-5804. DOI 10.1002/aepp.13378.

[2] Affandi, Y. (2015) "Unveiling Stubborn Inflation in lndonesia", *Economics and Finance in Indonesia*, Vol. 59, No. 1, p. 47. ISSN 2442-9260.

[3] Aji, P. (2015) "*Summary of Indonesia's Poverty Analysis*", *ADB Papers on Indonesia*, No. 4, Oct. 2015. Asian Development Bank.

[4] Akhmad, A. (2014) "Dampak Kompensasi Kenaikan Harga Bbm Terhadap Kemiskinan di Indonesia", *Jurnal Academica Fisip Untad*, Vol. 6, No. 2, pp. 1235-1252. ISSN 1411- 3341. (In Indonesian).

[5] Anugrah, D. F. and Pratama, R. (2018) "Core Inflation Determinants: Decomposition of Global and Domestic Factors Associated With Sectoral Inflation", Working Papers WP/31/2018, Bank Indonesia.

[6] Arellano, M. and Bover, O. (1995) "Another look at the instrumental variable estimation of error-components models", *Journal of Econometrics*, Vol. 68, No. 1, pp. 29-51. E-ISSN 1872-6895. DOI 10.1016/0304-4076(94)01642-D.

[7] Bhattacharya, R. (2017) "Effectiveness of monetary policy in stabilising food inflation: Evidence from advanced and emerging economies", Working Papers 17/209, National Institute of Public Finance and Policy.

[8] Bland, J. M. and Altman, D. G. (1996) "Statistics Notes: Transforming data", *Brittish Medical Journal*, p. 312. E-ISSN 1756-1833. DOI 10.1136/bmj.312.7033.770.

[9] Blundell, R. and Bond, S. (1998) "Initial conditions and moment restrictions in dynamic panel data models", *Journal of Econometrics*, Vol. 87, No. 1, pp. 115-143. ISSN 0304-4076. DOI 10.1016/S0304-4076(98)00009-8.

[10] Deski, S. W., Sartiyah, S. and Seftarita, C. (2022) "Investigating the Effect of Rice Commodity Price Volatility on Provincial Capital Poverty: An Evidence from Indonesia", *International Journal of Global Optimization and Its Application*, Vol. 1, No. 2, pp. 90-99. ISSN 2948-4030. DOI 10.56225/ijgoia.v1i2.18.

[11] Faharuddin, F., Yamin, M., Mulyana, A. and Yunita, Y. (2023) "Impact of food price increases on poverty in Indonesia: empirical evidence from cross-sectional data", *Journal of Asian Business and Economic Studies*, Vol. 30, No. 2, pp. 126-142. ISSN 2515-964X. DOI 10.1108/JABES-06-2021-0066.

[12] Feng, C., Wang, H., Lu, N., Chen, T., He, H., Lu, Y. and Tu, X. M. (2014) "Log-transformation and its implications for data analysis", *Shanghai Archives of Psychiatry*, Vol. 26, No. 2, pp. 105-109. ISSN 1002-0829. DOI 10.3969/j.issn.1002-0829.2014.02.009.

[13] Fujii, T. (2013) "Impact of food inflation on poverty in the Philippines", *Food Policy*, Vol. 39, pp. 13-27. ISSN 0306-9192. DOI 10.1016/j.foodpol.2012.11.009.

[14] Gujarati, D. N. and Porter, D. C. (2009) "*Basic Econometrics*", McGraw-Hill Companies (Ed.), Inc. New York (5th ed.). ISBN 978-0-07-337577-9.

[15] Hammoudeh, S., Nguyen, D. K. and Sousa, R. M. (2015) "US monetary policy and sectoral commodity prices", *Journal of International Money and Finance*, Vol. 57, pp. 61-85. ISSN 0261-5606. DOI 10.1016/j.jimonfin.2015.06.003.

[16] Headey, D. (2018) "Food prices and poverty", *World Bank Economic Review*, Vol. 32, No. 3, pp. 676-691. ISSN 0258-6770. DOI 10.1093/wber/lhw064.

[17] Headey, D. and Hirvonen, K. (2023) "Higher food prices can reduce poverty and stimulate growth in food production", *Nature Food*, Vol. 4, pp. 699-706. E-ISSN 2662-1355. DOI 10.1038/s43016-023-00816-8.

[18] Hill, H. (2021) "What's happened to poverty and inequality in indonesia over half a century?", *Asian Development Review*, Vol. 38, No. 1, pp. 68-97. E-ISSN 1996-7241. DOI 10.1162/adev_a_00158.

[19] Ivanic, M., Martin, W. and Zaman, H. (2012) "Estimating the Short-Run Poverty Impacts of the 2010-11 Surge in Food Prices", World Development, Vol. 40, No. 11, pp. 2302-2317. ISSN 2468-0532. DOI 10.1016/j.worlddev.2012.03.024.

[20] Izzati, R. A., Yusrina, A. and Suryahadi, A. (2023) "Estimating the effect of a fuel price increase on poverty and inequality: Evidence from a fuel subsidy reduction in Indonesia", SMERU Working Paper, Vol. 1, pp. 1-7. [Online]. Available: https://smeru.or.id/id/publication-id/mengestimasikan-dampak-kenaikan-harga-bbm-terhadap-kemiskinan-dan-ketimpangan-bukti [Accessed: Sept 25, 2025]. (In Indonesian).

[21] Kıroğlu, B. S. and Sezgin, Ş. (2021) "The Impact of Food Inflation on Poverty in Turkey", J*ournal of Economics, Business & Political Studies*, Vol. 8, No. 2, pp. 256-273. E-ISSN 2687-4695. DOI 10.48064/equinox.995580.

[22] LPEM FEB UI. (2025) "Indonesia Economic Outlook Q3 - 2025". [Online]. Available: https://en.lpem.org/indonesia-economic-outlook-q3-2025-kondisi-normal-baru-di-bawah-5/ [Accessed: Sept 25, 2025]. (In Indonesian).

[23] Luster, T., Gunderson, Z. J., Sun, S. and Loder, R. T. (2023) "Slipped Capital Femoral Epiphysis, Food Deserts, Poverty, and Urban/Rural Residence: Is There a Link?", *Journal of Pediatric Orthopaedics*, Vol. 43, No. 3. ISSN 1473-5865. DOI 10.1097/BPO.0000000000002315.

[24] Misdawita, H. D. and Nugroho, A. (2019) "Impacts of Food Prices on the Economy: Social Accounting Matrix and Microsimulation Approach in Indonesia", *Review of Urban and Regional Development Studies*, Vol. 31, No. 1-2, pp. 137-154. E-ISSN 1467-940X. DOI 10.1111/rurd.12099.

[25] Monge, M. and Lazcano, A. (2022) "Commodity Prices after COVID-19: Persistence and Time Trends", *Risks*, Vol. 10, No. 6. ISSN 2227-9091. DOI 10.3390/risks10060128.

[26] Munnell, A. H. (1992) "Policy Watch: Infrastructure Investment and Economic Growth", *Journal of Economic Perspectives*, Vol. 6, No. 4, pp. 189-198. E-ISSN 1944-7965. DOI 10.1257/jep.6.4.189.

[27] Mustika, C. and Nurjanah, R. (2021) "Rural and urban poverty models on Sumatra Island", *Jurnal Perspektif Pembiayaan Dan Pembangunan Daerah*, Vol. 9, No. 1, pp. 107-114. E-ISSN 2355-8520. DOI 10.22437/ppd.v9i1.10684.

[28] Pangesti, A. W., Darsono and Antriyandarti, E. (2023) "Causality Analysis of Rice Prices with Inflation Rate in Indonesia", *Agriekonomika*, Vol. 7, No. 3, pp. 539-549. ISSN 2301-9948. DOI 10.14710/agrisocionomics.v7i3.16460.

[29] Pratikto, R. and Ikhsan, M. (2016) "Inflasi Makanan dan Implikasinya terhadap Kebijakan Moneter di Indonesia" (Food Inflation and Its Implications for Monetary Policy in Indonesia), *Jurnal Ekonomi dan Pembangunan Indonesia*, Vol. 17, No. 1, pp. 58-74. E-ISSN 3031-3406. DOI 10.21002/jepi.v17i1.658. (In Indonesian).

[30] Prayoga, A. E., Nazamuddin, B. S. and Dawood, T. C. (2025) "Analysis of the Effect of Food Prices, GDP, Inflation, and Exchange Rate on Poverty in 4 ASEAN Countries", *Majority Science Journal*, Vol. 3, No. 2, pp. 135-146. ISSN 3025-1990. DOI 10.61942/msj.v3i2.372.

[31] Purwono, R., Yasin, M. Z. and Mubin, M. K. (2020) "Explaining regional inflation programmes in Indonesia: Does inflation rate converge?", *Economic Change and Restructuring*, Vol. 3, No. 4, pp. 571-590. ISSN 1574-0277. DOI 10.1007/s10644-020-09264-x.

[32] Rammohan, A. and Tohari, A. (2023) "Rural poverty and labour force participation: Evidence from Indonesia's Village fund program", *PLOS ONE*, Vol. 18, No. 6 June, pp. 1-15. ISSN 1932-6203. DOI 10.1371/journal.pone.0283041.

[33] Rivai, A. (2022) "The monetary policy impact on agricultural growth and food prices", *International Journal of Research in Business and Social Science*, Vol. 11, No. 9, pp. 158-165. ISSN 2147- 4478. DOI 10.20525/ijrbs.v11i9.2234.

[34] Samal, A., Ummalla, M. and Goyari, P. (2022) "The impact of macroeconomic factors on food price inflation: an evidence from India", *Future Business Journal*, Vol. 8, No. 1. E-ISSN 2314-7210. DOI 10.1186/s43093-022-00127-7.

[35] Shahbaz, M., Song, M., Ahmad, S. and Vo, X. V. (2022) "Does economic growth stimulate energy consumption? The role of human capital and R&D expenditures in China", *Energy Economics*, Vol. 105, No. December 2020, p. 105662. ISSN 1873-6181. DOI 10.1016/j.eneco.2021.105662.

[36] Shrestha, M. B. and Chaudhary, S. K. (2012) "The Impact of Food Inflation on Poverty in Nepal", *NRB Economic Review*, Vol. 24, No. 2, pp. 1-14. ISSN 1608-6627. DOI 10.3126/nrber.v25i1.52715

[37] Suryahadi, A., Al Izzati, R. and Suryadarma, D. (2020) "The Impact of COVID-19 Outbreak on Poverty: An Estimation for Indonesia (Draft)", *SMERU Working Paper*, April, pp. 1-20. [Online]. Available: https://smeru.or.id/sites/default/files/publication/wp_covid19impact_draft.pdf [Accessed: Oct 16, 2026].

[38] Valero-Gil, J. N. and Valero, M. (2008) "The effects of rising food prices on poverty in Mexico", *Agricultural Economics*, Vol. 39, No. Suppl. 1, pp. 485-496. ISSN 1805-9295. DOI 10.1111/j.1574-0862.2008.00354.x

[39] Warr, P. and Yusuf, A. A. (2014) "World food prices and poverty in Indonesia", *Australian Journal of Agricultural and Resource Economics*, Vol. 58, No. 1, pp. 1-21. ISSN 1467-8489. DOI 10.1111/1467-8489.12015.

[40] Wu, X. and Xu, J. (2021) "Drivers of food price in China: A heterogeneous panel SVAR approach", *Agricultural Economics (United Kingdom)*, Vol. 52, No. 1, pp. 67-79. ISSN 0169-515. DOI 10.1111/agec.12607.

[41] Xi, Z., Zhao, Y., Ren, Y., Zheng, Z. and Han, X. (2025) "The asymmetric price effects on food demand and nutrient intake of rural low-income households in China", *China Agricultural Economic Review*, pp. 1-22. ISSN 1756-1388. DOI 10.1108/CAER-05-2024-0166.

Appendix 1.

To get the full data that we used in this paper, please access this link: https://unej.id/-zZnBBe

Volume XVIII Number 3, 2026

Feeding Simulation Model: Data Acquisition to Achieve Climate Neutrality Goals

Mairita Zaķe, Andris Lapāns, Anda Mežgaile, Oskars Java

Vidzeme University of Applied Sciences, Valmiera, Latvia

Abstract

In line with the European Union's development plans, preserving a healthy environment for future generations is a key priority. Progress towards climate neutrality requires reliable, coherent, high-quality and interoperable data that capture historical trends and emerging societal dynamics while reflecting ecological, economic and technological systems. This paper addresses data acquisition and integration across five sectors: energy, agriculture, land use, land-use change and forestry, industry and waste management. The proposed framework supports Latvia's national climate neutrality simulation model with harmonised data. The analysis reveals substantial differences in data accessibility, formats, update regularity, and sectoral methodologies. Critical datasets are often available only in unstructured formats, such as PDFs of infographics, limiting automation in data collection. The developed modular workflow and harmonisation system improve model reliability and support evidence-based long-term climate policy design at national and global levels.

Keywords

Data mapping, data acquisition, climate neutrality, data model, data exchange.

Zaķe, M., Lapāns, A., Mežgaile, A. and Java, O. (2026) "Feeding Simulation Model: Data Acquisition to Achieve Climate Neutrality Goals", *AGRIS on-line Papers in Economics and Informatics*, Vol. 18, No. 3, pp. 137-146. ISSN 1804-1930. DOI 10.7160/aol.2026.180310.

Introduction

The European Union's climate policy aims to achieve climate-neutral Europe by 2050, which means the involvement of all member states. To monitor the progress of this process, several different models are developed, often covering one specific sector or a few of the sectors (for example, energy). It is emphasized that the need to work on harmonizing input data for better model use and provisions for cross-sector and cross-border integration (Mikropoulos et al., 2025; Samarasinghe et al., 2025). Assessing the impact of different drivers therefore requires climate-neutrality models capable of representing interactions across multiple sectors and system components. Such models are inherently complex, as they integrate diverse processes, assumptions, and data flows with a single analytical framework (Randall et al., 2007). As climate information may be collected from a variety of resources, based on different assumptions and different methodological limitations, many challenges for regional and national climate change, thus climate neutrality, may occur, such as quality of data and quality control, homogenization, data scarcity, gridding and structural uncertainty (IPCC, 2023).

Data acquisition challenges for climate neutrality simulation models manifest across multiple interconnected dimensions. The primary challenges span five domains: (1) data availability issues including inadequate collection and monitoring systems (Ulpiani et al., 2023) and inaccessible data formats (Zargar et al., 2024), (2) quality problems such as inaccurate emissions baselines (Ulpiani et al., 2023) and negative emission balance errors (O'Keeffe et al., 2023), (3) standardization gaps with inconsistent emission factor definitions across countries (Skujevska et al., 2024) and overly simplistic land-use representation (Prestele et al., 2017), (4) integration difficulties including incompatible data sources (Ghent et al., 2011) and lack of fully coupled models (Prestele et al., 2017), (5) and spatial-temporal resolution mismatches between model requirements and available data (Ghent et al., 2011; Prestele et al., 2017). These challenges stem from (1) methodological inconsistencies between countries (Prestele et al., 2017; Skujevska et al., 2024), (2) technical limitations in measurement systems (Schnase et al., 2016; Ulpiani et al., 2023), (3) institutional barriers in reporting frameworks (O'Keeffe et al., 2023; Zargar et al., 2024), (4) scientific uncertainties in process understanding (Figueroa & Farias, 2022; Ghent et al., 2011), and (5) resource constraints affecting infrastructure and capacity (Krysovatyy et al., 2024; Ulpiani et al., 2023). The long-term impacts of policies are often uncertain, making it challenging to design effective, far-reaching measures. Policy modelling tools can help to address all barriers to reduce the efforts of decision-making, but it is not possible to run a mathematical model without trusted data.

This publication is a part of a larger study that aims to develop a framework of a unified national policy decision-making support tool to simulate the impact of policy decisions on Latvia's progress toward climate neutrality by 2050. The project develops a dynamic simulation model using a system dynamics approach to interconnect key economic sectors and simulate policy impacts. Through multi-objective differential optimization, the model integrates technological, economic, social, and environmental dimensions to identify optimal pathways toward achieving climate neutrality. In this paper, we focus on data acquisition and analysis for five predefined economic sectors: (1) energy, (2) industry, (3) agriculture, (4) land use, land-use change and forestry (LULUCF), and (5) waste management.

Materials and methods

This study collects and integrates statistical data from Latvia, drawing on diverse datasets from various sources (institutional, regulatory, and statistical) for the period 2015-2023. The methodological approach, starting from the data availability analysis and finishing with the main principles of the data exchange prototype, is described in this paragraph.

Analysis of data availability

Data availability analysis includes both data sources mapping and evaluating whether specific data sources are suitable for use in modelling. It is determined whether the available data provides the necessary details, accuracy, and time coverage. The format of data availability is determined, for example, whether it is necessary to gather information from different reports, graphs, or tables, whether the data is available free of charge, for a license fee, or whether the data is classified as restricted access data. An analysis and description of what kind of data processing will be required is carried out.

Data collection and source verification

Data is collected from various official and reliable sources, such as the Central Statistical Bureau of the Republic of Latvia, the cadastral register of State Land Service of the Republic of Latvia, industry reports, and scientific studies. At this stage, the reliability of the sources, the relevance of the data, and their relevance to the requested modelling context are checked.

Metadata review is primarily performed to ensure that the data contains the information required for the modelling process.

Data categorisation and standardisation

The resulting data is often available in different classifications or formats. For example, one source may provide data on "healthcare buildings" while another source breaks them down as "buildings of ambulatory or inpatient medical facilities" and "hospitals". In this step, the data is transformed and combined to ensure its consistency.

Lastly, unit standardization is conducted (e.g., if different sources use different units, such as square meters or hectares, they are converted to a single unit suitable for the modelling process).

Data consolidation

If the data is divided into different categories or time periods, it is combined. For example, if the time dimensions in different sources do not match (e.g., one source offers data yearly, and another quarterly), data transformation is performed to ensure a uniform time dimension.

Combining data from multiple sources into one unified data set is performed. This can include both statistical summation and averaging to obtain representative data.

Data acquisition, transformation, and compilation methodology

If the data is available as a downloadable dataset – usually in CSV, JSON or Excel formats – importing into a database is accomplished as the first step. The workflow involves importing the data stored "as is". The next step is transformation, which provides data conversion for consolidation purposes. The values are not changed, rather measurement units are changed. If necessary, the recalculation based on measurement units is performed, or aggregate functions and grouping are performed.

Manual data entry into the database is conducted if the data is available in read-only mode as PDF reports or online publications.

Transformations are implemented using data processing tools (Python, SQL). Afterwards, integration and storage in a structured data model in the database takes place, which is followed by export and transfer for further use in modelling tools (individual CSV format file for each sector). Each CSV file is supplemented by metadata – detailed information about the used data sources, measurement units, time period, as well as the necessary actions to integrate the available output information into the common data model.

Data storage and exchange prototype

A data storage and exchange prototype is implemented to facilitate straightforward data consumption by end users and the simulation environment.

To ensure consistent access and interoperability, all cleaned and processed data is stored in a central database for long-term maintenance and controlled version management.

Results and discussion

Analysis of data availability

The sub-model for greenhouse gas (GHG) inventory is developed in accordance with the Intergovernmental Panel on Climate Change (IPCC) Guidelines and Standards (IPCC, 2019). The study primarily uses Latvia's officially calculated, internationally reported, and projected greenhouse gas (GHG) data, prepared in accordance with national and international legislation and the Cabinet on Ministers of Latvia Regulation No. 675 of 25 October 2022, "Procedures for the Establishment and Maintenance of the Greenhouse Gas Inventory, Projection, and Climate Change Adaptation Reporting Systems" (Cabinet of Ministers, 2022). The data acquisition is implemented in accordance with the Order of the Cabinet of Ministers of Latvia No. 610, Section 6.3 (Cabinet of Ministers, 2023). The sub-model identifies the data required for the operation of the decision-support system, with particular attention to the integration of new regulatory provisions and the analysis of policy proposals.

Data collection and source verification

The main data sources include the Latvian Environment, Geology and Meteorology Centre, the Central Statistical Bureau of the Republic of Latvia, and the Ministry of Climate and Energy of the Republic of Latvia, ensuring alignment with the IPCC Guidelines for National Greenhouse Gas Inventories. Table 1 summarizes the main data sources for climate neutrality monitoring in the analyzed sectors, indicating the data sources, time coverage, data format, and the most significant challenges in their use. The table provides an overview of data availability and identified limitations that need to be considered in climate policy modelling and monitoring.

The information summarized in the table shows that a wide range of official statistics and administrative data is available in Latvia for the purposes of monitoring climate neutrality. At the same time, common problems have been identified in several sectors, including data fragmentation between institutions, insufficient availability of machine-readable data, different time coverage, limited metadata and insufficient data interoperability between sectors. These challenges limit the use of data in integrated climate policy modelling and in assessing progress towards climate neutrality.

The metadata analysis led to the systematic identification of key data sources within each sector, accompanied by an assessment of their availability, level of detail, and compatibility with other datasets. The analysis revealed substantial inter-sectoral variation in data structure and degree of standardization. A unique identifier was assigned to each data source, enabling its consistent link to the metadata catalogue and the integrated database. This configuration supports precise attribution of data origin and ensures full traceability across all stages of data processing and integration. Consequently, a unified and structured overview of sector-specific data sources was established.

Sector	Main data sources	Temporal coverage	Data format	Key challenges
Energy	CSP statistical databases, State Construction Control Office energy performance data, State Land Service cadastral data, Public Utilities Commission heat supply reports and tariff database	Mostly annual time series; coverage varies by dataset, generally available for multiple years	CSV, Excel, online statistical tables; some data available only as PDF reports or administrative files	Fragmented data sources; some data available only in PDF format; limited access to confidential data; absence of systematic distinction between renovated and non-renovated buildings; inconsistent metadata and time-series documentation
Agriculture	CSP agricultural statistics, Rural Support Service administrative data, SUDAT/FADN data, agricultural annual reports, European Commission FADN/FSDN database	Varies across indicators; some key variables available only since 2019	CSV, Excel, online databases, PDF reports, administrative datasets	Incomplete long-term time series; missing indicators for manure management and enteric fermentation; limited methodological harmonisation; data dispersed across institutions; insufficient integration of economic and biophysical indicators
LULUCF	CSP forestry and land-use statistics, National Forest Register (State Forest Service), administrative datasets, EU climate finance databases, Eurostat fisheries statistics	Long-term statistical series available for major indicators; coverage varies for administrative datasets	CSV, Excel, online statistical tables, administrative databases, geospatial datasets	Limited access to detailed administrative and geospatial data; insufficient harmonisation of spatial datasets; methodological changes affecting time-series consistency; lack of unified access mechanisms for modelling purposes
Waste management	CSP waste statistics, Latvian Environment, Geology and Meteorology Centre waste databases, municipal waste management data, State Environmental Service registers, wastewater utility reports	Annual reporting cycles with varying historical coverage	Excel, CSV, databases, PDF reports	Data stored in multiple systems; limited automated access (lack of APIs); inconsistent integration with energy sector data; insufficient technical detail for modelling waste treatment processes and methane recovery
Industry	CSP industrial production statistics, energy balance statistics, fuel and biogas consumption data	Long-term annual statistical series	CSV, Excel, online statistical tables	Strong coverage of economic indicators but limited process-level activity data; insufficient detail on energy use and raw material consumption for emissions modelling; need for higher sectoral granularity
Macro-economics	CSP national accounts, demographic statistics, energy price statistics, regional GDP indicators	Long-term annual and periodic statistical series	CSV, Excel, online statistical tables	Missing or incomplete metadata; some data gaps and duplicate indicators; insufficient coherence with data structures of other sectors; limited possibilities for cross-sectoral integration

Source: Authors

Table 1: Main data sources for climate neutrality monitoring by sector.

Data categorisation and standardisation

Data quality was assessed with respect to completeness, consistency, comparability, and temporal coverage. Completeness was assessed by checking whether the dataset contained sufficient information for all required time periods and sectors, thereby eliminating incomplete or fragmentary information that could affect the results of the analysis. Consistency was analyzed across different data series and sources to identify potential discrepancies. Comparability was ensured by examining the alignment of the data with other national and international datasets, enabling reliable and verifiable analyses.

Data consolidation

Sectoral consistency was maintained by confirming that the classification and structuring of the data adhered to industry standards, ensuring compatibility with professionally used datasets. Additionally, time coverage was reviewed during the sourcing process to guarantee continuity throughout the entire period. Based on this assessment, an improvement plan was developed to address data gaps, refine inconsistent series, and define additional verification procedures. The resulting dataset and workflow are designed to remain interoperable with the Climate Policy and GHG Methodology Analysis modules.

Data acquisition, transformation, and compilation methodology

A dedicated analysis tier produces JSON-formatted output, which is processed in Python for data analysis, anomaly detection, and missing data imputation (Lapāns, 2025). For this purpose, a Row-Wise Z-Score Filtering method was developed. Row-level outliers are identified and removed using a configurable z_threshold, after which missing values are imputed column by column based on multivariate patterns across rows, treating each column with missing data as a target and the remaining columns as predictors. This procedure requires that rows contain compatible and comparable values, avoiding inconsistent units (e.g. mixing MW with MWh or EUR/MWh with kWh/m²).

As shown in the Table 2, the sectoral distribution of missing data reveals substantial heterogeneity in both completeness and temporal stability over the 2015–2023 period.

Sector	Range of missing data between 2015-2023 (%)	Average z-score (2015-2023)
Energy	27.45 - 73.86	1.66
Agriculture	5.30 - 64.80	-0.22
LULUCF	2.22 - 8.89	-0.52
Waste management	19.95 - 27.32	0.39
Industry	0.00 - 4.17	-0.76

Source: Authors

Table 2: Summary of missing data and z-score.

The industrial sector exhibits the most robust reporting profile, maintaining very low missingness throughout the study period and achieving complete data coverage in several years. The transport sector also shows a favourable trajectory, with missing data steadily decreasing from approximately 10% in the earlier years to full completeness in 2022, followed by only a minimal increase in 2023. In contrast, the waste management sector is characterised by persistently high missingness, remaining consistently above 19% and indicating long-term structural limitations in data collection or reporting. The LULUCF sector remains comparatively stable at a low to moderate level of missingness, although a gradual upward trend suggests a modest erosion in data completeness over time. The agricultural sector displays a markedly different pattern: while missingness remains moderate and relatively stable through 2022, the abrupt rise to 64.8% in 2023 constitutes a clear structural discontinuity and the most severe single-year deterioration among the analysed sectors. The energy sector represents the most consistently incomplete dataset overall, with exceptionally high missingness across all years; however, unlike the agricultural sector, it shows a gradual improvement over time, declining from 73.86% in 2015 to 27.45% in 2023. Taken together, these findings indicate that data quality is strongly sector-dependent, with the industrial sector and the transport sector demonstrating reliable reporting, the waste management and the energy sectors reflecting persistent systemic incompleteness, and the agricultural sector emerging as a critical outlier in the most recent year.

When using z-score to measure how much a certain score differs from the typical score (Berendrecht et. al., 2022), it indicates such large differences between years that additional attention should be paid to why there are inconsistencies in the data only for the energy sector, which is the only one where the z-score exceeds the threshold of 1.

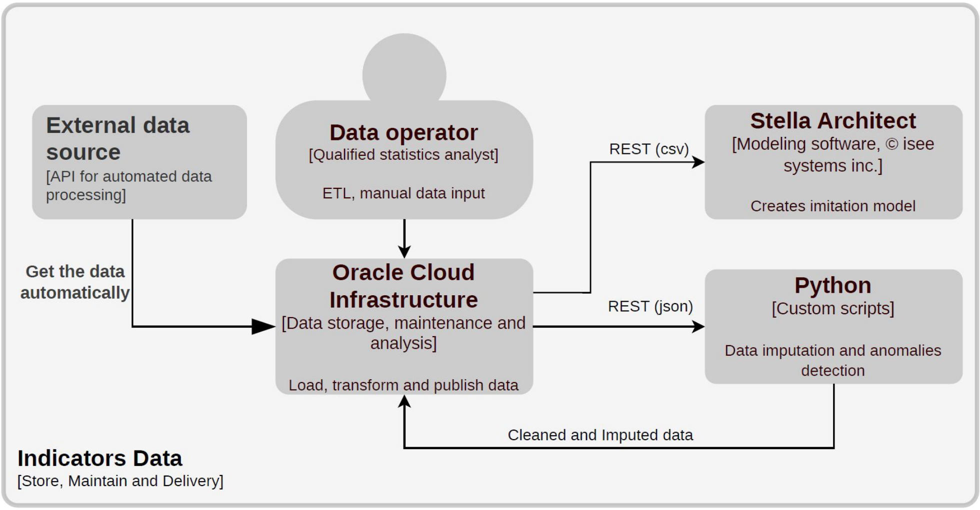
To handle heterogeneous data sources, both ETL (Extract, Transform, Load) and ELT (Extract, Load, Transform) workflows are employed. ETL is applied when sources are noisy or inconsistently structured, with transformations performed in Python before loading. ELT is used for well-structured sources, where data are first loaded into the database and then transformed using SQL and PL/SQL. Combining ETL and ELT ensures an adaptable and efficient pipeline for generating model-ready inputs.

Data storage and exchange prototype

Oracle Cloud Infrastructure (OCI) with an Oracle APEX 24.2.10 cloud-based database was selected as the integrated environment for pre-loading, processing, and preparing results for all indicator categories. This setup provides a unified, systematically organized data storage solution and supports simultaneous access by multiple users.

The integrated data retrieval mechanism (see Figure 1) offers two main outputs tailored to different use cases. First, a downloadable CSV file is generated and specifically formatted for direct use n the STELLA® modelling software, following the import and export specifications provided by Isee Systems (isee systems, 2025) to ensure error-free integration with the simulation model. Second, a visually rich HTML page presents the same data in a structured, user-friendly format, enabling interactive browsing and visual inspection of data quality. This interface allows users to assess completeness, identify anomalies, and explore trends across datasets.

Both outputs are delivered via RESTful web services, enabling online access and seamless integration into external systems or analytical workflows. This dual-output design maintains a clear separation between storage, processing, and application layers - data are securely stored in a central database, while REST-based interfaces



Source: Authors

Figure 1: A graphical overview of the end-to-end workflow for data collection, maintenance, and delivery, summarizing the processes described above in a single integrated scheme.

provide standardized, system-independent access for external tools. This architecture supports reliable data integration, scalability, and interoperability across the modelling workflow.

Key findings and data challenges

The findings of this study confirm that data acquisition for climate neutrality modelling in Latvia is a complex and resource-intensive task. This is mainly due to (1) fragmented data ownership, (2) heterogeneous formats, (3) limited availability of key datasets, and (4) restricted accessibility for researchers. Although various institutions collect important indicators, data resolution is often insufficient for comprehensive analysis. For example, total building floor area may be reported, but systematic data on technical condition, renovation status or energy performance class are not publicly available. Many relevant data sources are also published in PDF or Word formats (e.g., inventory reports) or as narrative text and infographics on institutional websites, which are unsuitable for automated extraction and must largely be processed manually. Automated web scraping is further constrained by dynamically rendered content and graphical images that hinder optical character recognition.

Additional challenges arise from data made available in forms that are not amenable to systematic, automated processing. For instance, information on renovated apartment buildings is often provided only as visual annotations on public GIS maps rather than downloadable structured files. Extracting this data requires bespoke scripts and introduces uncertainty, as points may be inaccurately positioned, lack metadata or don’t correspond to actual building addresses, thus complicating spatial analysis and linkage to other datasets.

Heterogeneous formats, outliers, and incomplete records during transformation required careful handling. Nevertheless, cleaning and imputation significantly improved usability by standardizing input, reducing noise, and producing a dataset further suitable for modelling.

The overall quality of the processed data remains mixed. Some variables show low levels of missing values and anomalies, whereas others require extensive imputation and are therefore less robust. These issues reveal persistent structural weaknesses in Latvia’s data governance environment, including the absence of unified national standards for statistical metadata and online publication formats, and limited integration between sectoral and cross-sectoral information systems. Although the data ecosystem has improved in recent years, substantial methodological and institutional advances are still needed to achieve fully interoperable, high-resolution datasets capable of supporting complex climate policy modelling.

Implications for climate neutrality modelling and policy

The structured data mapping system and metadata catalogue developed in this study provide a systematic overview of data inputs for climate neutrality modelling. By documenting (1) source institution, (2) temporal coverage, (3) territorial scope, (4) data format, (5) accessibility, and (6) methodological notes, of each dataset, the framework strengthens transparency and facilitates future extensions of the data infrastructure. This level of documentation is essential for reproducible modelling and for institutional learning about data governance.

The testing of anomaly detection and missing-value imputation frameworks using Row-Wise Z-score filtering and an iterative Random Forest–based imputer (Stekhovn and Bühlmann, 2012) demonstrates both the potential and the limits of automated quality enhancement. While historical patterns and cross-source correlations can improve completeness, their effectiveness is restricted by short and fragmented time series. For policy-relevant modelling, this implies that hybrid approaches—combining automated methods with rule-based corrections and expert judgement—are necessary to balance completeness and reliability. Importantly, gaps without sufficient contextual information were deliberately left flagged rather than filled, which supports honest communication of uncertainty.

The results show that re-imputed values are relatively robust for variables with limited data loss, but much less so for variables with high proportions of missing or flagged values. For climate neutrality modelling and policy analysis, this means that some indicators can be used with moderate confidence, while others should be treated as exploratory and subjected to sensitivity analysis or explicit caveats in scenario interpretation. Selective and data-conscious cleaning is therefore a critical component of responsible model building.

By compiling source-level and project-level metadata and linking them to a unified data model, the study creates a dataset that is (1) traceable, (2) interpretable, and (3) auditable. The integrated time series for 2015–2023 enables cross-source comparisons and cross-sectoral analyses and reveals an improving trend in metadata completeness over time. This strengthens the empirical foundation for an integrated national climate neutrality model.

Finally, the harmonized data from the five sectors—energy, industry, agriculture, land use, LULUCF, and waste management—provide the necessary inputs to parameterize and run the system dynamics (SD) model, capturing sectoral interactions and temporal dynamics. Implemented as a decision-impact support system with a user-friendly interface, the SD model will enable users to design potential climate policies and evaluate their effects on GHG emission reductions, required investments, and other key impacts. At the same time, the documented data limitations underscore the need for cautious interpretation of model outputs and for continuous improvement of the underlying data systems.

Looking ahead, the data framework and integrated dataset developed in this study provide a scalable foundation for further enhancement of Latvia's climate neutrality modelling capacity. Future work should focus on institutionalizing data mapping and quality assessment procedures, expanding machine-readable publication of key indicators, and progressively reducing reliance on manually processed sources. In parallel, iterative coupling between the SD model and the evolving data infrastructure can support continuous improvement of both modelling accuracy and the evidence base for climate policy design.

Conclusion

The results of this study highlight that improving data infrastructure for climate neutrality modelling in Latvia requires both institutional and technical measures. Based on the identified challenges and the experience gained in building the integrated dataset, the following recommendations are proposed.

Improve data standards and publication formats:

- Develop national standards for statistical metadata, units, and indicator definitions relevant to climate neutrality
- Require key datasets to be published in machine-readable, structured formats (e.g. CSV, JSON, GIS vector files) in addition to reports in PDF/Word
- Ensure better integration between sectoral and cross-sectoral systems through harmonized classifications and exchange formats

Implement automated and modular data ingestion:

- Begin with automated API ingestion using scheduled scripts or ETL tools to enable real-time or periodic updates from primary providers
- For structured files (CSV, XLSX, GIS exports), implement batch processing pipelines that validate, normalize, and convert data into a schema aligned with the integrated database model

- For unstructured sources, apply specialized parsing, OCR and, where appropriate, NLP for PDFs and reports; use web scraping frameworks with content extraction logic for web pages, subject to legal and ethical constraints
- Each ingestion module should include metadata tagging, error logging, and version control to ensure traceability and support quality control

Centralize and harmonize storage:

- Store all processed data in an integrated database, using standardized transformation rules and mapping logic to ensure coherence across formats and sources
- Maintain a clear separation between storage, processing, and application layers, allowing different modelling and analytical tools to access a single authoritative data source
- Strengthen quality management and transparency
- Apply a formal data quality assessment framework (completeness, consistency, comparability, accuracy) and clearly flag variables with extensive imputation or uncertainty
- Combine automated anomaly detection and imputation with expert review, specifically for critical indicators and fragmented time series.

Facilitate access for modelling and policy analysis:

- Provide model-ready outputs via RESTful web services and preformatted CSVs for specific tools (e.g. STELLA®)
- Offer simple web-based interfaces for exploring completeness, anomalies, and trends to support both modelers and policy users.

Taking together, these measures would address many of the structural and technical barriers identified in this study. Implementing them would move Latvia closer to having fully interoperable, high-resolution, and transparently documented datasets capable of supporting robust system dynamics modelling and evidence-based climate policy design.

Acknowledgements

This research was funded by the State Research Programme, Latvia project Climate Neutrality Decision Models in Action, VPP-KEM-Klimatneitralitāte-2023/1-0002.

Corresponding author:
Ph.D. Oskars Java
Vidzeme University of Applied Sciences, Tērbata street 10, Valmiera, LV-4202, Latvia
Phone: +37128678558, Email: oskars.java@va.lv

References

[1] Berendrecht, W., van Vliet, M. and Griffioen, J. (2022) "Combining statistical methods for detecting potential outliers in groundwater quality time series", *Environmental Monitoring and Assessment*, Vol. 195, No. 1, pp. 1-14. ISSN 1573-2959. DOI 10.1007/s10661-022-10661-0.

[2] Cabinet of Ministers. (2023) "*Par valsts pētījumu programmu "Klimatneitralitātes mērķu sasniegšanas lēmumu pieņemšanas atbalsta sistēma" 2023.– 2025*", gadam. Sep. 2023. [Online]. Available: https://likumi.lv/ta/id/345896-par-valsts-petijumu-programmu-klimatneitralitates-merku-sasniegsanas-lemumu-pienemsanas-atbalsta-sistema-2023-2025. [Accessed: Sep. 3, 2025]. (In Latvian).

[3] Intergovernmental Panel on Climate Change (IPCC). (2019) "*2019 refinement to the 2006 IPCC guidelines for national greenhouse gas inventories*". IPCC. [Online]. Available: https://www.ipcc.ch/report/2019-refinement-to-the-2006-ipcc-guidelines-for-national-greenhouse-gas-inventories/. [Accessed: Sep. 3, 2025].

[4] Intergovernmental Panel on Climate Change (IPCC). (2023) "Linking Global to Regional Climate Change", In *Climate Change 2021 – The Physical Science Basis: Working Group I Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*, pp. 1363-1512). ISBN 9781009157896. DOI 10.1017/9781009157896.

[5] Figueroa, V. S. and Farias, C. B. (2022) "Carbon Neutrality Scenario Projection Model under Uncertainty", *2022 IEEE International Conference on Automation/XXV Congress of the Chilean Association of Automatic Control (ICA-ACCA)*, Vol. 2022, pp. 1-6. E-ISBN 978-1-6654-9408-3. DOI 10.1109/ICA-ACCA56767.2022.10005931.

[6] Ghent, D., Kaduk, J., Remedios, J. and Balzter, H. (2011) "Data assimilation into land surface models: The implications for climate feedbacks", *International Journal of Remote Sensing*, Vol. 32, No. 3, pp. 617-632. E-ISSN 1366-5901. DOI 10.1080/01431161.2010.517794.

[7] Isee systems. (2025) "*Creating Import and Export Files*". [Online]. Available: https://www.iseesystems.com/resources/help/v3/#03-buildingmodels/importing_and_exporting_data/creating_import_and_export_files.htm. [Accessed: Oct. 17, 2025].

[8] Krysovatyy, A., Maksymova, I., Kurilyak, V., Radin, M. and Kurilyak, M. (2024) "International convergence towards a climate-neutral economy: modeling the agricultural sector", *AgEcon SEARCH*, Vol. 10, No. 2, pp. 52-79. ISSN 2414-584X. DOI 10.22004/ag.econ.355962.

[9] Lapāns, A. (2025) "*Energy Data Imputation and Anomaly Detection System*", Nov. 2025. [Online]. Available: https://Github.Com/Ala-Hub-25/Vpp-Kem/. [Accessed: Nov. 12, 2025].

[10] Mikropoulos, E., Roelfsema, M., Chen, H. H., Staffell, I., Oreggioni, G., Hdidouan, D., Thellufsen, J. Z., Chang, M. A., Fragkos, P., Giannousakis, A., Chatterjee, S., Ürge-Vorsatz, D., Pfenninger, S., Pickering, B., Victoria, M., Brown, T. and van Vuuren, D. P. (2025) "Examining pathways for a climate neutral Europe by 2050; A model comparison analysis including integrated assessment models and energy system models", *Energy*, Vol. 319. pp. 1-18. E-ISSN 1873-6785. DOI 10.1016/j.energy.2025.134809.

[11] O'Keeffe, S., O'Sullivan, D. and Bruton, K. (2023) "Development of a modelling tool to aid the transition to carbon neutral industrial manufacturing", *Journal of Cleaner Production*, Vol. 425, pp. 1-19. ISSN 0959-6526. DOI 10.1016/j.jclepro.2023.138604.

[12] Prestele, R., Arneth, A., Bondeau, A., de Noblet-Ducoudré, N., M Pugh, T. A., Sitch, S., Stehfest, E. and Verburg, P. H. (2017) "Current challenges of implementing anthropogenic land-use and land-cover change in models contributing to climate change assessments", *Earth System Dynamics*, Vol. 8. No. 2, pp. 369-386. E-ISSN 2190-4987. DOI 10.5194/esd-8-369-2017.

[13] Cabinet of Ministers. (2022) "*Procedures for Establishing and Maintaining the System for Greenhouse Gas Inventories, the Projection System, and the System for Reporting on the Adaptation to Climate Change*". [Online]. Available: https://likumi.lv/ta/en/en/id/336733-procedures-for-establishing-and-maintaining-the-system-for-greenhouse-gas-inventories-the-projection-system-and-the-system-for-reporting-on-the-adaptation-to-climate-change. [Accessed: Nov. 10, 2025].

[14] Randall, D. A., Wood, R. A., Bony, S., Colman, R., Fichefet, T., Fyfe, J., Kattsov, V., Pitman, A., Shukla, J., Srinivasan, J., Stouffer, R. J., Sumi, A. and Taylor, K. E. (2007) "Climate models and their evaluation", In Solomon, S., Qin, D., Manning, M., Chen, Z., Marquis, M., Averyt, K. M. Tignor, M. and Miller, H. L. (Eds.) "*Climate change 2007: The physical science basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change",* pp. 589-662. Cambridge University Press. ISBN 978 0521 88009-1.

[15] Samarasinghe, S. L., Moghimi, M. and Kaparaju, P. (2025) "A review of modelling tools for net-zero emission energy systems, based on model capabilities, modelling criteria and model availability", *Renewable Energy Focus*, Vol. 53, pp. 1-24. ISSN 1755-0084. DOI 10.1016/j.ref.2024.100659.

[16] Schnase, J. L., Lee, T. J., Mattmann, C. A., Lynnes, C. S., Cinquini, L., Ramirez, P. M., Hart, A. F., Williams, D. N., Waliser, D., Rinsland, P., Webster, W. F., Duffy, D. Q., McInerney, M. A., Tamkin, G. S., Potter, G. L. and Carrier, L. (2016) "Big Data Challenges in Climate Science", *IEEE Geoscience and Remote Sensing Magazine*, Vol. 4, No. 3, pp 10-22. E-ISSN 2168-6831. DOI 10.1109/MGRS.2015.2514192.

[17] Skujevska, D., Feofilovs, M. and Romagnoli, F. (2024) "Challenges in Standardizing Global Emission Factors for Peatlands", *CONECT. International Scientific Conference of Environmental and Climate Technologies*, pp. 44-45. E-ISSN 2592-9704. DOI 10.7250/conect.2024.028.

[18] Stekhoven, D. J. and Bühlmann, P. (2012) "MissForest: non-parametric missing value imputation for mixed-type data", *Bioinformatics*, Vol. 28, No. 1, pp. 112-118. E-ISSN 1367-4811. DOI 10.1093/bioinformatics/btr597.

[19] Ulpiani, G., Vetters, N., Melica, G. and Bertoldi, P. (2023) "Towards the first cohort of climate-neutral cities: Expected impact, current gaps, and next steps to take to establish evidence-based zero-emission urban futures", *Sustainable Cities and Society*, Vol. 95. pp. 1-20. E-ISSN 2210-6715. DOI 10.1016/j.scs.2023.104572.

[20] Zargar, S., Kannangara, M., Gonzales-Calienes, G., Yang, J., Shadbahr, J., Decès-Petit, C. and Bensebaa, F. (2024) "Data Hub for Life Cycle Assessment of Climate Change Solutions—Hydrogen Case Study", *Data*, Vol. 9, No. 11. ISSN 2306-5729. DOI 10.3390/data9110129.

Agris on-line Papers in Economics and Informatics

Volume XVIII Number 3, 2026

Agris on-line Papers in Economics and Informatics

The international reviewed scientific journal issued by the Faculty of Economics and Management of the Czech University of Life Sciences Prague.

http://online.agris.cz

ISSN 1804-1930