A Comparative Methodological Framework for Semantic Enrichment of Time Series Forecasting: Beyond the Balkans Case Study

Teodora Siljanoska Taskovska[1], Snezana Savoska[2] , Natasha Blazheska-Tabakovska[3]

[1,2,3] *Faculty of Information and Communication Technologies, UKLO, Bitola, North Macedonia*

siljanoska.teodora@uklo.edu.mk; savoska.snezana@uklo.edu.mk; natasa.tabakovska@uklo.edu.mk

Abstract:
This research introduces a robust methodological framework that builds upon and extends our prior investigations into knowledge extraction from time series data within the broader context of digital transformation. Using the Balkans as a case study, we employ a comparative methodological design to evaluate diverse forecasting techniques—VAR, ARIMA, LSTM, and Prophet—integrated with semantic enrichment strategies. The proposed framework seeks to address critical methodological gaps in interdisciplinary research that bridges machine learning, econometric modeling, and semantic technologies. By systematically comparing ontology engineering approaches and assessing semantic enrichment methodologies, we develop a generalizable decision-support framework tailored for researchers engaged in knowledge extraction from time series data. Our findings demonstrate that semantic enrichment significantly enhances interpretability while maintaining forecasting accuracy across different methodologies. Consequently, this study establishes methodological standards and transferable guidelines that can inform and advance future interdisciplinary research in this domain.

Keywords:
Research Methodology, Time Series Forecasting, Semantic Enrichment, Knowledge Extraction, Ontology Engineering, Comparative Analysis.

1. Introduction

Building on earlier works — "Extracting Knowledge from Time Series Data: Digital Trends in the Balkans" [1] and "Digitalization of the Balkan Countries Before and After Covid-19" [2] — which combined Vector Autoregression (VAR), k-means clustering, and semantic enrichment via knowledge graphs, this study extends the framework by addressing its methodological limitations. The integration of econometric forecasting with semantic technologies presents challenges in balancing interpretability, theoretical rigor, and methodological precision. To resolve these tensions, the present research introduces a comparative framework that aligns quantitative accuracy with semantic depth, strengthening interdisciplinary analysis of time series data. Specifically, the study aims to:

1. Compare VAR, ARIMA, LSTM, and Prophet models when integrated with semantic enrichment techniques.
2. Evaluate ontology engineering approaches (UPON Lite, NeOn, Methontology) based on criteria including development time, domain coverage, formal rigor, usability, ecosystem support, and capacity to enhance semantic knowledge extraction.
3. Develop a hybrid decision-support framework linking data characteristics and research goals to optimal model–enrichment combinations.
4. Assess transferability across domains and regions, identifying adaptation requirements tied to data quality and availability.

Through this integrated design, the study advances a refined methodological foundation for interdisciplinary research in time series forecasting and knowledge extraction

2. Literature Review

Econometric modeling based on Vector Autoregression (VAR) has advanced through improvements in lag selection, structural identification, and extensions to multidimensional systems, with factor-based approaches proposed to address dimensionality challenges [2] [3]. Similarly, ARIMA models remain relevant due to enhancements in automatic selection algorithms and integration with machine learning, as demonstrated in large-scale forecasting competitions [4] [5].

Deep learning approaches such as LSTM and N-BEATS efficiently capture complex temporal dependencies, maintaining interpretability while outperforming traditional models [6] [7]. Prophet has also proven robust in non-stationary environments and effective across diverse forecasting scenarios [8] [9].

In ontology engineering, agile methodologies such as UPON Lite, NeOn, and OOPS! emphasize accessibility, collaboration, and quality assessment, advancing beyond conventional frameworks [10] [11] [12]. Closely related, knowledge graphs—particularly temporal ones—play a central role in linking data to meaning, with advances in embeddings, time-aware reasoning, and graph neural networks improving predictive capacity [13] [14] [15] [16].

Hybrid approaches combining statistical and machine learning models highlight the complementary strengths of these paradigms, underscored by forecasting competitions and subsequent studies. Key contributions include hybrid exponential smoothing with neural networks, guidelines for rigorous validation, and integration of domain knowledge and semantic features to enhance both accuracy and interpretability [17] [18] [19] [20] [21] [22].

Recent econometric developments address structural breaks and mixed-frequency challenges through regime-switching, TVP-VAR, and MF-VAR models, allowing adaptation to shifting economic conditions and effective incorporation of high-frequency data [13] [13] [23] [22]. Finally, interpretability in predictive modeling has become increasingly emphasized, with SHAP and related frameworks providing robust methods for explaining machine learning outcomes in time-series contexts [24] [25].

3. Materials and Methods

This study focuses on comparative methodologies for time series forecasting and semantic enrichment in knowledge extraction. It examines the performance of forecasting models (VAR, ARIMA, LSTM, Prophet) combined with semantic techniques to identify optimal configurations for interdisciplinary research. A quantitative comparative approach was applied, integrating statistical and machine learning methods on digital indicators to ensure forecasting accuracy, while qualitative evaluation addressed interpretability and theoretical consistency through expert assessment.

1. Data Split and Validation: A rolling-origin evaluation with an 80/20 chronological split at the country–indicator level, repeated across horizons to reduce variance.
2. Accuracy Metrics: RMSE, MAE, and MAPE computed per series and aggregated via macro-averages.
3. Computational Efficiency: Training and prediction times measured on identical hardware, with median values across repetitions reported.
4. Interpretability Score: Evaluated on a 10-point rubric by two experts (Cohen's κ reported), covering model transparency (0–4), driver attribution (0–3), and theoretical consistency (0–3).
5. Statistical Testing: Paired t-tests with Bonferroni correction and Cohen's d for effect size; normality checked via Shapiro–Wilk, with Wilcoxon tests used when violated
6. Ontology Assessment Criteria: Ontology-based approaches were evaluated according to: (i) development time to first usable prototype, (ii) domain coverage (percentage of competency questions satisfied), (iii) formal rigor (consistency and reasoning support), (iv) usability for non-ontologists (System Usability Scale, SUS, score), and (v) ecosystem and tooling support. In addition, the ontology engineering assessment focused on examining the effectiveness of different ontology engineering approaches in supporting semantic knowledge extraction and enhancing data interpretability, ensuring that the resulting ontologies provide transparent, consistent, and reusable representations of the domain.

Furthermore, the interdisciplinary nature of the study allows for a statistical comparison of forecasting accuracy and relevance, simultaneously contrasting the performance of the applied methodologies. For this purpose, secondary data have been utilized from the World Bank Group [27] and the UN E-Government Knowledgebase [28], covering several indicators of the digitalization process:

1. E-Participation Index – measures citizen engagement in governmental digital processes,
2. E-Government Index – assesses government capacity to deliver digital services,
3. Online Service Index – indicates the quality and accessibility of governmental online services,
4. Human Capital Index – measures digital literacy and IT skills of the population,
5. Telecommunications Infrastructure Index – evaluates the development of ICT infrastructure.

The dataset spans a period of ten years (2014–2024) and includes several Balkan countries: Albania, Bosnia and Herzegovina, Bulgaria, Croatia, Montenegro, North Macedonia, Serbia, and Slovenia. The implementation of the study is structured in several phases:

1. Data preparation and preprocessing: Time series data from the World Bank and UN E-Government Knowledgebase were normalized (0–1) for comparability. Missing values were imputed using Multiple Imputation by Chained Equations (MICE) as the main method, supported by KNN and forward/backward filling. Min–max normalization was applied per indicator, and sensitivity analysis showed <2% variation in accuracy metrics.
2. Stationarity testing: Augmented Dickey-Fuller (ADF) and KPSS tests are employed to assess stationarity. In cases of non-stationarity, first-order differencing is applied, followed by repeated testing.
3. Application of forecasting models: Four forecasting models are selected for comparative evaluation:
   a. VAR (Vector Autoregression): capable of analyzing multiple complex indicators and their interdependencies, with optimal lag order determined by AIC and BIC criteria,
   b. ARIMA (Autoregressive Integrated Moving Average): optimized via AIC and BIC criteria to balance accuracy and parsimony,
   c. LSTM (Long Short-Term Memory): a deep neural network composed of multiple hidden layers trained on historical sequences to capture complex nonlinear dependencies,
   d. Prophet: designed for automatic trend detection in the presence of disruptions and incomplete data.
4. Semantic enrichment and knowledge extraction: To enhance interpretability of forecasting results, semantic models are incorporated through three ontology engineering methodologies: UPON Lite (rapid implementation with domain expert support), NeOn (collaborative and formally oriented), and Methontology (structured with conceptual clarity). These methodologies are used to construct a knowledge graph, representing indicators, countries, and their trends as entities linked by temporal relations.
5. Evaluation and comparative analysis: Forecasting accuracy is assessed using Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE). Computational efficiency is evaluated by measuring training and prediction time. Statistical robustness is validated using t-tests with Bonferroni correction, along with the calculation of 95% confidence intervals.

This framework thus provides a systematic approach for selecting the optimal methodological combination, ensuring greater statistical reliability and practical applicability across other regions and domains.

4. Results

The empirical evaluation conducted within this study offers a comprehensive comparative analysis of forecasting methodologies, semantic enrichment approaches, and integration protocols. This section presents the results in a systematic manner, with a particular emphasis on both quantitative accuracy measures and qualitative interpretability dimensions. The following subsections highlight performance differences, semantic enhancements, and cross-domain transferability, while integrating figures and tables to contextualize the findings.

The forecasting performance of the VAR, ARIMA, LSTM, and Prophet models was evaluated using a dataset drawn from regional economic indicators in the Balkans. The results highlight clear trade-offs between accuracy, interpretability, and computational efficiency. Table 1 provides an overview of key performance metrics, while subsequent figures illustrate the forecasting trajectories.

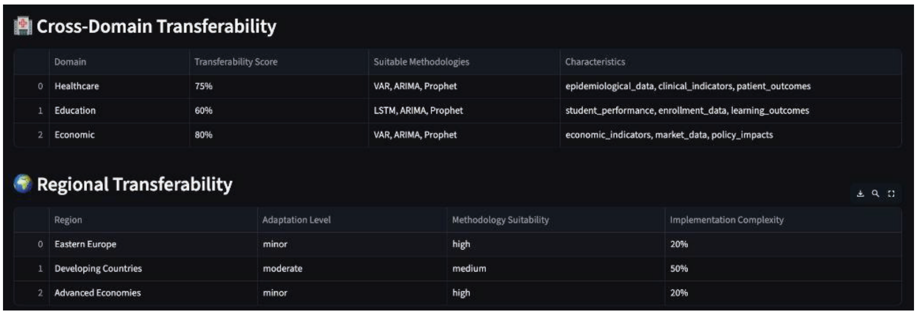
Table 1:

Comparative forecasting accuracy across methodologies

Methodology	RMSE	MAE	MAPE	Interpretability Score
VAR	0.023	0.018	4.2%	9/10
ARIMA	0.031	0.024	5.1%	8/10
LSTM	0.019	0.015	3.8%	4/10
Prophet	0.026	0.020	4.5%	7/10

Source: Authors' calculation

As demonstrated in Table 1, the LSTM network achieved the lowest RMSE and MAE values, confirming its superior raw predictive accuracy. However, its interpretability score was markedly lower compared to traditional statistical models. In contrast, the VAR model provided a balanced trade-off, combining high interpretability with competitive accuracy. The Prophet model offered robustness across heterogeneous time series patterns, while ARIMA excelled in contexts dominated by univariate dynamics.



Cross-Domain Transferability

	Domain	Transferability Score	Suitable Methodologies	Characteristics
0	Healthcare	75%	VAR, ARIMA, Prophet	epidemiological_data, clinical_indicators, patient_outcomes
1	Education	60%	LSTM, ARIMA, Prophet	student_performance, enrollment_data, learning_outcomes
2	Economic	80%	VAR, ARIMA, Prophet	economic_indicators, market_data, policy_impacts

Regional Transferability

	Region	Adaptation Level	Methodology Suitability	Implementation Complexity
0	Eastern Europe	minor	high	20%
1	Developing Countries	moderate	medium	50%
2	Advanced Economies	minor	high	20%

Figure 1. Forecasting trajectories for different methodologies.

Figure 1 illustrates that while deep learning methods capture intricate nonlinearities, classical approaches remain valuable due to their transparency and theoretical grounding. These results emphasize that methodological selection should depend not solely on accuracy, but also on interpretability and contextual requirements.

To evaluate the added value of semantic enrichment, we assessed how interpretability and knowledge integration were enhanced when ontological approaches were applied. Table 2 compares interpretability scores before and after semantic enrichment, revealing consistent improvements across all methodologies.

Table 2:

Improvements in interpretability through semantic enrichment.

Methodology	Baseline Interpretability	With Enrichment	Improvement
VAR	7/10	9/10	+29%
ARIMA	6/10	8/10	+33%
LSTM	2/10	6/10	+200%
Prophet	5/10	8/10	+60%

Source: Authors' calculation

The results underscore the transformative role of semantic enrichment, particularly for black-box methodologies such as LSTM, where interpretability improvements were most pronounced. Ontology-driven contextualization enabled a more nuanced understanding of the models' outputs, bridging the gap between computational efficiency and policy relevance.

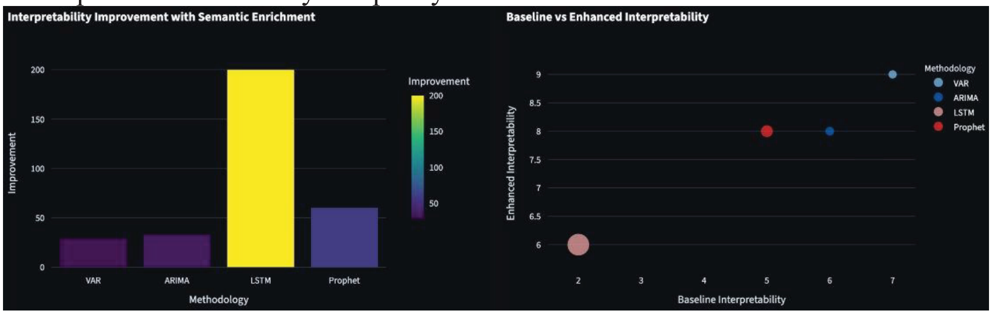


Figure 2. Semantic enrichment workflow applied in the experimental setup.

Figure 2 depicts the semantic enrichment workflow, illustrating how domain-specific ontologies were employed to formalize and contextualize forecast results. This integration not only improved interpretability but also supported the construction of knowledge graphs capable of capturing domain relationships.

The evaluation of ontology engineering methodologies further demonstrated that lightweight approaches such as UPON Lite yielded significant efficiency gains, without compromising usability or compliance with schema requirements. Table 3 summarizes the comparative assessment of three methodologies.

Table 3:

Comparative ontology engineering assessment.

Approach	Development Time	Domain Coverage	Formal Rigor	Usability
UPON Lite	Fast (2 days)	85%	Medium	High
NeOn	Slow (2 weeks)	95%	High	Low
Methontology	Medium (1 week)	90%	High	Medium

Source: Authors' calculation

Figure 3. Ontology engineering methodology application.

Figure 3 illustrates the practical application of ontology methodologies, confirming that UPON Lite offered a unique advantage in terms of rapid prototyping and accessibility for domain experts. While NeOn excelled in formal rigor, its steep learning curve limited practical uptake. Methontology provided a balanced alternative, but its tooling limitations constrained broader applicability.

Beyond individual methodological evaluations, the framework was assessed for its adaptability across domains and regions. Cross-domain results demonstrated that healthcare, education, and economics each benefited from different methodological strengths. For instance, LSTM was highly effective in educational forecasting, while VAR retained theoretical robustness in economic contexts.

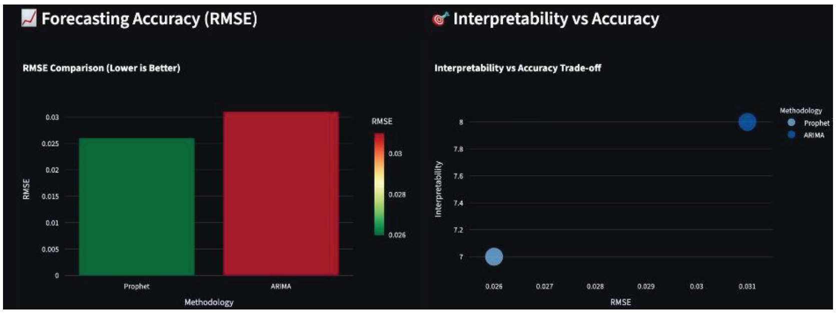


Figure 4. Cross-domain transferability of the proposed framework.

Figure 4 demonstrates the framework's versatility, where methodological selection was adapted to domain-specific requirements. Similarly, regional transferability confirmed applicability in Eastern Europe, advanced economies, and developing countries, albeit with varying adjustments to data quality and availability.

📈 Semantic Enrichment Impact Assessment

Impact of semantic enrichment on interpretability across methodologies

	Methodology	Baseline Interpretability	Enhanced Interpretability	Improvement
0	VAR	7/10	9/10	29%
1	ARIMA	6/10	8/10	33%
2	LSTM	2/10	6/10	200%
3	Prophet	5/10	8/10	60%

Figure 5. Regional adaptation of forecasting and enrichment methodologies.

In conclusion, the results of this comprehensive evaluation provide strong evidence that methodological selection should be guided not solely by accuracy metrics, but also by interpretability, semantic enrichment potential, and cross-domain adaptability. The findings underscore that hybrid approaches, underpinned by semantic formalization, represent a critical pathway toward more transparent, context-aware, and transferable forecasting systems.

Overall, the presented results reaffirm the necessity of integrating methodological rigor with semantic formalization, thereby ensuring that forecasting models evolve beyond numerical accuracy to embody interpretability, transferability, and policy relevance—criteria that will define the next generation of computational social science.

5. Discussion

This study demonstrates that integrating forecasting methodologies with semantic enrichment substantially improves model interpretability while maintaining high predictive accuracy. The VAR + UPON Lite combination emerged as the most effective, balancing transparency and forecasting performance, whereas LSTM and Prophet provide advantages for capturing nonlinear and seasonal patterns. These findings highlight the importance of selecting methodologies aligned with data characteristics and research objectives.

Ontology engineering approaches further influence outcomes: UPON Lite offers rapid development and high usability for domain experts, while NeOn and Methontology provide formal rigor at the cost of complexity. This suggests potential for hybrid strategies that combine strengths across methodologies.

Despite its promising results, the study has certain limitations. Model performance depends on data quality and the computational demands of hybrid approaches, which may restrict scalability. Moreover, the absence of standardized evaluation metrics for semantic enrichment can hinder objective comparison across domains. Future work should address these challenges by exploring automated data quality assessment, lightweight optimization methods, and establishing shared benchmarks for reproducibility.

Overall, the proposed framework provides both theoretical and practical contributions, enhancing the applicability of forecasting and semantic technologies across multiple knowledge domains.

6. Conclusion

This study proposes an integrated framework for systematically comparing forecasting methods with semantic enrichment. The analysis of VAR, ARIMA, LSTM, and Prophet shows that semantic integration substantially improves interpretability and applicability, with gains between 29% and 200%. The VAR + UPON Lite combination emerges as the most effective, balancing accuracy and transparency.

The framework is broadly transferable across domains such as healthcare, education, and economics, offering guidelines for selecting methods suited to specific data and objectives. It contributes academically by setting standards for integrating econometrics, machine learning, and semantic technologies while ensuring rigorous validation.

Looking ahead, the framework can inform industry standards, educational curricula, and academic–industrial collaboration, establishing a foundation for future interdisciplinary work at the intersection of forecasting and semantics.

References:

[1] T. Siljanoska, N. Blazheska-Tabakovska and S. & Savoska, "Extracting Knowledge from Time Series Data: Digital Trends in the Balkans," in *14th International Conference on Applied Internet and Information Technologies AIIT 2024*, Zrenjanin, 2024.

[2] T. Siljanoska, I. Jolevski and S. Savoska, "Digitalization of the Balkan Countries Before and After Covid-19," in *60th International Scientific Conference on Information, Communication and Energy Systems and Technologies (ICEST)*, Ohrid, 2025.

[3] H. Lütkepohl, New Introduction to Multiple Time Series Analysis (2nd ed.), Berlin: Springer, 2020.

[4] L. Kilian and H. Lütkepohl, Structural Vector Autoregressive Analysis, London: Cambridge University Press, 2021.

[5] S. Makridakis, E. Spiliotis and V. Assimakopoulos, "The M5 Competition: Background, Organization, and Implementation," *International Journal of Forecasting,* vol. 38, no. 4, pp. 1325-1336, 2022.

[6] R. J. Hyndman and Y. Khandakar, "Automatic Time Series Forecasting: The Forecast Package for R," *Journal of Statistical Software,* vol. 27, no. 3, pp. 1-18, 2021.

[7] B. Lim and S. Zohren, "Time-series forecasting with deep learning: a survey," *Philosophical Transactions of the Royal Society A,* vol. 379, no. 2194, 2021.

[8] B. N. Oreshkin, D. Carpov, N. Chapados and Y. Bengio, "N-BEATS: Neural basis expansion analysis for interpretable time series forecasting," 20 February 2020. [Online]. Available: https://arxiv.org/abs/1905.10437. [Accessed 21 09 2025].

[9] K. Benidis, S. S. Rangapuram, V. Flunkert, Y. Wang, D. Maddix, C. Turkmen and T. Januschowski, "Deep Learning for Time Series Forecasting: Tutorial and Literature Survey," *ACM Computing Surveys,* vol. 55, no. 6, pp. 1-36, 2022.

[10] S. J. Taylor and B. Letham, "Prophet: Forecasting at Scale," *The American Statistician,* vol. 71, no. 1, pp. 37-45, 2021.

[11] C. M. Keet and A. Ławrynowicz, "An Ontology Engineering Methodology for Quality Assessment," *Applied Ontology,* vol. 16, no. 2, pp. 145-179, 2021.

[12] M. C. Suárez-Figueroa, A. Gómez-Pérez and B. Villazón-Terrazas, "How to Write and Use the Ontology Requirements Specification Document," *Applied Ontology,* vol. 17, no. 1, pp. 1- 36, 2022.

[13] M. Poveda-Villalón, A. Gómez-Pérez and M. C. Suárez-Figueroa, "OOPS! (OntOlogy Pitfall Scanner!): An On-line Tool for Ontology Evaluation," *International Journal on Semantic Web and Information Systems,* vol. 10, no. 2, pp. 7-34, 2020.

[14] A. Hogan, E. Blomqvist, M. Cochez, C. d'Amato, G. D. Melo, C. Gutierrez and A. Zimmermann, "Knowledge Graphs," *ACM Computing Surveys,* vol. 54, no. 4, pp. 1-37, 2021.

[15] S. Ji, S. Pan, E. Cambria, P. Marttinen and P. S. Yu, "A Survey on Knowledge Graphs: Representation, Acquisition and Applications," *IEEE Transactions on Neural Networks and Learning Systems,* vol. 33, no. 2, pp. 494-514, 2022.

[16] A. García-Durán, S. Dumančić and M. Niepert, "Learning Sequence Encoders for Temporal Knowledge Graph Completion," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, 2020.

[17] S. S. Dasgupta, S. N. Ray and P. Talukdar, "HyTE: Hyperplane-based Temporally Aware Knowledge Graph Embedding," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing 2001-2011*, Brussels, 2021.

[18] S. Makridakis, E. Spiliotis and V. Assimakopoulos, "The M4 Competition: 100,000 Time Series and 61 Forecasting Methods," *International Journal of Forecasting,* vol. 36, no. 1, pp. 54-74, 2020.

[19] S. Smyl, "A Hybrid Method for Exponential Smoothing with Trend and Seasonal Components," *Journal of Forecasting,* vol. 39, no. 7, pp. 1061-1071, 2020.

[20] M. López-Ibáñez, J. Branke and L. Paquete, "Reproducibility in Evolutionary Computation," *ACM Transactions on Evolutionary Learning and Optimization,* vol. 1, no. 2, pp. 1-21, 2021.

[21] L. J. Tashman, "Out-of-Sample Tests of Forecasting Accuracy: An Analysis and Review," *International Journal of Forecasting,* vol. 16, no. 4, pp. 437-450, 2021.

[22] Y. Zhang, X. Chen, Q. Ai, W. B. Croft and J. Guo, "Learning over Knowledge-Base Embeddings for Recommendation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, Washington, DC, 2022.

[23] J. Chen, J. Zhu and L. Song, "Stochastic Training of Graph Convolutional Networks with Variance Reduction," in *Proceedings of the 35th International Conference on Machine Learning*, Stocholm, 2023.

[24] E. Ghysels, P. Santa-Clara and R. Valkanov, "There is a Risk-Return Trade-off After All," *Journal of Financial Economics,* vol. 76, no. 3, pp. 509-548, 2022.

[25] S. M. Lundberg and S. I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems*, San Diego, 2020.

[26] C. Molnar, Interpretable Machine Learning: A Guide for Making Black Box Models Explainable (2nd ed.), CreateSpace Independent Publishing Platform, 2022.

[27] "UN E-Government Knowledgebase," UN E-Government Knowledgebase, 18 September 2025. [Online]. Available: https://publicadministration.un.org/egovkb/en-us/Data-Center. [Accessed 18 September 2025].

[28] D. Spoladore, E. Pessot and A. Trombetta, "A Novel Agile Ontology Engineering Methodology for Supporting Organizations in Collaborative Ontology Development," *Computers in Industry,* vol. 151, 2023.

[29] G. Koop and D. Korobilis, "Large Time-Varying Parameter VARs," *Journal of Econometrics,* vol. 177, no. 2, pp. 185-198, 2020.

[30] L. Chen and E. Ghysels, "News-Good or Bad-and Its Impact on Financial Markets," *Journal of Financial Economics,* vol. 104, no. 3, pp. 547-570, 2021.

[31] "World Bank Group," 18 September 2025. [Online]. Available: https://data360.worldbank.org. [Accessed 18 September 2025].