Enhancing data accuracy in agri-food forecasting: Methods and implications for informed decision-making*

László Várallyai[1,†], Szilvia Botos[1,*,†], Levente P. Bálint[1,†], Viktor L. Takács[1,†] and Róbert Szilágyi[1,†]

[1] *University of Debrecen, Faculty of Economics, Institute of Methodology and Business Digitalization, 138 Böszörményi út, 4032 Debrecen, Hungary*

Abstract

In our article, we aim to demonstrate how enterprises can apply secondary economic data and use methodologies for trend analysis and forecasting. By utilizing secondary databases, organisations can effectively evaluate market stability and conduct comprehensive industry analyses. This approach not only enhances the accuracy of their assessments but also supports strategic decision-making processes.

Keywords

Forecast methodology, food industry, trend analysis

1. Introduction

Europe's Digital Decade policy and the Digital Europe Programme both clearly aim to integrate digital solutions (like AI or BigData analytics) into business processes throughout the EU, in order to enhance the operational performance of business organisations. With the help of open source softwares companies will be able to overcome the general digital technology implementation barriers to help them in achieving a new digital corporate strategy.

2. Literature review

According to the targets of the European Union, several strategies, policies and initiatives have been prepared and the development of the digital economy and society is a significant part of it. With the concept of Industry 5.0 and Agriculture 5.0, the stakeholders operating in agriculture and the food industry also have the potential to become one of the biggest users of technologies based on open-source solutions with various advantages.

Missing data is a frequent issue across various fields, not just in academia, which can occur due to several factors and can lead to misleading information and incorrect decisions or results (Gjorshoska et al., 2022). In recent years, there has been an increasing interest in applying predictive analytical methods (like SARIMA) in the field of supply chain management (Kumari & Muthulakshmi, 2024). The primary aim of using these technics was demand and supply forecasting (for prices and volumes) in purchasing functions (Falatouri et al., 2022). The presence of missing data in a time series can greatly affect model performance by interrupting data continuity, making it an

* *Short Paper Proceedings, Volume I of the 11th International Conference on Information and Communication Technologies in Agriculture, Food & Environment (HAICTA 2024), Karlovasi, Samos, Greece, 17-20 October 2024.*

* Corresponding author.

† These authors contributed equally.

varallyai.laszlo@econ.unideb.hu (L. Várallyai); botos.szilvia@econ.unideb.hu (S. Botos); balint.peter.levente@econ.unideb.hu (L. Bálint); takacs.viktor@econ.unideb.hu (V. Takács); robert.szilagyi@econ.unideb.hu (R. Szilágyi)

0000-0002-0795-9527 (L. Várallyai); 0000-0003-4873-1032 (S. Botos); 0009-0008-4913-2449 (L. Bálint); 0000-0001-8433-6115 (V. Takács); 0000-0002-1783-6483 (R. Szilágyi)

important area of research (Lee et al., 2024). Understanding and pinpointing the reasons behind missing data is always essential when working with any datasets.

To prepare our literature review, we first performed a qualitative analysis (Figure 1) on the relevant keywords (agriculture, open source and python), as this method is suitable for determining the direction of the research. We used an international literature database (Scopus) for the analysis. We defined the most relevant keywords closely related to the research.

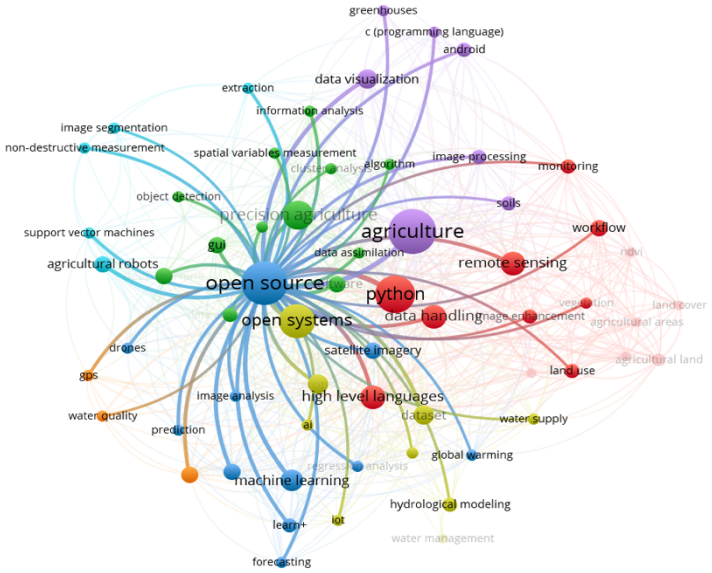

Figure 1: The relationship among the author keywords. Source: own construction based on Scopus (2024) results

3. Individual readiness for data analysis and opportunities

Because of the many operating processes with digitalized data (transaction data, Internet of Things data etc.) recording there are a lot of data ready to be transformed into data warehouses and analyzed. The analysis of these data helps in identifying inefficient points of the business processes and finding bottlenecks. The quantity and quality of these types of data can also be used in advanced data analytics, such as Machine Learning. These data support many decisions related to optimization of material, financial and information resources, increasing cooperation and forecasting.

However, for advanced data analytics, there are many requirements from the side of the human resources as well. Figure 2 and 3 show two relevant indicators that express the readiness levels of individuals for advanced data analysis.

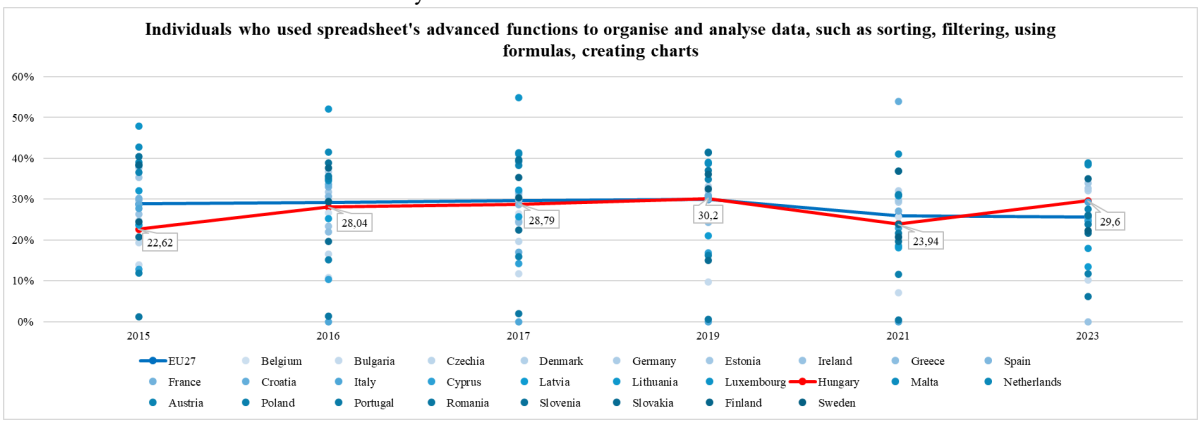

Figure 2: Individual's readiness for advanced data analysis. Data source: own construction based on Eurostat (2024)

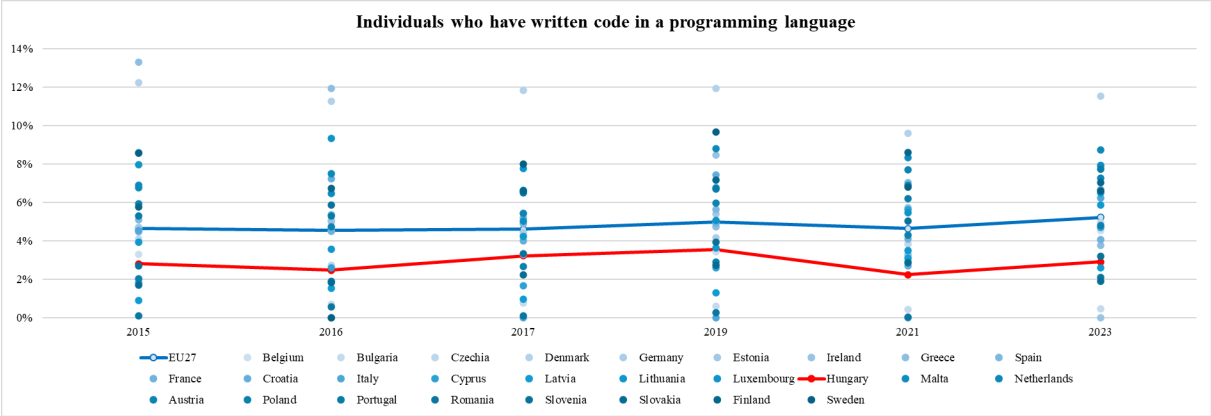

Figure 3: Individual's readiness for writing programming codes. Data source: own construction based on Eurostat (2024)

Enterprises have access to a variety of open-source tools designed for data analytics. These applications offer budget-friendly options and can be adapted to fit unique business requirements. Based on the numbers on Figure 3, we can conclude that there is a need for writing programming codes that can be considered as a necessary digital skill in the near future.

4. SARIMA results

In this article we present a possible forecasting model that can be used with time series where seasonal variability might occur. In our results we describe a performance analysis of the applied seasonal autoregressive integrated moving average methods and how we implemented Python-based data-preprocessing algorithms.

The obtained secondary product volume time series data from the Hungarian agricultural product information system is stored in a publicly available database. As there were changes made in how the collected data is stored in categories over the years, some transformations were applied on the data source, which were carried out using Python Data Analysis Library (Pandas).

The SARIMA methodology employed is this paper is a statistical model that analyzes time series data to enhance the understanding of the dataset or forecast future trends better (Abeladi et al., 2023).

SARIMA is a type of regression analysis that assesses the relationship between one dependent variable and other varying factors.

The SARIMA model includes several parameters to measure seasonality, like frequency, integration, MA(Q) and AR(p) orders.

Before applying the forecasting method, it is crucial to verify the stationarity of the original dataset. This can be done using the ADfuller test, which provides a significance level through hypothesis testing. The test results in a p-value that helps determine whether the time series is stationary. If the p-value is below 0.05, the time series can be considered stationary. Conversely, if the p-value exceeds 0.05, the time series is deemed non-stationary (Tatarintsev et al. 2021).

The data source has been checked for suitability for the SARIMA model and is applicable to use. The data was obtained from and trained on the Hungarian Agriculture Information Portal (2024)'s dataset.

4.1. Performing the SARIMA

In our study, to choose the most appropriate model, the AIC (Akaike Information Criterion) was utilized (Vadim & Alchakov, 2023). After the model optimalization process the ARIMA(2,1,1)(1,0,1)[12] model was selected. Figure 4 shows our original dataset and the predicted model data.

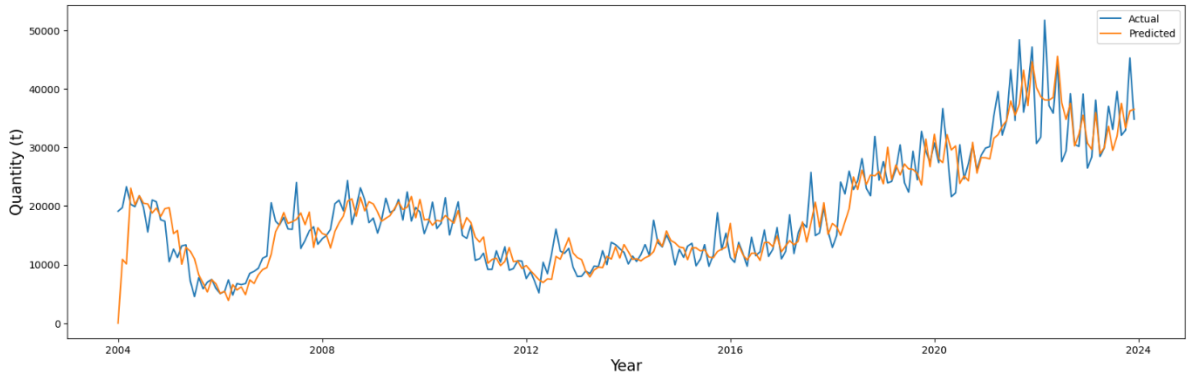

Figure 4: Quantity values Actual data vs Predicted SARIMA model. Source: own editing based on model results

Holt-Winters forecasting is a statistical method used for time series data that exhibits trends and seasonality. It's particularly effective when the underlying pattern in the data is not linear and has recurring seasonal fluctuations. Holt-Winter's Exponential Smoothing method is used to examine the data to identify if it exhibits trends or seasonal patterns by analyzing the overall pattern (Pongdatu & Putra, 2018). There are 3 key components: level- represents the overall average value of the time series, trend - captures the upward or downward trend in the data over time, seasonality- accounts for the periodic fluctuations that occur at regular intervals. In this process we created a Holt-Winters forecasting model with a multiplicative trend and seasonality.

The model was fitted to the training data ("training Q data"), and the test set was generated ("test q data") for the forecast. Finally, the predicted values were created ("predicted Q data") Figure 5.

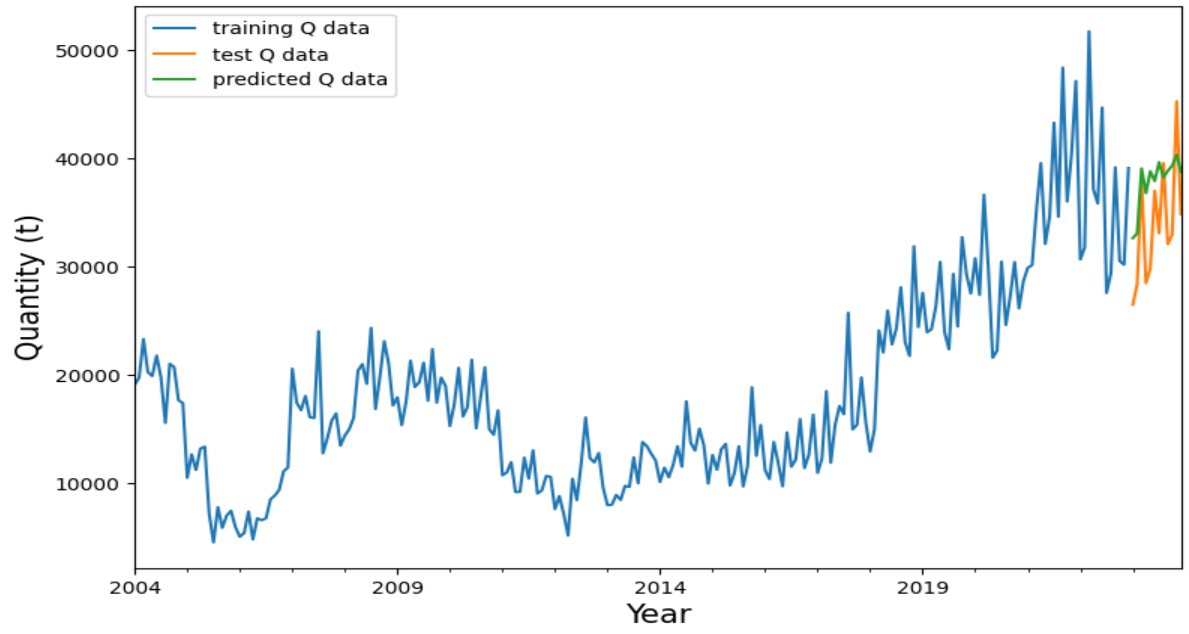

Figure 5: Quantity values by Holt-Winters Forecasting, training, test and predicted data. Source: own editing

5. Model assessment

Table 1 summarizes the assessment of our model (accuracy) with five statistical metrics.

Table 1
Assessment table

Statistic metric	Value
R^2	0.6687
MSE	31859073.0
RMSE	5644.0
MAE	4993.0
MAPE	15.8236

Source: own construction based on calculations

Based on these metrics, the SARIMA model shows moderate accuracy: the R^2 of 0.6687 indicates a reasonable fit, while MSE, RMSE, MAE, and MAPE values suggest that there is some level of error in predictions, but the model's performance is generally acceptable with room for improvement.

6. Conclusions

In conclusion, making well-informed, data-driven decisions is essential for modern businesses, particularly in the food industry, where data analytics plays a crucial role in enhancing sustainability, innovation, competitiveness, and resilience.

In this paper, a quantity forecast was performed for wheat flour (BL55) using Python's in-build SARIMA model, as an open-source digital solution.

Our article illustrates how enterprises can leverage secondary economic data and employ trend analysis and forecasting methodologies to improve market stability and strategic decision-making. By addressing and filling missing values in production-related data, organizations can avoid biases and distortions that could otherwise lead to flawed conclusions and ineffective resource allocation.

For instance, in agriculture, real-time data from sensors can provide insights into crop health, soil conditions, and environmental factors, which are essential for optimizing yields and managing resources efficiently. Accurate market price and volume information helps in forecasting demand, setting prices, and adjusting production levels accordingly.

Our application of the SARIMA method on historical market data demonstrates how accurate estimation of missing information supports better market forecasts, optimizes production and inventory strategies, and enhances cost planning and investment decisions.

Ultimately, this approach fosters more informed decision-making and contributes to greater efficiency and profitability in the agri-food sector.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] Abeladi, K., Zafar, B., & Mueen, A. (2023). Time Series Forecasting using LSTM and ARIMA. International Journal of Advanced Computer Science and Applications, 14(1), 313–320. https://doi.org/10.14569/IJACSA.2023.0140133.

[2] Eurostat, 2024. Overview. URL: https://ec.europa.eu/eurostat/web/digital-economy-and-society/database/comprehensive-database

[3] Hungarian Agriculture Information Portal, 2024. URL: https://adat.aki.gov.hu/

[4] Ivana Gjorshoska, Tome Eftimov and Dimitar Trajanov, 2022. Missing value imputation in food composition data with denoising autoencoders. Journal of Food Composition and Analysis. Vol. 112. article 104638, ISSN 0889-1575. https://doi.org/10.1016/j.jfca.2022.104638.

[5] Kyungjae Lee, Hyunwoo Lim, Jeongyun Hwang, and Doyeon Lee, 2024. Evaluating missing data handling methods for developing building energy benchmarking models. Energy. Vol. 308. article 132979. ISSN 0360-5442. https://doi.org/10.1016/j.energy.2024.132979.

[6] Pongdatu, G.A.N. and Putra, Y.H. (2018): Time Series Forecasting using SARIMA and Holt Winter's Exponential Smoothing. IOP Conf. Ser. Mater. Sci. Eng. 407, 012153. https://doi.org/10.1088/1757-899X/407/1/012153.
[7] Shabnam Kumari, and P. Muthulakshmi, 2024. SARIMA Model: An Efficient Machine Learning Technique for Weather Forecasting. Procedia Computer Science. Vol. 235. pp. 656-670. ISSN 1877-0509. https://doi.org/10.1016/j.procs.2024.04.064.
[8] Taha Falatouri, Farzaneh Darbanian, Patrick Brandtner and Chibuzor Udokwu, 2022. Predictive Analytics for Demand Forecasting – A Comparison of SARIMA and LSTM in Retail SCM. Procedia Computer Science. Vol. 200. pp. 993-1003. ISSN 1877-0509. https://doi.org/10.1016/j.procs.2022.01.298.
[9] Tatarintsev, M., Korchagin, S., Nikitin, P., Gorokhova, R., Bystrenina, I. and Serdechnyy, D. (2021). Analysis of the forecast price as a factor of sustainable development of agriculture. Agronomy, 11(6). https://doi.org/10.3390/agronomy11061235.
[10] Vadim K. and Alchakov V. 2023. "Time-Series Forecasting of Seasonal Data Using Machine Learning Methods" Algorithms 16, no. 5: 248. https://doi.org/10.3390/a16050248.