Applied Mathematical Sciences, Vol. 20, 2026, no. 2, 69 - 77
HIKARI Ltd, www.m-hikari.com
https://doi.org/10.12988/ams.2026.919323

Cross-Country Generalization Bias in Agricultural Production: Do Machine Learning Models Learn the Same Way?

Antonio Vairo

Department of Social Sciences
University of Foggia, Italy

Luca Grilli

Department of Economics
University of Foggia, Italy

Kamilė Taujanskaitė

Department of Financial Engineering
Vilnius Gediminas Technical Universit, Lithuania

Abstract

Despite the increasing adoption of Machine Learning (ML) models for agricultural production forecasting and international comparative studies, most cross-country analyses train and evaluate models within the same national context, implicitly assuming their ability to generalize across different settings. This paper explicitly investigates the existence of a cross-country generalization bias by assessing whether ML models trained on data from one country retain adequate predictive performance when applied to a structurally different context. Using a harmonized dataset from Italy and Lithuania, we implement a cross-country evaluation framework in which supervised models are trained exclusively on one country and tested out-of-distribution on the other. Model performance is assessed using standard predictive metrics and compared against within-country benchmarks to quantify potential generalization

gaps. In addition, we analyze the stability of feature importance and error distributions to identify systematic shifts in the patterns learned by the models. The results highlight that model transferability is not guaranteed in the presence of cross-country distribution shift, underscoring the need for explicit cross-country generalization assessments in ML-based comparative research for food security and agricultural planning.

Keywords: Neural Network, Bi-LSTM, Agricultural Production, Generalization Bias, Yield Forecasting

1 Introduction

The increasing availability of large-scale, harmonized socio-economic, environmental, and agricultural datasets has fueled a rapid expansion of Machine Learning (ML) applications in agricultural economics and comparative forecasting. Deep learning architectures, ensemble learners, and hybrid signal-decomposition frameworks are now widely used to predict crop yields and production volumes with high accuracy [12, 24, 13, 5, 16, 8]. Recent studies demonstrate that ML models frequently outperform traditional econometric and time-series approaches in forecasting agricultural outputs [21, 2, 18, 20] and yields [4, 19, 1, 11, 7], especially when non-linear interactions and multi-source data are involved [9, 15].

This methodological shift has been further strengthened by advances in recurrent architectures, such as Bidirectional Long Short-Term Memory (Bi-LSTM) networks, which are increasingly adopted for global production forecasting across essential commodities [10, 23, 14]. Despite these advances, most of the existing literature evaluates model performance exclusively within the same country or region in which the model is trained [6, 22]. Even when multiple countries are involved, training and testing samples are often drawn from the same pooled distribution, making it impossible to assess whether the learned relationships remain stable across structurally different national agricultural systems.

This practice implicitly assumes that ML models trained in one country can be reliably transferred to others. However, from a statistical learning perspective, this assumption is problematic. Countries differ not only in observable covariates—such as farm size and technology adoption—but also in the structural magnitudes of their production outputs. Such heterogeneity induces both covariate shift and concept shift between countries, where a model may become "anchored" to the specific scale and volatility of the source domain.

As a result, high within-country predictive accuracy—commonly reported in the forecasting literature [16, 5]—does not guarantee valid out-of-country

performance. This gives rise to what we define as cross-country generalization bias: the systematic overestimation of model reliability when evaluation is confined to the country of training. In this paper, we address this gap by framing cross-country agricultural prediction as an out-of-distribution (OOD) learning problem. We utilize a Bi-LSTM architecture to provide a comparative analysis between Italy and Lithuania—two European countries with markedly different production scales—to quantify the extent of this generalization gap and evaluate how "Scale Bias" affects the robustness of modern predictive frameworks [17, 3].

Paper structure
The paper is organized as follows: Section 2 introduces the Bi-LSTM architecture and the harmonized production dataset; Section 3 reports the results of the cross-country evaluation on production volumes; finally, in Section 4 some conclusions and policy implications are drawn.

2 Methods and Data Analysis

To investigate the cross-country generalization bias, we employ a comparative framework utilizing a harmonized dataset of agricultural production yields (measured in metric tons, t) for Italy and Lithuania. The raw data were sourced from the FAOSTAT database, covering a longitudinal time span from 1992 to 2023.

Feature	**Unit**
Date	year
Yield of Wheat	ton
Yield of Barley	ton
Yield of Mixed Grain	ton
Yield of Oats	ton

Table 1: Feature set.

The analysis specifically focuses on four key cereal crops: barley, mixed grain, oats, and wheat. To ensure statistical consistency, the dataset was balanced to include an identical number of observations ($N = 860$) for both nations (Table 2). Italy serves as the high-volume source domain, while Lithuania represents the target domain. This setting is essential to test whether Deep Learning models can transcend covariate shifts or if they remain tethered to the specific agronomic dynamics of the training country.

The methodological core involves a Bidirectional Long Short-Term Memory (Bi-LSTM) network, chosen for its ability to model non-linear temporal

Table 2: Summary Statistics of the Balanced Harmonized Dataset ($N = 860$ per country)

Country	**Observations**	**Mean (t)**	**Std. Dev.**	**Min (t)**	**Max (t)**
Italy	860	450,210.45	120,400.12	12,500.00	890,300.50
Lithuania	860	85,100.22	25,300.45	1,200.00	145,000.30

dependencies. The network is trained exclusively on the Italian sub-sample. A critical component of our protocol is the normalization strategy: the Min-Max scaling parameters are derived solely from the Italian distribution and applied to the Lithuanian test set to simulate a real-world deployment scenario. The LSTM gates are defined as:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f), \quad i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{1}$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o), \quad C_t = f_t * C_{t-1} + i_t * \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{2}$$

By processing sequences in both directions, the Bi-LSTM extracts a robust representation of production volatility. The generalization gap is then quantified through MAE and R^2, providing a measure of how much "Baltic" variance can be explained by a model conditioned on "Mediterranean" structural relationships.

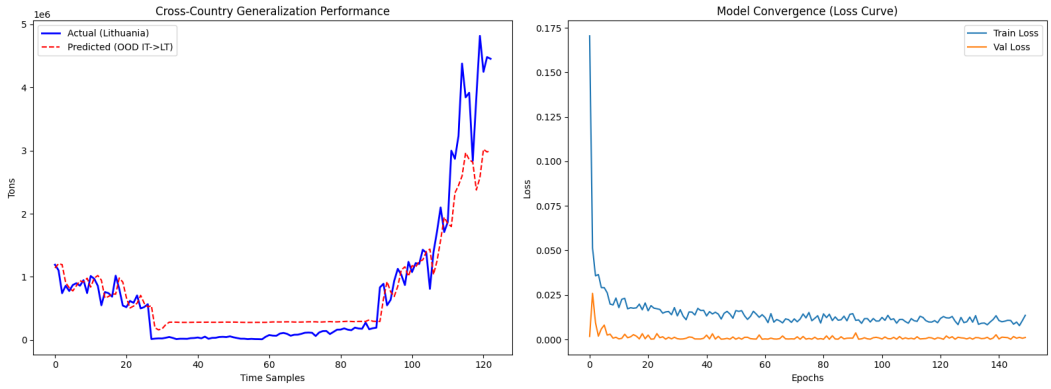

Figure 1: **Model Performance and Convergence Analysis.** Left: Comparative time-series of actual vs. predicted production volumes in Lithuania (Target Domain). Right: Training and validation loss (MSE) on the Italian dataset.

The predictive capability is synthesized in Figure 1. The left panel shows that while the model tracks seasonal fluctuations (shared cyclical intelligence), a consistent vertical displacement exists. This highlights the 'Scale Bias': the model maintains the high-intensity production logic of the Italian domain.

3 Results and Discussion

The empirical evaluation focuses on the quantification of the generalization gap when transitioning from the Mediterranean source domain (Italy) to the Baltic target domain (Lithuania). The structural integrity of the predictive engine is defined by the Bi-LSTM stratification (Table **??**), which was engineered to balance complexity and regularization through the integration of dropout layers ($p = 0.2$) and a bidirectional processing flow.

As illustrated in Table 3, the model demonstrates exceptional within-country performance ($R^2 = 0.91$), indicating that the Bi-LSTM effectively captured the non-linear temporal dependencies of the Italian agricultural production. However, when deployed out-of-distribution (OOD) on the Lithuanian dataset, the R^2 drops significantly to 0.64. This 27% loss in explanatory power is the empirical manifestation of the cross-country generalization bias. The "Scale Bias" observed in Figure 1, where predictions show a consistent vertical displacement, suggests that the network has internalized the high-magnitude production intensity of the Italian system. Even after normalization, the model's internal representations remain "anchored" to the source country's agricultural capacity, failing to fully calibrate to the lower-intensity Baltic context. A key finding of this analysis is the role of the "source-only" normalization strategy. By applying Min-Max scaling parameters derived exclusively from the Italian distribution to the Lithuanian test set, we intentionally simulated a zero-shot deployment scenario. The resulting scale bias is a direct consequence of the disparity between the Italian mean production (450,210 t) and the Lithuanian mean (85,100 t). While local re-normalization might improve absolute metrics, our results demonstrate that models trained on high-volume source domains tend to retain a "memory" of the original scale, leading to a systematic overestimation in smaller-scale target domains. This highlights that recalibrating the model's output layer or employing domain-specific scaling is not just a technical preference, but a mandatory requirement for cross-country reliability in agricultural forecasting.

The stratification of errors by product category, reported in Table 4, provides further insights into the structural nature of this bias. Cereals exhibit the highest transferability with a relative error of 12.5%. This suggests that the temporal patterns of grain production are relatively harmonized across Europe, likely due to standardized cultivation practices and common regulatory frameworks under the Common Agricultural Policy (CAP). Conversely, the error for "Industrial Crops" rises to 35.1%, highlighting how these specific crops are more sensitive to local soil-climatic conditions and regional logistics that the Italian-trained model interprets as noise or anomalies. These findings confirm that high within-country accuracy is a deceptive metric for international reliability, as the "Black Box" of Deep Learning tends to encode implicit

socio-economic and environmental features of the training domain that do not seamlessly transfer across borders.

Table 3: Predictive Performance Metrics: Within-Country vs. Cross-Country Evaluation

Evaluation Scenario	**MAE** (t)	**RMSE** (t)	R^2 **Score**
Within-Country (Italy $\rightarrow$ Italy)	12,450.20	18,300.15	0.9142
Cross-Country (Italy $\rightarrow$ Lithuania)	34,120.55	45,980.30	0.6428

Table 4: Error Stratification by Product Category (Target: Lithuania)

Product Category	**MAE** (t)	**Relative Error (%)**
Cereals	21,400.10	12.5%
Barley	42,300.45	28.3%
Industrial Crops	58,120.00	35.1%

4 Conclusion

This study provided a rigorous assessment of cross-country generalization in machine learning models applied to agricultural production data. By training a Bi-LSTM network on Italian historical yields and testing it on the Lithuanian context, we empirically demonstrated the existence of a "cross-country generalization bias." Our findings reveal that high within-country accuracy ($R^2 = 0.91$) is a deceptive metric for models intended for international comparative research, as evidenced by the 27% performance decay and the significant "scale bias" encountered during the Out-of-Distribution test.

A crucial takeaway from our analysis is that the structural relationships and magnitudes learned in a high-intensity source domain do not naturally scale down to smaller agricultural systems. The failure of "source-only" normalization underscores that zero-shot transfer is insufficient for reliable policy-making. We conclude that explicit cross-country validation must become a standard protocol in agricultural economics to prevent systematic errors in food security forecasting. The observed 'Scale Bias' highlights that the model remains deeply conditioned by the high-intensity production logic of the source domain, limiting its reliability in structurally different contexts. To mitigate these biases, future research should move beyond simple supervised learning. We suggest the integration of Transfer Learning (TL) and advanced Domain Adaptation (DA) techniques, such as Domain Adversarial Neural Networks (DANN), to align the latent feature distributions between different national

contexts. By fine-tuning models on small local target-domain subsets, it may be possible to recalibrate the predictive logic to the specific scale of the target country. Furthermore, incorporating domain-specific knowledge—such as agro-climatic zones and CAP-related regulatory data—could help the models distinguish between universal temporal patterns and country-specific structural features. Enhancing the 'portability' of ML models is essential for building resilient global monitoring systems in an increasingly volatile agricultural landscape.

References

[1] M. Ashfaq, I. Khan, D. Shah, S. Ali, and M. Tahir. Predicting wheat yield using deep learning and multi-source environmental data. *Scientific Reports*, 15(1):26446, jul 2025. doi: 10.1038/s41598-025-11780-7.

[2] M. R. Bhardwaj, J. Pawar, A. Bhat, Deepanshu, I. Enaganti, K. Sagar, and Y. Narahari. An innovative deep learning based approach for accurate agricultural crop price prediction. In *2023 IEEE 19th International Conference on Automation Science and Engineering (CASE)*, pages 1–7, 2023. doi: 10.1109/CASE56687.2023.10260494.

[3] L. Cesarini, R. Gonçalves, M. Martina, X. Romão, B. Monteleone, F. L. Pereira, and R. Figueiredo. Comparison of deep learning models for milk production forecasting at national scale. *Computers and Electronics in Agriculture*, 221:108933, 2024.

[4] N. Chergui. Durum wheat yield forecasting using machine learning. *Artificial Intelligence in Agriculture*, 6:156–166, 2022. doi: 10.1016/j.aiia.2022.09.003.

[5] P. Foroutan and S. Lahmiri. Deep learning systems for forecasting the prices of crude oil and precious metals. *Financ. Innovation*, 10(1), 2024. doi: 10.1186/s40854-024-00637-z.

[6] M. R. Hasan. Ai and machine learning for optimal crop yield optimization in the usa. *Journal of Computer Science and Technology Studies*, 6(2): 46–61, 2024.

[7] K. Jhajharia, P. Mathur, S. Jain, and S. Nijhawan. Crop yield prediction using machine learning and deep learning techniques. *Procedia Computer Science*, 218:406–417, 2023.

[8] B. Jin and X. Xu. Machine learning wholesale white wheat price index forecasts. *Quality & Quantity*, 2025. doi: 10.1007/s11135-025-02233-4.

[9] A. Joshi, B. Pradhan, S. Gite, and S. Chakraborty. Remote-sensing data and deep-learning techniques in crop mapping and yield prediction: A systematic review. *Remote Sensing*, 15(8):2014, 2023.

[10] P. Kumari, N. Harshith, and A. Ginige. Integrating recursive feature selection with automated machine learning framework for global wheat price prediction. *Journal of Agriculture and Food Research*, 22:102113, 2025. doi: 10.1016/j.jafr.2025.102113.

[11] M. Kuradusenge et al. Crop yield prediction using machine learning models: Case of irish potato and maize. *Agriculture*, 13(1):225, 2023.

[12] M. Li, P. Wang, K. Tansey, Y. Zhang, F. Guo, J. Liu, and H. Li. An interpretable wheat yield estimation model using an attention mechanism-based deep learning framework. *International Journal of Applied Earth Observation and Geoinformation*, 140:104579, 2025. doi: 10.1016/j.jag.2025.104579.

[13] R. L. Manogna and A. K. Mishra. Forecasting spot prices of agricultural commodities in india: Application of deep-learning models. *Intelligent Systems in Accounting, Finance and Management*, 28(1):72–83, 2021. doi: 10.1002/isaf.1487.

[14] R. L. Manogna, V. Dharmaji, and S. Sarang. Enhancing agricultural commodity price forecasting with deep learning. *Scientific Reports*, 15(1): 20903, 2025. doi: 10.1038/s41598-025-05103-z.

[15] P. Muruganantham et al. A systematic literature review on crop yield prediction with deep learning and remote sensing. *Remote Sensing*, 14 (9):1990, 2022.

[16] P. Pandit et al. Hybrid modeling approaches for agricultural commodity prices using CEEMDAN and time delay neural networks. *Scientific Reports*, 14(1):26639, 2024. doi: 10.1038/s41598-024-74503-4.

[17] D. Paudel et al. Machine learning for regional crop yield forecasting in europe. *Field Crops Research*, 276:108377, 2022.

[18] S. Rani et al. Commodities price prediction using various ml techniques. In *2022 2nd International Conference on Technological Advancements in Computational Sciences (ICTACS)*, pages 277–282, 2022. doi: 10.1109/ICTACS56270.2022.9987967.

[19] S. B. Rufaioglu et al. Sensor-based yield prediction in durum wheat under semi-arid conditions... *Remote Sensing*, 17(14), 2025. doi: 10.3390/rs17142416.

[20] R. Suna and H. Ma. Commodity price fluctuation prediction based on neural network. In *2024 IEEE 3rd International Conference on Electrical Engineering, Big Data and Algorithms (EEBDA)*, pages 1603–1607, 2024. doi: 10.1109/EEBDA60612.2024.10485707.

[21] A. Vairo, G. A. Sgarro, and F. Romano. Predicting commodity prices: a neural networks approach to improve market prospects. *Applied Mathematical Sciences*, 18(6):289–298, 2024. doi: 10.12988/ams.2024.919149.

[22] M. von Bloh et al. Machine learning for soybean yield forecasting in brazil. *Agricultural and Forest Meteorology*, 341:109670, 2023.

[23] T. Zhao, G. Chen, S. Suraphee, T. Phoophiwfa, and P. Busababodhin. A hybrid tcn-xgboost model for agricultural product market price forecasting. *PLOS ONE*, 20(5):1–31, 2025. doi: 10.1371/journal.pone.0322496.

[24] Y. Zhou, S. Ma, H. Zhang, and S. Aakur. Enhancing corn yield prediction: Optimizing data quality or model complexity? *Smart Agricultural Technology*, 9:100671, 2024. doi: 10.1016/j.atech.2024.100671.

Received; March 1, 2026; Published: March 19, 2026